Model Selection Criteria: An Investigation of Relative Accuracy, Posterior Probabilities, and Combinations of Criteria

Roland T. Rust • Duncan Simester • Roderick J. Brodie • V. Nilikant
Owen Graduate School of Management, Vanderbilt University, Nashville, Tennessee 37203
Graduate School of Business, University of Chicago, 1101 *E.* 58*th St., Chicago, Illinois* 60637
University of Auckland, Auckland, New Zealand
University of Canterbury, Canterbury, United Kingdom

We investigate the performance of empirical criteria for comparing and selecting quantitative models from among a candidate set. A simulation based on empirically observed parameter values is used to determine which criterion is the most accurate at identifying the correct model specification. The simulation is composed of both nested and nonnested linear regression models. We then derive posterior probability estimates of the superiority of the alternative models from each of the criteria and evaluate the relative accuracy, bias, and information content of these probabilities. To investigate whether additional accuracy can be derived from combining criteria, a method for obtaining a joint prediction from combinations of the criteria is proposed and the incremental improvement in selection accuracy considered.

Based on the simulation, we conclude that most leading criteria perform well in selecting the best model, and several criteria also produce accurate probabilities of model superiority. Computationally intensive criteria failed to perform better than criteria which were computationally simpler. Also, the use of several criteria in combination failed to appreciably outperform the use of one model. The Schwarz criterion performed best overall in terms of selection accuracy, accuracy of posterior probabilities, and ease of use. Thus, we suggest that general model comparison, model selection, and model probability estimation be performed using the Schwarz criterion, which can be implemented (given the model log likelihoods) using only a hand calculator.

(*Model Selection; Cross-Validation; Akaike's Criterion; Schwarz' Criterion; AIC; BIC; BCVL; Bayesian Methods; Combination of Forecasts*)

1. Introduction

Research in forecasting and other areas of management science often involves the selection of the best available model from among a candidate set. For example, forecasting market share may involve selecting between an attraction model and a multiplicative model or the selection between a naive extrapolation model and an econometric model. Often the competing models reflect different or even conflicting assumptions, and the selection of a model thus amounts implicitly to acceptance of the assumptions and theoretical structure of the accepted model.

We focus on the selection of models using empirical criteria. There are many such criteria available to the researcher, each of which evaluates the relative consistency of the alternative models with the available data. We use a simulation based on empirically observed parameter values to test the performance of the leading criteria. We begin by asking which criterion is the most accurate at identifying the correct model. We then show

0025-1909/95/4102/0322$01.25

how posterior probabilities of the superiority of the alternative models can be derived from the various criteria, and evaluate the relative accuracy, bias, and information content of these probabilities. Finally, following the approach adopted in the combination of forecasts literature, we develop a method for combining the different criteria and evaluate the incremental accuracy that results.

Previous research has focused solely on identifying which criterion is the most accurate. This work has been both analytical and empirical. The analytical work has been used largely to motivate the development of new criteria. For example, Allenby (1990) introduces his criterion by arguing that the Akaike and Schwarz criteria over-emphasizes parsimony, while his does not. This finding conflicts with Shibata's (1976) earlier finding that Akaike under-emphasizes parsimony. Our results will help to resolve this apparent contradiction.

The previous empirical work has appeared in different fields and has resulted in few consolidating generalizations. Rust and Schmittlein (1985) compared how well different criteria were able to discriminate between alternative piecewise linear regression models. Their results gave some indication that the simple penalty function approaches of Akaike and Schwarz were capable of performing just as well as more complicated criteria. However, they considered only nonnested models of equal dimension. Clayton et al. (1986) also restricted themselves to models of equal dimension, investigating competing exponential and normal models. This study also gave preliminary indications that a simple penalty function approach (Akaike) could perform as well as or better than the more cumbersome cross-validation procedures. Homburg (1991) considered competing nonnested covariance structure models. The Akaike and Schwarz criteria performed substantially better than the cross-validation approaches, with the Schwarz criterion doing the best of all.

When evaluated separately, these studies offer few general conclusions beyond the particular model specifications that are considered. When viewed together, the studies appear to suggest that the simple model selection approaches, such as the Akaike and Schwarz criteria, are every bit as good as (or better than) the more complicated approaches. In each of these studies, the simulations relied upon model parameters arbitrarily chosen by the researchers. We seek to construct a truer test, by basing our simulations on actual data sets and empirically estimated parameters. We also enrich our model comparison by including both nested and nonnested alternatives and comparisons of models of both equal and unequal dimensions.

While it previously has been shown that it is possible to derive posterior probabilities of model superiority from many of the criteria (Rust and Schmittlein 1985), we extend this finding to a more general class of criteria. We then compare these posterior probabilities with the identity of the correct model, to identify which of the resulting posterior predictions are the most accurate, which show evidence of bias (over-confidence or under-confidence), and which provide the most information to the researcher.

In the forecasting literature, it is well established that combining the forecasts of several forecasters can improve forecasting accuracy (Clemen 1989, Makridakis and Winkler 1983, Bates and Granger 1969). Researchers have also shown that when forecasts are positively correlated, the incremental benefit from additional forecasts may be surprisingly small (Clemen and Winkler 1985, Morrison and Schmittlein 1991). Also, when the costs of obtaining the forecasts is considered, the optimal number of forecasts becomes even smaller (Chen and Anandalingam 1990). Combination of forecasts is generally recognized to be a subset of the general topic of the combination of models. The present research is concerned not with the combination of models, but rather with the combination of criteria for evaluating models. Thus we wish to select the single best model, but are willing to consider a combination of model selection criteria. The combination of forecasts literature suggests that if the criteria are sufficiently independent, use of multiple criteria may improve model selection. Very similar criteria are probably redundant and do not justify the additional effort required. If all the criteria are highly dependent, then little will be gained from using multiple criteria. Theoretical considerations are not available to address these issues. Empirical exploration is necessary but has not previously been attempted.

The different model selection criteria are introduced in §2 together with a description of the approaches used to derive posterior probabilities from each criteria.

The accuracy and any apparent biases of these criteria in identifying the correct model specification are investigated in §3. Also discussed in §3 are the relative accuracy, bias, and information content in the posterior probabilities. Section 4 contains an assessment of the merits of combining criteria, and the paper concludes with a review of the findings and a recommendation as to which criterion should be used.

2. Model Selection Criteria

We consider only *general* model selection criteria—general enough to require only that the competing models have a likelihood function and a finite number of estimated parameters. As a result, we do not limit the scope of the research to criteria capable only of evaluating nested models. For the purposes of this paper, we further limit our attention to criteria which have been prominent in the recent literature. All the criteria we investigate have appeared in the last 25 years.

To assist in introducing and describing each criterion, the criteria are classified into *split sample*, *jackknife*, and *full sample* criteria. Split sample criteria require that the empirical data be divided into two parts: an estimation sample and a validation sample. Model parameters are estimated on the estimation sample, and then model performance is tested on the validation sample. Jackknife criteria (Stone 1974, Geisser and Eddy 1979, Cooil et al. 1987) do a similar cross-validation, one observation at a time. Each data point is held out in turn, the model parameters are estimated on the rest of the points, and the likelihood of the holdout point is evaluated. A "pseudo-likelihood," the product of the individual point likelihoods, is then computed and used to compare model performance.

Full sample criteria calculate the maximum likelihood and then adjust for parsimony by subtracting a penalty term which is an increasing function of the number of estimated parameters. These criteria are the easiest computationally, and have gained widespread popularity as a result.

2.1. Split Sample Criteria

The two major criteria proposed are the Predictive Sample Reuse Quasi-Bayes (PSRQB) criteria of Geiser and Eddy (1979) and the Cross-Validated Likelihood (CVL) criteria (Stone 1974, Geisser and Eddy 1979, Rust and Schmittlein 1985). Both these criteria can be used on either a split sample basis or a jackknife basis, so we will refer to the split sample versions as PSRQB(Split) and CVL(Split).

2.1.1. PSRQB(Split). The general idea behind PSRQB (Geisser and Eddy 1979) is to begin with diffuse priors for the parameters, and then update the parameter distribution, based on the estimation sample. This posterior distribution for the parameters is then used to obtain the likelihood of the validation sample. In general, this approach leads to a discouraging amount of computational complexity, including the necessity of performing numerical integration in what could be a high-dimensional space.

However, for linear regression models a closed form expression for the likelihood may be obtained, based on results giving the conditional predictive densities of regression models (Zellner 1971, p. 235). If X is the estimation sample predictor variable matrix (rows corresponding to observations and columns corresponding to variables), Z is the validation sample predictor variable matrix, $\hat{\beta}$ is the vector of estimated coefficients, W is the vector of holdout Y's (where Y is the dependent variable), S is the sample sum of square error, k is the number of observations in the estimation sample, p is the number of observations in the validation sample, and v is the number of predictor variables, then the predictive likelihood is, following Zellner (1971, p. 235) and using results of Dickey (1967),[1]

$$\text{PSRQB(Split)} = C|S + (W - Z\hat{B})'(I - ZM^{-1}Z') \times (W - Z\hat{B})|^{(k+p-v-1)/2}, \quad (1)$$

where

$$M = X'X + Z'Z,$$

$$C \propto \frac{S^{(k-v-1)/2}|I - ZM^{-1}Z'|^{1/2}}{f(k+p-v-1, p, 1)}, \quad \text{and}$$

$$f(x, y, z) = \frac{\Pi^{yz/2}\Gamma_z[(x-y)/2]}{\Gamma_z[x/2]},$$

$$\Gamma_z(\lambda) = \Pi^{z(z-1)/4}\Gamma(\lambda)\Gamma\left(\lambda - \frac{1}{2}\right)\cdots\Gamma\left(\lambda - \frac{z}{2} + \frac{1}{2}\right).$$

2.1.2. CVL(Split). CVL (Stone 1974, Geisser and Eddy 1979, Rust and Schmittlein 1985) is a very general

[1] The authors are grateful to Bruce Cooil for his help in deriving the "constant" of proportionality for the PSRQB linear regression case.

and computationally simpler quasi-Bayesian model comparison criterion. The central idea is to estimate parameters on the estimation sample, and then evaluate the likelihood on the validation sample. For the case of one criterion variable Y, with a normal error term, for example, the cross-validated likelihood is:

$$\text{CVL(Split)} = \prod_{i=1}^{p} \left\{ (2\Pi)^{-1/2} \hat{\sigma}^{-1} \exp\left[-\frac{1}{2} \left(\frac{(Y_i - \hat{Y}_i)}{\hat{\sigma}} \right)^2 \right] \right\}. \quad (2)$$

2.2. Jackknife Criteria

Again, the two major criteria are based on PSRQB and CVL. We refer to the jackknife versions of these criteria, following previous notational conventions (Stone 1974, Rust and Schmittlein 1985) as PSRQB(L^*) and CVL(L^*). The L^* designation refers to the cross-validated pseudo-likelihood, using a resampling procedure which treats each data point in turn as the validation sample and the remaining points as the estimation sample (Stone 1974, Geisser and Eddy 1979).

2.2.1. PSRQB(L^*). To calculate PSRQB(L^*) we again potentially will be forced to conduct numerical integrations in high-dimensional spaces, among other computational indignities. Only this time there is an integration for every holdout point! Again, things become manageable for the case of linear regression models. Let $Y_{(i)}$ be the criterion variable vector, excluding point i, $X_{(i)}$ be the independent variable matrix excluding point i, $\hat{\beta}_{(i)}$ be the coefficient vector estimated from all points besides i, $S_{(i)}$ be the sample sum of square error, excluding point i, N is the total sample size, and all other notation as defined previously. Then the quasi-Bayes likelihood, again following Zellner and Dickey, is

$$\text{PSRQB}(L^*) = \prod_{i=1}^{N} \{ C | S_{(i)} + (Y_i - X_i \hat{\beta}_{(i)})' \times (I - X_i M_i^{-1} X')(Y_i - X_i \hat{\beta}_{(i)}) | \}^{(N-v-1)/2}, \quad (3)$$

where

$$M_i = X'_{(i)} X_{(i)} + X'_i X_i,$$

$$C \propto \frac{S_{(i)}^{N-v-2} | I - ZM^{-1}Z' |^{1/2}}{f(N - v - 1, 1, 1)},$$

and $f(\cdot)$ is defined as previously.

2.2.2. CVL(L^*). The simplifying assumptions of CVL make CVL(L^*) much easier to compute (in general) than PSRQB(L^*). Let $L_{(i)}$ be the likelihood of Y_i, obtained from the parameter estimates from an estimation sample which deletes point i. Define:

$$L^* = \sum_{i=1}^{N} \ln L_{(i)}. \quad (4)$$

Then the cross-validated pseudo-likelihood used by CVL(L^*) is

$$\text{CVL}(L^*) = \exp(L^*). \quad (5)$$

2.3. Full Sample Criteria

Full sample criteria do not rely on data splitting to obtain their fit statistics, but rather are based on the log likelihood on the data as a whole. Three full sample criteria we will consider are those proposed by Akaike, Schwarz, and Allenby.

(1) Akaike. Akaike (1974) proposed an astonishingly simple model comparison criterion, based on an information theoretic rationale. This criterion uses a penalty term to penalize the log maximum likelihood for lack of parsimony. If ln L is the log maximum likelihood, Akaike's criterion is computed as[2]

$$A = \ln L - (\text{number of parameters}). \quad (6)$$

In the regression case, with v independent variables, there are $v + 1$ total estimated regression coefficients, yielding:

$$A = \ln L - v - 1. \quad (6a)$$

(2) Schwarz. Schwarz (1978) criticized Akaike's criterion as being asymptotically nonoptimal and provided a simple alternative, based on a Bayesian argument. His mathematical results lead to a revised form of the penalty function, but again one which is simple computationally. His criterion is:

$$B = \ln L - \left[\frac{\ln n}{2} \cdot (\text{number of parameters}) \right] \quad (7)$$

(3) Allenby. A third full sample criterion has been

[2] Both Akaike's criterion and Schwarz's criterion often appear in slightly different form, multiplied by a numerical constant (usually $-\frac{1}{2}$).

proposed by Allenby (1990). Using an approximation based on a Bayesian approach, he arrives at the criterion

$$C = \ln L - \left[\frac{\ln 2}{2} \cdot (\text{number of parameters})\right]. \quad (8)$$

2.4. Posterior Probability Criteria

Consider the case in which a researcher is selecting either model 1 or model 2. Suppose that selection of the correct model leads to an expected benefit of one unit of utility (there is something to be gained from choosing the correct model). Suppose also that the difference in cost between employing model 1 and employing model 2 is ΔC. Let $P(1)$ and $1 - P(1)$ be the respective posterior probabilities assigned to models 1 and 2.

Under what circumstances should the researcher select model 1? He/she should select model 1 only if the difference in expected benefits exceeds the difference in costs. That is, model 1 should be selected iff:

$$P(1) - [1 - P(1)] > \Delta C, \quad \text{or}$$

$$P(1) > \frac{1 + \Delta C}{2}. \quad (9)$$

Thus we see that selecting the most appropriate model depends on both the model posterior probabilities and the difference in cost of implementation. Note that if costs of implementation are equal, then the model with the higher posterior probability should be chosen. In such circumstances the posterior probabilities provide a measure of the extent to which a criterion favors the preferred model. If ΔC is large enough, then model 1 should never be chosen. Conversely, if ΔC is negative enough, then model 1 should always be chosen. Thus model posterior probabilities may be used, in conjunction with implementation costs, to trade off the benefit of selecting the better model against the comparative cost of its implementation.

As it turns out, all the criteria described earlier in this section may also be used to generate posterior odds or, if one assumes the correct model is in the candidate set, posterior probabilities. In this section and the remainder of the paper we will, for the sake of consistency and the familiarity of Bayesian posterior probability calculations, make the required assumption and refer to the *probability of model correctness or superiority*.[3]

[3] Not making the required assumption merely results in a renormalization, but does not affect the relative results.

The criteria of §§2.1 and 2.2 all produce cross-validated likelihoods for the competing models. These are then easily converted to posterior odds or posterior probabilities. For example, if L_j denotes the cross-validated likelihood (or pseudo-likelihood) of model j then the posterior probability of model j^* is:

$$P(j^*) = \frac{L_{j^*}}{\sum_j L_j}, \quad (10)$$

if the prior probabilities of the models are equal. Otherwise, if p_j is the prior probability of model j, then:

$$P(j^*) = \frac{p_{j^*} L_{j^*}}{\sum_j p_j L_j}. \quad (11)$$

Although it is not widely realized, all the criteria of §2.3 also may be used to produce model posterior probabilities. For example, Stone (1977) showed that Akaike's criterion is asymptotically equivalent to the jackknifed pseudo-likelihood whenever the model is correct. This suggests the use of Akaike's criterion to produce posterior probabilities, given a set of correct or approximately correct models (Rust and Schmittlein 1985). Preliminary tests have shown this approach to work well (Fornell and Rust 1989). The approximate "posterior" model likelihood is given as:

$$L = \exp(A). \quad (12)$$

Similar approximations also result directly from the criteria of Schwarz and Allenby.[4] Because both criteria are approximations to the posterior log likelihood, the likelihood approximations are simply

$$L = \exp(B). \quad (13)$$

$$L = \exp(C). \quad (14)$$

This use of the Schwarz criterion is new, but follows logically from its derivation.

2.5. Transformed Dependent Variables

If the dependent variable is transformed in one of the competing models, then the likelihood itself must be transformed, through the use of a Jacobian. Let us consider, for the sake of illustration, the competing models

$$Y = \alpha + \beta_1 x_1 + \beta_2 X_2 + \epsilon,$$

$$\ln Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon. \quad (15)$$

[4] Allenby (1990) explicitly shows how his approximation may be used to generate posterior odds.

Let L_1 and L_2 be the respective likelihoods for the two models, in terms of Y and ln Y, respectively. Then L must be transformed, because it is in terms of the transformed dependent variable ln Y. Then

$$L_2(Y) = L_2 \cdot \prod_i \left| \frac{1}{Y_i} \right|. \quad (16)$$

In general, if $w(y)$ is a transformation of Y, then the untransformed likelihood is

$$L_2(Y) = L_2 \cdot \prod_i J(Y_i), \quad (17)$$

where $J(Y_i)$ is the Jacobian of the transformation, evaluated at data point i.

3. Which Criteria Are Best?

In this section we address the question of which of the general model selection criteria perform best in selecting the correct model, when the correct model is in the consideration set. We assess the performance of the different criteria by how often they identify the correct model and whether there is any evidence of bias in the selections. The section also contains an assessment of the relative accuracy, presence of bias, and information content of the posterior probabilities derived from each criterion.

3.1. An Empirically Derived Simulation

We based our simulation on 15 empirical data sets which had previously been used to test the forecasting accuracy of market share models (Brodie and DeKluyver 1984). The data sets represented 15 brands from three New Zealand markets. Three of the brands were chocolate biscuits (cookies), five were liquid detergents, and seven were toothpastes. For each brand there were 28 bimonthly observations (1975–1980) of market share, advertising, price, distribution, and other marketing variables, extracted from Nielsen retail audits and a national advertising audit.

We examined two sets of models. The first set of models is a static model versus a dynamic model of market share. The static model is[5]

$$MS_t = \alpha + \beta_1 P_t + \beta_2 D_t + \beta_3 A_t + \epsilon \quad \text{where} \quad (18)$$

MS_t = market share at period t;
P_t = price at t;
D_t = distribution at t;
A_t = advertising at t;
ϵ = error term, assumed i.i.d. normal; and
α = constant.

The dynamic model includes a lagged dependent variable, MS_{t-1}

$$MS_t = \alpha + \beta_1 P_t + \beta_2 D_t + \beta_3 A_t + \beta_4 MS_{t-1} + \epsilon. \quad (19)$$

This set of models is nested.

The second set of models is a multiplicative (log) model versus a linear model. The multiplicative model is

$$\ln MS_t = \alpha + \beta_1 \ln P_t + \beta_2 \ln D_t + \beta_3 \ln A_t + \beta_4 \ln MS_{t-1} + \epsilon, \quad (20)$$

while the linear model is equivalent to the above dynamic model (equation 19). The multiplicative model is linearized using logs, and thus both models are inherently linear, although they are not nested.

Both competing models were estimated on all 15 data sets, for both sets of models. We then constructed 675 simulated data sets, 15 replications for each of the 15 sets of empirical data, for each of the three distinct conditions (true model is static/linear, dynamic, or log). For example, for the static model condition, the coefficient estimates and error variance estimate were assumed to be the true values, for each of the 15 data sets. Then the 15 replication data sets were constructed by simulating a random error term around the predicted dependent variable value. Thus, the simulated data sets are reflective of the actual data, even though the dependent variable values are simulated. All the independent variable intercorrelations are retained.

We tested how accurately each criterion chose the correct model and produced posterior probabilities of model correctness. For the purpose of comparison, we also tested several nongeneral criteria: a nested F test and a likelihood ratio chi-square (both at the 95% level) for the static vs. dynamic comparison; the Box-Cox criterion for the linear vs. log comparison; and adjusted R^2 (ADJR) for both comparisons (not general because a single dependent variable is required). We can thus see whether or not the general criteria do as well as

[5] See Brodie and DeKluyver (1984) for detailed descriptions of the operationalizations of the variables.

Table 1 Model Selection Accuracy: Comparison of Individual Criteria

Criteria	Static (18)*	Dynamic (19)	Linear (20)	Log (19)	Total	% Correct	Rank
Schwarz	201[1]	190	174	173	738[2]	82.0%	1
Akaike	179	193	174	173	719	79.9%	2
CVL(L^*)	199	182	172	165	718	79.8%	3
PSRQB(L^*)	192	161	178	170	701	77.9%	4
ADJR	148	205	169	167	689	76.6%	5
Allenby	124	210	174	173	681	75.7%	6
PSRQB(Split)	141	172	159	153	625	69.4%	7
CVL(Split)	123	139	150	136	548	60.9%	8
Log L(95%)	209	188	n/a	n/a			
F(95%)	216	178	n/a	n/a			
Box-Cox	n/a	n/a	174	173			

[1] Number correct out of 225 replications.

[2] Number correct out of 900 replications.

* Equation number (see text).

criteria which were designed specifically for the kind of model comparison encountered.

3.2. Model Selection Accuracy

Table 1 summarizes the performance of the various criteria, for the four true model conditions. There were a total of 225 selections in each cell. We see that the Schwarz and Akaike criteria, as well as both the jackknife criteria, do a consistently good job of selecting the correct model. In the dynamic-static comparison, the traditional F-test and likelihood ratio test do a very good job of picking the correct model, although they are not significantly better than Schwarz or CVL(L^*).

Dividing the analysis into the static-dynamic and linear-log cases provides further insights. In Table 2 we test whether any of the criteria are significantly biased in favor of one of the model specifications. Table 2 reports the z-scores for tests of whether the success rates are significantly different for the competing models (assuming a *Bernoulli process*). In the static-dynamic case we see that again the Schwarz, Akaike, and CVL(L^*) criteria perform best. Allenby (1990) had suggested that the Akaike and Schwarz criteria would overly emphasize parsimony, while his would not. The comparison of proportions test indicates that the hypotheses that Akaike and Schwarz each produce equal selection probabilities for the static vs. more complicated dynamic model and cannot be rejected at the 95% level. Allenby's criterion, on the other hand, significantly favors the less parsimonious model, which would suggest that the approximation on which Allenby's criterion is based may contain a systematic bias. This may be caused by Allenby's use of the approximation of the fit on the entire sample as a proxy for cross-validated fit. This

Table 2 Selection Bias by Model Selection Scenario

Criteria	Dynamic vs. Static (z-score)	Linear vs. Log (z-score)
Schwarz	−1.54[1]	0.11[2]
Akaike	0.22	0.11
CVL(Split)	1.53	1.37
PSRQB(L^*)	−3.55*	0.90
CVL(L^*)	−2.24*	0.76
PSRQB(Split)	−3.21*	0.61
Allenby	9.27*	0.11
ADJR	6.87*	0.22
F	−5.61*	n/a
Box-Cox	n/a	0.11

[1] The minus sign indicates static is preferred, but in this case not significantly.

[2] The positive sign indicates linear is preferred, but in this case not significantly.

* Significant at 0.05 level. A z-score greater than 1.96 (2.58) would suggest that the probability that a method significantly prefers one of the two models is less than 5% (1%).

results in an advantage for more complicated models, because their fit has fewer degrees of freedom.

The linear vs. log comparison shows that the full sample criteria all do equivalently well because model parsimony does not influence selection (the log and linear models have the same number of parameters). The jackknife criteria also do well, and the split criteria do somewhat worse. Table 2 shows that none of the criteria has a significant bias with regard to the linear-log comparison.

3.3. Posterior Probability Accuracy

The results in Table 3 allow us to evaluate the relative accuracy, bias, and information content in the posterior probabilities derived from each criterion. Each of the criteria was used to generate 900 posterior probability estimates.[6] To calculate actual probabilities of success for each criterion, these 900 estimates were sorted in order of confidence between 0.5[7] and 1.0. The estimates were then sequentially allocated into 45 ordered groups (20 estimates in each group). For each of these 45 groups a *mean posterior probability* was calculated by finding the mean of the 20 posterior probability estimates. *Mean actual success* was measured by the proportion of times (out of 20) that the criterion chose the true model (gave probability greater than or equal to 0.5 that the true model was correct).

Accuracy and bias in the posterior probability estimates from each criterion were evaluated by regressing the mean posterior probability estimate against mean actual success. Perfect predictions would result in a constant of zero, coefficient of one, and all the variance in the dependent variable being explained. Investigation of the information content in the various posterior probability estimates recognizes that a criterion can perform well in this regression test by making few confident predictions. If the criterion reports that both alternatives are equally likely, its posterior probability estimate does not provide any more information than the priors, as each specification is correct the same number of times. Only posterior probability estimates that differ from 0.5 yield additional information and offer an opportunity for incorrect predictions. The information content in each criterion's posterior probabilities was evaluated by summing the log of each criterion's posterior probability estimates:

$$X = \sum_{i=1}^{900} L_i . \qquad (21)$$

The higher the value of X, the more confident were the posterior probabilities. The sum of the logs of the priors is equal to −623.8, which provides a lower bound for this measure.

Schwarz, Akaike, CVL(L^*), and PSRQB(Split) provided the most accurate posterior probabilities, with no significant biases and around 70% of the variance in mean actual success explained by the posterior probability estimates. Schwarz and Akaike also provided impressive information content, while CVL(L^*) and PSRQB(Split) contributed the least information of all seven of the criteria.

The lack of information content in the PSRQB(Split) estimates offers an explanation for why there was no

Table 3 Posterior Probability Accuracy: Comparison of Criteria

Criteria	Results of Regressing *Mean Posterior Probabilities* on *Mean Actual Success*: Constant (α)	Coefficient (β)	R^2	Sum of Logs* X
Schwarz	−0.0339	1.0268	0.75	−84.2
	(0.0749)	(0.0896)		
Akaike	0.0355	0.9436	0.70	−97.1
	(0.0777)	(0.0952)		
CVL(L^*)	0.0494	0.8495	0.64	−112.0
	(0.0859)	(0.0971)		
PSRQB(Split)	0.0017	1.0033	0.70	−159.4
	(0.0700)	(0.0998)		
Allenby	0.1843	0.7392[2]	0.48	−107.4
	(0.0945)	(0.1173)		
CVL(Split)	0.2280[1]	0.4981[2]	0.43	−112.0
	(0.0701)	(0.0881)		
PRSQB(L^*)	0.4379[1]	0.4000[2]	0.10	−78.6
	(0.1571)	(0.1844)		

The sample size for each regression was 45.

The terms in brackets under each parameter are standard errors.

* See equation (21).

[1] Significantly different from zero (α = 0.05).

[2] Significantly different from one (α = 0.05).

[6] For each of 15 simulations, 15 data sets, and four simulation scenarios.

[7] Given that there are only 2 alternative model specifications, in each case the preferred specification had a minimum posterior probability of 0.5.

significant bias in the posterior probabilities from this split sample approach while the jackknifed version's posterior probabilities were significantly biased. The jackknifing procedure on PSRQB resulted in posterior probabilities that were more confident than any of the other six criteria, while the estimates from the split sample approach were the least confident. When the posterior probabilities do not differ from the priors, the posteriors provide no additional information and afford no opportunity to introduce bias. Therefore, biases will be more apparent when the posterior probabilities are more confident.

The posterior probability estimates for Allenby and CVL(Split) were also systematically biased and were poor predictors of actual success. The β parameter estimates (significantly less than one) indicate that the posterior probability predictions are overly optimistic, while the R^2 reported for these three criteria are noticeably smaller than those reported for the other criteria.

3.4. Time Requirements and Ease of Use

The Schwarz, Akaike, and Allenby criteria require only the log maximum likelihood, plus the simple calculation of a penalty term. Their calculation is virtually instantaneous (about 0.1 seconds CPU time for each—all runs performed on a 486 PC), and the criteria are very easy to implement. The CVL procedures require several steps. First, the sample must be split (either once or many times). Model coefficients must be estimated for each sample split. Finally, the likelihood must be calculated for each validation sample. These criteria are considerably more trouble than the Akaike, Schwarz, and Allenby approaches. While CPU time is not a major cost, it does offer an approximation of the relative complexity of the different criteria. CVL(Split) required an average of about 0.2 seconds CPU time, and CVL(L^*) required an average of 4.1 seconds.

Most complicated and difficult to use are the PSRQB approaches. They require all the effort of the CVL approaches, plus the necessity of numerical integration, in most instances. Often the numerical integration will need to be accomplished in a high-dimensional space, which can be computationally forbidding. Even when, as in our simulations, closed-form solutions exist, these criteria are relatively slow, although not infeasible timewise. PSRQB(Split) required an average of 0.3 seconds of CPU time, and PSRQB(L^*) an average of 9.0 seconds.

3.5. Caveats

We must qualify all our performance results with the observation that we have investigated only a small fraction of the model comparison scenarios which are possible. Thus, it is always possible that (for reasons unforeseen) there may exist conditions under which criteria which performed poorly in our study may instead perform very well. There is considerable room for further empirical and theoretical study to determine the conditions under which particular criteria may show a differential advantage.

4. How Many Criteria Should Be Used?

One might wonder whether combining two or more criteria might produce better results than using any single criterion. To investigate what advantages can be gained from combining model selection criteria, we propose a method for combining criteria and explore whether combinations of criteria are more accurate at identifying the correct model specification.

4.1. A Method for Combining Criteria

In this subsection we propose a method for combining model selection criteria. Our situation differs from the usual combination of forecasts scenario because selection of a model is categorical, while a forecast is quantitatively scaled. Let PL_{jm} denote the posterior likelihood of model j according to criterion m, as calculated in subsection 2.4, except normalized to sum to one across models. Assume that there exists a linear combination

$$X_j = \sum_m \beta_m PL_{jm}, \tag{22}$$

for which choosing the model with the highest X_j will maximize the likelihood of choosing the correct model. Specifically, assume that

$$Y_j = X_j + \epsilon, \tag{23}$$

where Y_j is a true model indicator, and ϵ is an extreme value distributed error term which reflects that the linear combination is fallible in selecting the true model. We assume that the model j for which Y_j is largest is the true model. These are the standard logit assumptions. By calibrating the model (estimating the β coefficients) and testing its predictive performance, we should be

Table 4 Estimation and Prediction Results: Combinations of Criteria

Criteria	Which Criteria Were in Each Optimal Combination? Size of Optimal Combination						
Schwarz	✓	✓	✓	✓	✓	✓	✓
PSRQB(L^*)		✓	✓		✓	✓	✓
CVL(Split)			✓		✓	✓	✓
CVL(L^*)				✓	✓	✓	✓
Allenby				✓	✓	✓	✓
PSRQB(Split)						✓	✓
Akaike							✓
Constant		✓	✓	✓	✓	✓	✓

✓ Indicates that this criteria was included in the relevant optimal combination.

	Incremental Accuracy From Combining Criteria						
−2 Log Likelihood		340.31	340.07	322.12	321.58	316.12	296.65
%Correct (Est)	84.8	86.9	87.2	86.9	86.7	86.3	85.2
%Correct (Val.)	78.1	80.0	80.6	81.4	81.4	82.5	80.0

able to ascertain the benefits of combining several comparison criteria.

4.2. Research Design

For the four simulation scenarios (static correct, dynamic correct, linear correct, log correct) we randomly selected nine of the fifteen data sets to serve as an estimation sample, and reserved the remaining six data sets for validative testing. We constructed a dependent variable, which was one when the model was correct and zero otherwise. To create variance in the dependent variable, we considered the wrong model in four randomly chosen data sets of the nine estimation data sets, and three randomly chosen data sets of the six validation data sets.[8]

We then estimated a logit model with model correctness (one or zero) as dependent variable and PL_{jm}'s as independent variables. The wrong model PL_{jm} was used if the dependent variable was zero, while the correct model was used if the dependent variable was one. Using the estimation sample and considering *all possible* combinations, we first found the best pair of criteria, then the best triple, quadruple, quintuple, and sextuple. We also found the best individual criterion, and the coefficients for the model which included all seven criteria. We then tested the predictive ability of the combinations on the validation sample, and recorded the proportion of the time the correct model was chosen, for each combination of criteria.

4.3. Results

Table 4 shows the accuracy of the combinations of criteria on both the estimation and the validation sample. The Schwarz criterion was the single best selection criterion (for the estimation sample), choosing the correct model 84.8% of the time. It is striking that, even in the estimation sample, there is very little improvement with the number of criteria. In fact, the correct selection percentage actually declines somewhat, even though the log likelihood continues to improve. This is possible because the combinations can become "more sure" of which model is correct (increasing the posterior probabilities).

The relative performance of the combinations of criteria could be tested by obtaining posterior likelihoods of the different numbers of criteria, assuming that one of the combinations is the true model and that they have equal prior probabilities. The posterior likelihoods will then be proportional to the proportion of correct selections. However, it is obvious that this Bayesian test will not differentiate between the combination of criteria very much, since the performance of the worst combination (78.1%) is still about 95% of the performance of the best combination (82.5%).

[8] Otherwise, there would be no variance in the dependent variable.

An alternative test may be conducted on the estimation sample. Some of the combinations are nested within others, permitting a likelihood ratio chi-square test. By this test, using all seven criteria results in a −2 log likelihood of 296.647, compared to 381.578 for one criterion. The chi-square value is thus 84.931, which exceeds the 0.05 critical value of 12.592. Similar tests show the 7-combination's superiority over all the others. This gives some evidence that using a combination of criteria may be superior. However, given our large sample size, statistical significance is not surprising. Given a large enough sample size, *any* unique information added by a criterion will produce significance. It is clear that the *practical* gains from employing multiple criteria may be small. For example, using all seven criteria, as would be indicated by the likelihood ratio tests, improved prediction in the validation sample only from 78.1% to 80.0% (versus using only the Schwarz criterion).

5. Conclusions

We conclude that the use of more than one model selection criterion may be unwarranted. This is apparently due to the fact that all good model selection criteria are highly intercorrelated. Thus, following the logic of Clemen and Winkler (1985), very little additional information is imparted, and even a very large number of criteria might not add very much in selection accuracy.

Given that it is best to use only one criterion, which criterion should it be? Based on our analyses, as well as those of several previous studies, we conclude that the Schwarz and Akaike criteria both do a good job of selecting the best model. Of the two, the Schwarz criterion appears to be the most consistently accurate. The estimated posterior probabilities from the Schwarz criterion were also the most accurate predictions of the criterion's actual accuracy. Thus, by using the Schwarz criterion, we may accurately select the correct model, and at the same time generate posterior probabilities of model correctness, conditional on the correct model being in the consideration set. What's more, the necessary calculations to implement this criterion are very simple and can be performed in seconds on a hand calculator, given the log maximum likelihoods of the competing models.

In summary, we conclude that the best available general approach to selecting quantitative models is also one which is startingly easy to implement. Using the Schwarz criterion is likely to result in accurate model selections and an accurate evaluation of the relative likelihood that the selected model is best.[9]

[9] The authors appreciate the helpful comments of John D. C. Little and John R. Hauser, and the statistical assistance of Bruce Cooil.

References

Akaike, H., "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19 (1974), 716–723.

Allenby, G. M., "Cross-Validation, The Bayes Theorem, and Small-Sample Bias," *Business and Economic Statistics*, 8 (April, 1990), 171–178.

Bass, F. M., "Misspecification and the Inherent Randomness of the Model Are at the Heart of the Brodie and DeKluyver Enigma," *International Forecasting*, 3 (1987), 441–444.

Bates, J. M. and C. W. J. Granger, "The Combination of Forecasts," *Oper. Res. Quarterly*, 20 (December, 1969), 451–468.

Brodie, R. and C. A. DeKluyver, "Attraction Versus Linear and Multiplicative Market Share Models: An Empirical Evaluation," *J. Marketing Res.*, 21 (May, 1984), 194–201.

Chen, L. and G. Anandalingam, "Optimal Selection of Forecasts," *J. Forecasting*, 9 (1990), 283–297.

Clayton, M. K., S. Geisser, and D. E. Jennings, "A Comparison of Several Model Selection Procedures," in A. Zellner and J. B. Kadane (Eds.) *Bayesian Inference and Decision Techniques*, North Holland, Amsterdam, 1986.

Clemen, R. T., "Combining Forecasts: A Review and Annotated Bibliography," *International J. Forecasting*, 5 (1989), 559–583.

—— and R. L. Winkler, "Limits for the Precision and Value of Information from Dependent Sources," *Oper. Res.*, 33 (March-April, 1985), 427–442.

Cooil, B., R. Winer, and D. Rados, "Cross-Validation for Prediction," *J. Marketing*, 24 (August, 1987), 271–279.

Dickey, J. M., "Matricvariate Generalizations of the Multivariate *t* Distribution and the Inverted Multivariate *t* Distribution," *Annals of Mathematical Statistics*, 38 (April, 1967), 511–518.

Diebold, F. X., "Forecast Combination and Encompassing: Reconciling Two Divergent Literatures," *International J. Forecasting*, 5 (1989), 589–592.

Fornell, C. and R. T. Rust, "Incorporating Prior Theory in Covariance Structure Analysis: A Bayesian Approach," *Psychometrika*, 54 (June, 1989), 249–259.

Geisser, S. and W. F. Eddy, "A Predictive Approach to Model Selection," *J. American Statistical Association*, 74 (1979), 153–160.

Homburg, C., "Cross-Validation and Information Criteria in Causal Modeling," *J. Marketing Res.*, 28 (May, 1991), 137–144.

Makridakis, S. and R. Winkler, "Averages of Forecasts: Some Empirical Results," *Management Sci.*, 29 (1983), 987–996.

Morrison, D. G. and D. C. Schmittlein, "How Many Forecasters Do You Really Have?: Mahalanobis Provides the Intuition for the Surprising Clemens and Winkler Result," *Oper. Res.*, 39 (May-June 1991), 519–523.

Rust, R. T. and D. C. Schmittlein, "A Bayesian Cross-Validation Likelihood Method for Comparing Alternative Specifications of Quantitative Models," *Marketing Sci.*, 4 (Winter, 1985), 20–40.

Schwarz, G., "Estimating the Dimension of a Model," *Annals of Statistics*, 6 (1978), 461–464.

Stone, M., "Cross-Validatory Choice and Assessment of Statistical Predictions," *J. Royal Statistical Society B*, 36 (1974), 111–147.

Stone, M., "On Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *J. Royal Statistical Society B*, 39 (1977), 44–47.

Winkler, R. L. and S. Makridakis, "The Combination of Forecasts," *J. Royal Statistical Society A*, 146 (1983), 150–157.

Zellner, A., *An Introduction to Bayesian Inference in Econometrics*, John Wiley, New York, 1971.

Accepted by Gabriel R. Bitran; received June 20, 1992. *This paper has been with the authors* 4 *months for* 1 *revision.*

[Footnotes]

[5] **Attraction versus Linear and Multiplicative Market Share Models: An Empirical Evaluation**

Roderick Brodie; Cornelis A. de Kluyver

Journal of Marketing Research, Vol. 21, No. 2. (May, 1984), pp. 194-201.

Stable URL:

http://links.jstor.org/sici?sici=0022-2437%28198405%2921%3A2%3C194%3AAVLAMM%3E2.0.CO%3B2-J

References

Attraction versus Linear and Multiplicative Market Share Models: An Empirical Evaluation

Roderick Brodie; Cornelis A. de Kluyver

Journal of Marketing Research, Vol. 21, No. 2. (May, 1984), pp. 194-201.

Stable URL:

http://links.jstor.org/sici?sici=0022-2437%28198405%2921%3A2%3C194%3AAVLAMM%3E2.0.CO%3B2-J

Matricvariate Generalizations of the Multivariate t Distribution and the Inverted Multivariate t Distribution

James M. Dickey

The Annals of Mathematical Statistics, Vol. 38, No. 2. (Apr., 1967), pp. 511-518.

Stable URL:

http://links.jstor.org/sici?sici=0003-4851%28196704%2938%3A2%3C511%3AMGOTMD%3E2.0.CO%3B2-H

LINKED CITATIONS

Cross-Validation and Information Criteria in Causal Modeling
Christian Homburg
Journal of Marketing Research, Vol. 28, No. 2. (May, 1991), pp. 137-144.
Stable URL:
http://links.jstor.org/sici?sici=0022-2437%28199105%2928%3A2%3C137%3ACAICIC%3E2.0.CO%3B2-G

A Bayesian Cross-Validated Likelihood Method for Comparing Alternative Specifications of Quantitative Models
Roland T. Rust; David C. Schmittlein
Marketing Science, Vol. 4, No. 1. (Winter, 1985), pp. 20-40.
Stable URL:
http://links.jstor.org/sici?sici=0732-2399%28198524%294%3A1%3C20%3AABCLMF%3E2.0.CO%3B2-C

Estimating the Dimension of a Model
Gideon Schwarz
The Annals of Statistics, Vol. 6, No. 2. (Mar., 1978), pp. 461-464.
Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28197803%296%3A2%3C461%3AETDOAM%3E2.0.CO%3B2-5