*Please do not quote. This research is still in progress.*

# A COMPARISON OF A LARGE NUMBER OF MODEL SELECTION CRITERIA

by

Xiaochuan Qin and W. Robert Reed*

## Abstract

This paper uses Monte Carlo analysis to compare the performance of a large number of model selection criteria (MSC), including information criteria, General-to-Specific modelling, Bayesian Model Averaging, and portfolio models. We use Mean Squared Error (MSE) as our measure of MSC performance. The decomposition of MSE into Bias and Variance provides a useful decomposition for understanding MSC performance. We find that the overall best MSC is the "small-sample corrected version" of the Schwarz Information Criterion. However, there is much variation in MSC performance and no MSC works best in all circumstances. We identify two important determinants of MSC performance: (i) overall fit of the equation as measured by $R^2$, and (ii) the ratio of relevant variables to total candidate variables. We show that these two factors matter because they relate to MSC "overfitting" and "underfitting." We proceed to identify circumstances in which some MSC are likely to dominate others.

JEL Categories: C52, C15

Keywords: Model Selection Criteria, Information Criteria, General-to-Specific modeling, Bayesian Model Averaging, Portfolio Models, AIC, SIC, AICc, SICc, Monte Carlo Analysis

May 26, 2008

*This paper was prepared for the ESAM08 Conference, "Markets and Models: Policy Frontiers in the AWH Phillips Tradition",* Wellington, New Zealand, July 9-11, 2008.

*The contact author is W. Robert Reed, Professor, Department of Economics, University of Canterbury, Private Bag 4800, Christchurch 8020, New Zealand; email: bobreednz@yahoo.com.

> "In those days there was no king in Israel; everyone did what was right in his own eyes." -- Judges 21:24

## I. INTRODUCTION

This study examines a problem that faces many researchers: How should one pick the "best" regression equation, or a set of "best" regression equations? Without some objective standard, efforts to select a "best" regression equation may, at best, innocently miss superior specifications; or, at worst, strategically select results to support the researcher's preconceived biases.

A substantial literature has grown that demonstrates that model selection matters. For example, many studies of economic growth find that results that are economically and statistically significant in one study are not robust to alternative specifications (cf. Levine and Renelt, 1992; Fernandez et al., 2001; Sala-i-Martin et al., 2004; Hoover and Perez; 2004). For these and related reasons, a literature has grown up that addresses the question of how to choose the best model(s).

A non-exhaustive list of the associated proposals include choosing a single best model based upon an information criterion such as the Akaike Information Criterion (AIC) or the Schwarz Information Criterion (SIC) (cf. McQuarrie and Tsai, 1998); following a General-to-Specific (GETS) algorithm of eliminating insignificant variables (cf. Hoover and Perez, 1999; Hendry and Krolzig, 2005); selecting a "portfolio" or best subset of models (cf. Poskitt and Tremayne, 1985); and combining models using Bayesian Model Averaging (cf. Hoeting et al., 1999; Sala-i-Martin, 2004).

Not only is there a variety of proposed MSC, but there is also a variety of measures to determine "best" MSC performance. A non-exhaustive list of performance

measures include counts of the number of times the MSC correctly picks the true DGP, number of times the MSC "overfits' (selects too many variables) or "underfits" (selects too few variables) (cf. McQuarrie and Tsai, 1999); whether the size of the statistical tests on the "best" model conform to the nominal significance levels (Hendry and Krolzig, 2005); and whether the MSC selects the model with greatest predictive efficiency (Kuha, 2004; Burnham and Anderson, 2004).

This study makes the following contributions to this literature: We empirically compare a larger number of MSC's than previous studies, including MSC's based on information criteria, General-to-Specific modelling, Bayesian Model Averaging, and portfolio models. We propose a different measure of estimator performance – MSE of estimated coefficients – and demonstrate its usefulness in explaining the relative performances of competing MSC. Our Monte Carlo analyses find that the best overall MSC is the "small-sample corrected" version of the SIC (cf. McQuarrie, 1999). However, we also find that no one MSC works best in all – or even most – circumstances.

It is well-known that MSC differ with respect to "overfitting" and "underfitting". Our simulations identify two important determinants of MSC performance: (i) overall fit of the equation as measured by $R^2$, and (ii) the ratio of relevant variables to total candidate variables. We show that these two factors matter because they relate to MSC "overfitting" and "underfitting," which in turn relate to the Bias and Variance components of MSE. We proceed to identify circumstances in which some MSC are likely to dominate others.

## II. A FRAMEWORK FOR COMPARING MODEL SELECTION CRITERIA

The Problem. We investigate the following problem. We have a data set consisting of $N$ observations on variables $Y, X_1, X_2, \ldots, X_L$. We assume that the data generating process (DGP) producing these observations is given by:

(1) $$Y_n = \alpha + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots \beta_L X_{Ln} + \varepsilon_n, \; n = 1,2,\ldots,N,$$

where $K$ of the $\beta$'s are nonzero and $L$-$K$ are zero, $1 \le K \le L$; and the $\varepsilon_n$ are i.i.d., with $\varepsilon_n \sim N\left(0,\sigma_\varepsilon^2\right)$. We want to choose the "best" MSC, where "best" is defined as the MSC that results in the most accurate estimates of the $\beta$'s. We define this more precisely below.

The Model Selection Criteria (MSC). We study 15 different MSC drawn from a variety of approaches. These are listed in TABLE 1, along with a brief description. The first four are based on information criteria (IC). While there are many information criteria, most of these are asymptotically related to either the Akaike Information Criterion (*AIC*) or the Schwarz Information Criterion (*SIC*) (Weakliem, 2004). Both the *AIC* and the *SIC* have the same general form: $-2\ell + Penalty$, where $\ell$ is the maximized value of the log-likelihood function for the given specification, and *Penalty* is a function that monotonically increases in the number of coefficients to be estimated. In both cases, smaller is better, and the specification with the smallest *AIC/SIC* value is considered to be "best." The *SIC* generally penalizes the inclusion of parameters more harshly than the *AIC*, and thus favors more parsimonious models.

The *AIC* and *SIC* have asymptotic justification. The *SIC* is consistent. That is, if the true DGP is included among the set of candidate models, the *SIC* will select the true DGP with probability approaching one as the sample size increases. The *AIC* is

asymptotically efficient. It assumes that the true DGP is not included in the set of candidate models. It selects the model having the smallest expected prediction error with probability approaching one as the sample size increases (Kuha, 2004).

It is well-known that both the *AIC* and *SIC* tend to "overfit" (i.e., include more variables than the DGP) in small samples. As a result, small-sample corrections for these have been developed by Hurvich and Tsai (1989) and McQuarrie (1999), respectively. These are denoted in TABLE 1 as *AICC* and *SICC*, where the last "C" denotes that it is the "corrected" version of the respective information criterion.

For each of these four MSC, IC values are calculated for all $2^L$ possible models. Coefficient estimates are taken from the model with the lowest IC value. If a variable does not appear in that model, then the associated estimate of that coefficient is set equal to zero.

The fifth MSC uses General-to-Specific (*GETS*) modeling. At its simplest, *GETS* starts with a fully-parameterized model and sequentially deletes insignificant variables. Modern versions of *GETS* emphasize multiple reduction paths, which greatly reduce path dependence with little abuse of nominal significance levels (Hoover and Perez, 1999; Campos et al. 2003; Hendry and Krolzig, 2005). Our version of *GETS* employs a 5% significance level and sequentially eliminates the variable with the smallest *t*-statistic until all insignificant coefficients are removed. As noted by Campos et al. (2003), this procedure should produce results very close to the more sophisticated, multi-path versions of *GETS* when the explanatory variables are orthogonal, as will be the case in our Monte Carlo experiments. Individual coefficient estimates are taken from the final

model. If a variable does not appear in that model, then the associated estimate of that coefficient is set equal to zero.

The next eight MSC are based on the idea of selecting – not a single "best" model – but a "portfolio" of models that are all "close" as measured by their information criterion (IC) values. Poskitt and Tremayne (1987) derive a measure based on the posterior odds ratio, $\Re_m = exp\left[-\frac{1}{2}\left(IC_{min} - IC_m\right)\right]$, where $IC_{min}$ is the minimum *IC* value among all $2^L$ models, and $IC_m$ is the value of the respective *IC* in model *m*, $m=1,2,\ldots,2^L$. They suggest forming a portfolio of models all having $\Re_m \leq \sqrt{10}$. Alternatively, Burnham and Anderson (2004) suggest a threshold $\Re_m$ value of *2*. We use both values. The MSC *AIC < 2*, *AICC < 2, SIC < 2,* and *SICC < 2* each construct portfolios of models that have *AIC*, *AICC*, *SIC*, and *SICC* values that lie within 2 of the minimum value model. The next four MSC ($AIC < \sqrt{10}$, $AICC < \sqrt{10}$, $SIC < \sqrt{10}$, and $SICC < \sqrt{10}$) do the same for models lying within $\sqrt{10}$ of the respective minimum value model.

In our experiments, coefficient estimates are set equal to zero for all variables that never appear in the portfolio. For variables that appear at least once in the portfolio of models, we calculate the respective coefficient estimates as the arithmetic average of all nonzero coefficient estimates. The reason for not including zero values will be discussed below.

The last two MSC are examples of Bayesian Model Averaging (Hoeting, Madigan, Raftery, and Volinsky, 1999). In Bayesian Model Averaging, a composite model is constructed by taking a weighted average of a set of models, which might consist of all possible models, with weights consisting of the posterior model

probabilities. In the composite model, each of the variable coefficients equals the weighted average of the individual estimated coefficients for that variable. We calculate model weights using the maximized value of the log-likelihood function. The MSC *LL_Weighted* uses the full set of $2^L$ models to construct weighted average, coefficient estimates. The MSC *LL_Weighted*( $\hat{\beta}_k \neq 0$ ) restricts itself to the set of all $2^{L-1}$ models where the given variable is included in the model.

Monte Carlo Experiments and the Performance Measure. Our experiments all use the DGP, $Y_n = \alpha + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots \beta_L X_{Ln} + \varepsilon_n$, $n = 1,2,...,N$, where $\alpha = 5$, $\beta_1 = \beta_2 = \cdots = \beta_K = 1$, $\beta_{K+1} = \beta_{K+2} = \cdots = \beta_L = 0$, $1 \le K \le L$. The $\varepsilon_n$ are i.i.d. and normally distributed with mean *0* and variance a function of *K*, as will shortly be described. Each experiment has *K* "relevant" variables and *L-K* "irrelevant" variables, with relevancy is defined according to whether that variable has a nonzero coefficient in the DGP. We simulate 1000 data sets for each experiment.

For given *L*, we run *L* consecutive experiments where *K* starts at *1* and progresses through *L*. The individual *X* realizations are distributed i.i.d. both within and across variables, such that $X_{kn} \sim N\left(0, \sigma_X^2 = 1\right)$, $k = 1,2,...,L$, $n = 1,2,...,N$. The *X* variables are held constant throughout all *L* experiments, so that the only things that change in the DGP across experiments are the values of the $\beta_k$'s and the distribution of the error terms.

We alter the distribution of the $\varepsilon_n$'s across experiments because we are interested in studying how MSC performance changes as a function of the $R^2$ of the estimated equation. Ceteris paribus, as *K* increases, and more variables are "relevant," $R^2$ will increase. In response, we increase the variance of the error terms in an ad hoc fashion so

that the $R^2$ values remain relatively constant as $K$ increases.[1] For given $L$, we study four $R^2$ values: $R^2 = 90\%$, $R^2 = 70\%$, $R^2 = 50\%$, and $R^2 = 35\%$.

Thus, for each value of $L$, there are $L$ experiments for each $R^2$ level, producing a total of $4L$ experiments. We set $L=5$, $10$, and $15$, producing a total of 120 experiments. Finally, we set $N$ equal to 75 observations in all our experiments. This was done primarily for technical reasons: We encountered computing problems as $N$ increased, both in terms of the time it took to run our programs, and in constructing the weights for the Bayesian Model Averaging MSC.[2]

As our basis for measuring MSC performance we use Mean Squared Error (MSE). For each data set/replication $r$, and each MSC, we have a set of estimates, $\left(\hat{\beta}_{1,r}^{MSC}, \hat{\beta}_{2,r}^{MSC}, ..., \hat{\beta}_{L,r}^{MSC}\right)$. Let $\beta_k$ be the true value of the slope coefficient for $X_k$. For the 1000 replications of the experiment, we calculate a coefficient-specific MSE as follows:

$$(2) \qquad MSE_k^{MSC} = \frac{\sum_{r=1}^{1000}\left(\hat{\beta}_{k,r}^{MSC} - \beta_k\right)^2}{1000}, \; k = 1,2, \ldots, L.$$

Because MSE is not comparable across coefficients, we assign a coefficient-specific ranking from $1$ to $15$, with the MSC producing the lowest MSE for that coefficient receiving a rank of $1$, the MSC with the next smallest MSE receiving a rank of $2$, and so on. These rankings are then averaged across all $L$ coefficients to produce an overall MSC ranking for that experiment. For example, if $L = 5$ and a given MSC has individual coefficient rankings $\{10,10,12,13,10\}$, it would receive an average rank of $11$.

---

[1] We use the general formula, $\sigma_\varepsilon = aK^b$. The values of $a$ and $b$ were determined experimentally, with $b$ held constant for given $L$, and $a$ adjusted with $K$ so that the $R^2$ of the estimated equations remained approximately constant as $K$ increased.

[2] A typical program with $L=15$ and a given $R^2$ took 5-6 days to run on our laptops.

There are several advantages to using MSE as a measure of MSC performance. First, it coincides with a key goal of estimation: that of producing accurate coefficient estimates. For example, researchers examining the effects of various policy variables want accurate estimates of those effects. Further, it is important to obtain accurate estimates for both "relevant" and "irrelevant" variables.

Another advantage to using MSE is that is can be decomposed into (i) Bias and (ii) Variance components. Some of the MSC are weak on one dimension, but strong on the other, so that their relative performance depends on tradeoffs between Bias and Variance. This can provide insights as to the conditions under which particular MSC are likely to be effective.

For example, it is well-known that model-averaging over all possible ($2^L$) models produces biased coefficient estimates for relevant variables (Hendry and Krolzig, 2005; Reed, 2008). The logic is this: In each model in which the variable $X_k$ appears, the OLS estimate of the associated coefficient is unbiased, $E\left(\hat{\beta}_k = \beta_k\right)$. However, $X_k$ only appears in half of all possible models ($2^{L-1}$). In the other $2^{L-1}$ models, where $X_k$ is excluded, $\hat{\beta}_k$ is set equal to $0 \neq \beta_k$. It follows that the corresponding weighted average over all possible models will be biased. This, by the way, is the motivation for why some of the MSC do not include *0* values when averaging coefficient estimates across models. On the other hand, using an average of estimated coefficients, rather than a single coefficient, reduces Variance. Thus it is possible for an MSC that produces biased coefficient estimates to perform better on MSE than an MSC that produces unbiased estimates.

A related example concerns MSC that underfit. If $\beta_k \neq 0$, an MSC that tends to underfit will set $\hat{\beta}_k = 0$ every time it does not include variable $X_k$ in its selected model. While these estimates are biased, they could reduce the Variance. In the extreme, if the MSC sets $\hat{\beta}_k = 0$ for every replication, Variance($\hat{\beta}_k$) = 0. The net effect could be a lower MSE than produced by an MSC that was more successful in selecting the true DGP.

Finally, we include a second performance measure, Mean Absolute Deviation (MAD).

$$(3) \qquad MAD_k^{MSC} = \frac{\sum_{r=1}^{1000} \left| \hat{\beta}_{k,r}^{MSC} - \beta_k \right|}{1000}, \; k = 1,2, \ldots, L.$$

MAD also measures coefficient accuracy, but without the convenient decomposition between Bias and Variance. We include it as a consistency check on our MSE results. As with MSE, we calculate an experiment-specific, average MAD rank across all $L$ coefficients to measure overall MSC performance.

## III. RESULTS

TABLE 2 summarizes the results from the 120 experiments. The first three columns use Mean Squared Error (MSE) to measure MSC performance. The next three use Mean Absolute Deviation (MAD). There is little difference between the two. This result holds not only in the aggregate, but also at the level of individual experiments. As a result, we focus the subsequent discussion on MSE.

The numbers in the table report average rankings. A smaller number represents a higher ranking, with 1 being the best. The individual unit of observation is the

experiment. For example, the mean MSE ranking for the *AIC* MSC over all 120 experiments is 7.22. The highest ranking achieved by this MSC in any one experiment is 2.60 (for the experiment $L=5, K=4, R^2=70\%$). This number is itself an average rank over the 5 coefficients in that experiment. The lowest ranking achieved by this MSC is 13.67 (for the experiment $L=15, K=15, R^2=35\%$).

In terms of overall performance, the top six MSC in descending order are:

1. *SICC* (5.91)
2. *SIC* (6.24)
3. *GETS* (6.44)
4. *LL_Weighted* (6.53)
5. *AICC* (6.81)
6. *AIC* (7.22)

However, these rankings disguise substantial variation in MSC performance. For example, *SICC*'s minimum rank is 1.00, achieved in 20 of the 120 experiments. The interpretation of this number is that *SICC* had the lowest MSE value for every coefficient in each of these 20 experiments. On the other hand, *SICC*'s maximum rank is 15.00, achieved in 5 experiments. It had the highest MSE for every coefficient in these experiments.

In terms of overall mean rankings, the portfolio MSC all perform worse than their non-portfolio analogs. For example, the mean ranks for $AIC < 2$ and $AIC < \sqrt{10}$ (8.68 and 9.10, respectively) are worse than for *AIC* (7.22). We also find that model averaging over all possible models (*LL_Weighted*) is generally superior to model averaging over only those models in which the respective variable appears (*LL_Weighted*$(\hat{\beta}_k \neq 0)$), even though the former produces biased coefficient estimates. That being said, there are

scenarios where portfolio MSC and *LL_Weighted(* $\hat{\beta}_k \neq 0$ *)* do better. We discuss these further below. In the meantime, we focus discussion on the top 6 MSC identified above.

FIGURE 1 plots the rankings of these 6 MSC as a function of *K,* $R^2$*,* and *L*; where *L* represents the total number of variables available to the researcher, *K* represents the number of variables in the DGP, and $R^2$ (approximately) measures the average $R^2$ value from estimating a fully specified model in a given experiment. The individual data points consist of the same, experiment-specific, average rankings summarized in TABLE 2.

There are a total of 12 graphs in the figure. Each graph shows how MSC ranking changes with *K*, holding *L* and $R^2$ constant. Moving from top to bottom in a given column shows the effect of $R^2$ decreasing from 90% to 70% to 50% to 35%, holding *L* constant. Moving from left to right in a given row shows the effect of *L* increasing from 5 to 10 to 15, holding $R^2$ constant.

The factors driving the changing, relative performances of the MSC are best illustrated through a closer look at the individual experiments. Consider the case $R^2$ = *70%* and *L=5* in FIGURE 1. Note that *SICC* goes from best MSC to worst MSC as the number of relevant variables (*K*) increases from 1 to 5. TABLE 3 allows one to study this result in greater detail.

The table is divided into 4 panels, each summarizing the results of 1000 replications of an experiment whose DGP includes *K* relevant variables and *L-K* irrelevant variables (*K=2* to *K=5*). The numbers in the table represent coefficient-specific ranks for each of the MSC using the MSE values calculated from Equation (2). Since *L=5*, there are 5 rows for each panel/experiment. The first *K* rows of each panel

correspond to the relevant variables. The following *L-K* rows correspond to the irrelevant ones, with the solid line within each panel separating the two sets of coefficients.

For example, when *K=3*, the *AIC* rank for the first coefficient is 2. The interpretation is that the 1000 estimates of $\beta_1$ taken from the models selected by *AIC* in that experiment produced an MSE value that ranked 2nd lowest among all 15 MSC. The *AIC* rank for the second coefficient in that experiment is 4. Thus, the MSE calculated from the respective 1000 $\hat{\beta}_2$ values was 4th lowest among all MSC. Since $X_1$ and $X_2$ are both relevant variables, this difference in ranks must be due to sampling variability.

We now consider the top panel (*K=2*) of TABLE 3. The latter three rows of this panel report results for the variables that are excluded from the DGP (the "irrelevant" variables). When it comes to correctly estimating the coefficients of these variables, *SICC* always performs best. The average rank of *SICC* for the irrelevant variables equals 1.0 in every panel of TABLE 3. Further, this result is robust across all experiments, not just the ones reported in TABLE 3.

The reason for *SICC*'s top performance with irrelevant variables is due to its penalty function. *SICC* has the largest marginal cost for adding additional variables (followed by *SIC, AICC*, and *AIC*, in that order). It selects, on average, the fewest number of variables. Therefore, it is the MSC most likely to choose model specifications that correctly leave out irrelevant variables. All of the MSC produce unbiased coefficient estimates for the irrelevant variables. However, *SICC*-selected models have lowest variance, since omitted variables are assigned coefficient values of 0.

When *K* is small, *SICC* also produces the most accurate coefficient estimates for the relevant variables. Its average performance rank for relevant variables when *K=1*

(not shown) and *K=2* is 1.0. When *K=3*, it rises to 2.7, which is still lowest among the MSC. The explanation again is due to the variance component of MSE. When *K* is relatively small, *SICC*-selected models are the most likely to be correctly specified, and thus the most likely to produce accurate estimates. In contrast, the model specifications of other MSC's are more likely to include irrelevant variables (overfit). This reduces the precision of the estimated, relevant coefficients due to multicollinearity.

When the number of relevant variables increases further, *SICC*'s performance worsens relative to the other MSC. When *K=4*, *Average(Relevant)* = 7.0, higher than all other MSC's. When *K=5*, *Average(Relevant)* = 15.0, which means that *SICC* performs dead last among all MSC. The reason is twofold: As *K* increases, *SICC* becomes more likely to leave out relevant variables from the regression (underfit). This biases coefficient estimates of the relevant variables. Further, as $L - K$ gets smaller, the opportunity narrows for other MSC to include irrelevant variables (i.e., less overfitting). This makes their estimates of the relevant variables more precise. The combination of these two effects causes *SICC*'s relative performance to deteriorate quickly as *K* gets close to *L*. Indeed, this behaviour is evident in all 9 panels of FIGURE 1.

The changing, relative performance of *SICC* in TABLE 3 is illustrative of the factors that influence MSC performance. Underfitting biases coefficient estimates of relevant variables. Overfitting increases the variance of both relevant and irrelevant variables. The importance of these effects is weighted by the ratio of relevant to irrelevant variables (*K/L*) in the DGP and, as we shall see below, the "signal-to-noise" ratio as represented by the $R^2$ of the equation (cf. McQuarrie and Tsai, 1998). These factors combine to determine overall MSC performance.

Turning back to FIGURE 1, one observes some correspondence in the relative performance of the MSC within rows. The MSC rankings are related to the position of *K* relative to *L*. This is consistent with the performance effects we identify above. It further confirms *K/L* as an important parameter in MSC performance.

$R^2$ also matters. This can also be related to the effects identified above. As the error term grows larger, the respective MSC differ in their abilities to (i) correctly include and (ii) correctly exclude variables. As discussed above, this feeds into the bias and variance components of MSE to determine overall performance.[3]

TABLE 4 reports the results of a simple regression where MSC performance is regressed on *K/L* and $R^2$. The units of observation are the 120 experiments summarized in TABLE 2. Except for *LL_Weighted*, these two parameters explain about half of the overall variation in MSC rankings. $R^2$ is significant in every regression. *K/L* is significant in every regression except the *LL_Weighted* equation.

Based on the preceding experimental results, we make the following observations:

1. When *K/L* is relatively small (i.e., when $K/L \leq 0.40$), *SICC* is almost always the best MSC. Forty-eight of the 120 experiments fit these criteria, and *SICC* performs better than the other MSC in all but three of these experiments.

2. When $R^2 = 90\%$, *SICC* outperforms all other MSC except when *K* is very close to *L* (i.e., when $K/L \leq 0.90$). Twenty-six of the 120 experiments fit these criteria, and SICC performs better than the other MSCs in every one of these experiments.

3. The only other MSC which is consistently superior for given ranges of *K/L* and $R^2$ values is *LL_Weighted*. *LL_Weighted* tends to do well when $R^2 < 90\%$ and *K/L* is relatively large. However, the performance of *LL_Weighted* never dominates to

[3] The reader might notice that MSC performance is less consistent for the smallest $R^2$ cases (bottom row of FIGURE 1). This may be explained as follows: As discussed above, as we increase *K*, we also increase the variance of the error term in the DGP in order to keep $R^2$ relatively constant. However, we were unable to push $R^2$ much below 35%. Increasing the variance of the error term beyond a certain point had virtually no effect on $R^2$. Practically, that meant that there was not a one-to-one mapping between the variance of the error term and $R^2$, which makes conformity within rows more tenuous. This is not a problem as long as $R^2$ is above its lower threshold.

the degree that *SICC* does. For example, there are 48 experiments where $R^2 < 90\%$ and $K/L > 0.50$. *LL_Weighted* does best in about three-fourths of these.

Results for Portfolio Models. As discussed above, several authors argue that it is better to select a set of "good" models, as opposed to a single, best model (Poskitt and Tremayne, 1987; Kuha, 2004; Burnham and Anderson, 2004). An argument in favor of this portfolio approach is that the MSC are themselves random variables. "Single best" MSC will select inferior model specifications as a result of sampling variability. Portfolio MSC have a greater probability of selecting the true DGP as additional models are selected.

There is an additional argument in favor of portfolio MSC that relates to the use of MSE as a measure of MSC performance. Averaging coefficient estimates can reduce variance.[4] If the additional coefficient estimates are unbiased, a reduction in MSE is possible. We average over nonzero coefficient estimates only. If the variable is relevant, and a given model in the portfolio does not include this variable, then including the associated zero coefficient will bias the average, possibly mitigating – and even reversing – the advantage of averaging. The "less extreme bounds analysis" employed by Reed (2008) relies on the same motivation for identifying "robust" variables.

TABLE 2 makes clear that this approach does not, in general, result in improved MSC performance. All of the portfolio models are dominated in mean overall performance by their non-portfolio counterparts. However, when $R^2$ is low, and $K/L$ is relatively high, portfolio models can consistently outperform "single best" models.

TABLE 5 illustrates the main issues using the $L=5$ experiments for the *AIC* MSC. The left hand side of the table reports coefficient-specific performance for each of the

[4] This assumes that the additional coefficient estimates do not have larger variances.

respective MSC when $R^2$ is high. The right hand side does the same when $R^2$ is low. When $R^2 = 90\%$, ("single best") *AIC* does a relatively good job of selecting the correct DGP. The additional models included by the portfolios are inferior/misspecified models. As a result, the portfolio *AIC*'s generally do a worse job estimating coefficients for both relevant and irrelevant variables.

In contrast, when $R^2$ is low ($R^2 = 35\%$), ("single best") *AIC* tends to do a relatively poor job of including the correct variables. When the correct variable is not included in the best *AIC* specification, its coefficient is estimated to be zero, contributing to the bias-component of MSE. In contrast, portfolio models have a greater likelihood of including the correct variable. Not only will each of these estimates be unbiased, but averaging multiple coefficient estimates should reduce the variance-component of MSE.

TABLE 5 provides evidence of this. With only one exception ($K=1$), $AIC<2$ and $AIC<\sqrt{10}$ have larger MSE's for the relevant variables than ("single best") *AIC* when $R^2=90\%$. In contrast, when $R^2=35\%$, $AIC<2$ and $AIC<\sqrt{10}$ perform better than *AIC* across all relevant variables, with no exceptions.

The advantage enjoyed by portfolio MSC on relevant variables when $R^2$ is low does not extend to irrelevant variables. The imposed restriction that portfolio MSC average only over non-zero coefficients inflates the variance-component of MSE whenever the single-best *AIC* would have chosen a model that did not include the irrelevant variable. This is illustrated in TABLE 5. A comparison of $AIC<2$ and $AIC<\sqrt{10}$ with *AIC* shows little difference in relative performance for the irrelevant variables between high and low $R^2$ experiments.

When $R^2$ is low and ($K/L$) increases, the advantage that portfolio models have on relevant variables causes their overall ranking to improve. This is evidenced by the three columns on the right-hand side of TABLE 5. When $K=1$, the average rank of the *AIC* MSC is 6.6, compared to 9.7 for the two portfolio models. As one moves down the panels, the relative rank of the *AIC* MSC gets larger, while those of the portfolio models get smaller. When $K=5$, *AIC* has an average rank of 11 compare to 5.1 for the portfolio models.

While the preceding discussion has focused on *AIC*, the arguments carry over to the other portfolio models. TABLE 6 reports that, over all 120 experiments, the portfolio models for the *AIC*, *AICC*, *SIC*, and *SICC* do better than their single-best counterparts about a third of the time. When the sample of total experiments is restricted to those where (i) $R^2$ is 35% or 50% and (ii) ($K/L$) > 0.50 this proportion rises to about three-fourths.

The last row shows that a portfolio model was top-ranked of all 15 MSC in approximately one out of 6 experiments. For the sample of experiments where (i) $R^2$ was 35% or 50% and (ii) ($K/L$) > 0.50, a portfolio model was best in approximately one out of 2 experiments.

To summarize, our results lead us to make the following observation:

4. Portfolio models are most likely to do better than their single-best counterparts when $R^2$ is relatively low and the proportion of relevant variables ($K/L$) is relatively high.

Results for Bayesian Model Averaging. Like the portfolio models discussed above, Bayesian Model Averaging (BMA) relies on averaging coefficient estimates

across a set of models. However, BMA uses a weighted average. Individual coefficient estimates are weighted by their posterior model probabilities.

In a recent paper, Sala-i-Martin et al. (2004) propose a procedure for obtaining an overall estimate of $\beta_k$ in which all possible models are weighted by their respective log-likelihood values. Given $L$ candidate variables, there are $2^L$ total variable combinations, and hence $2^L$ possible models. $2^{L-1}$ of these models contain $X_k$, and $2^{L-1}$ do not. Models that do not contain $X_k$ are assigned a coefficient estimate of zero. The $2^{L-1}$ nonzero coefficient estimates and the $2^{L-1}$ zero values are combined to produce a single, weighted average estimate of $\beta_{\mathrm{k}}$.

This procedure will produce a biased coefficient estimate for any variable whose true coefficient is nonzero ($\beta_k \neq 0$). Each of the $2^{L-1}$ models containing $X_k$ produces an unbiased estimate of $\beta_k$. Therefore, any linear combination of these with zero values biases the estimate of $\beta_k$ towards zero. This bias is mitigated to the extent that the models that (incorrectly) exclude $X_k$ receive small weights.

When it comes to MSE, there are both bias and variance considerations. For relevant variables, these two conflict. Adding the $2^{L-1}$ models that do not include $X_k$ reduces variance; both because more models are included in the average, and because these additional models assign the scalar 0 as their coefficient estimate. For relevant variables, the net effect on MSE of including the $2^{L-1}$ (incorrect) models that do not contain $X_k$ is ambiguous.

For irrelevant variables, there is no ambiguity. $E\left(\hat{\beta}_k\right) = 0$ for each of the $2^L$ models. As a result, there is no bias penalty from calculating a weighted average using

all possible models. At the same time, there is an unambiguous decrease in the variance, for both of the reasons identified above. As a result, MSE is unambiguously lower for irrelevant variables when the weighted average of $\beta_k$ is calculated using all $2^L$ models.

To summarize, we cannot predict how *LL_Weighted* and *LL_Weighted*($\hat{\beta} \neq 0$) will compare for relevant variables, but *LL_Weighted* should perform best for irrelevant variables. Panel A of TABLE 7 illustrates these effects for the experiment *L=10*, $R^2$*=50%,* and *K=4*. The first four variables have nonzero population coefficients in the DGP, and the remaining 6 do not. *LL_Weighted*($\hat{\beta} \neq 0$) outperforms *LL_Weighted* for each of the four relevant variables. This is an example of the bias advantage of *LL_Weighted*($\hat{\beta} \neq 0$) MSC outweighing its variance disadvantage. However, as predicted, *LL_Weighted* is the better MSC for the six irrelevant variables. Overall, the superior performance of *LL_Weighted* for the irrelevant variables is sufficient to produce a better overall ranking (6.7 versus 10.2).

Panel B of TABLE 7 continues the comparison of *LL_Weighted* and *LL_Weighted*($\hat{\beta} \neq 0$), but focuses solely on the relevant variables. When $R^2$ is large, and thus the variance of the estimated coefficients is relatively small, the bias effect dominates and *LL_Weighted*($\hat{\beta} \neq 0$) outperforms *LL_Weighted*. As $R^2$ decreases, the variance of $\hat{\beta}_k$ increases, and the performance of *LL_Weighted*($\hat{\beta} \neq 0$) deteriorates.

Finally, we consider the overall performance of these MSC. When performance is aggregated over both relevant and irrelevant variables, *LL_Weighted* does better than *LL_Weighted*($\hat{\beta} \neq 0$) in 92 of the 120 experiments. As suggested by Panel B,

*LL_Weighted*( $\hat{\beta} \neq 0$ ) performs best when $R^2$ is high and *K/L* is large. This leads us to our final observation:

5. Bayesian Model Averaging using all possible models generally does better than using only those models having nonzero estimated coefficients, except when both $R^2$ and *K/L* are relatively high.

## IV. CONCLUSION

This study compares the performance of a large number of model selection criteria (MSC). It is distinguished from previous studies in a number of ways. It includes a larger number of MSC than previous studies (15). It includes a wider variety of MSC. We examine (i) conventional information criteria such as the AIC and SIC, (ii) General-to-Specific modelling, (iii) portfolio modeling where a set of models are chosen rather than a "single best" model; and (iv) Bayesian model averaging where models are weighted by their posterior probabilities to produce a composite model.

An innovation of our study is that we use the Mean Squared Error of individual coefficients as our performance measure. This has the advantage of allowing us to classify MSC performance in terms of bias and variance. We show how this provides a useful framework for understanding the relative performances of MSC.

A further contribution of our study is that we relate MSC performance to two "observable" regression characteristics: (i) $R^2$ and (ii) the ratio of relevant to total number of candidate variables (*K/L*). While the latter cannot be directly observed, it can be indirectly approximated from regression results. In any case, we argue that this is an improvement over the methodology of Sala-i-Martin (2004) whose approach requires that the researcher assume the absolute number of relevant variables.

Our experimental results indicate that there is no best MSC for all $R^2$ and *K/L* values. However, for certain ranges of $R^2$ and *K/L* values, we find that some MSC perform consistently better than others. Our main experimental results are summarized as follows:

1. We find that the overall best model selection criterion is the small-sample corrected version of the Schwarz Information, *SICC* (McQuarrie, 1999). However, there is much variation in MSC performance and no model selection criterion works best in all circumstances.

2. *SICC* consistently outperforms other MSC when either $R^2$ is very high or *K/L* is relatively small.

3. Bayesian Model Averaging, where all possible models are weighted by their maximized log-likelihood values, tends to outperform other MSC when $R^2$ is relatively low and *K/L* is relatively large.

4. Portfolio models are usually dominated by their single-best counterparts; however, they can do better when $R^2$ is relatively low and the proportion of relevant variables (*K/L*) is relatively high

5. Bayesian Model Averaging using all possible models generally does better than using only those models having nonzero estimated coefficients, except when both $R^2$ and *K/L* are relatively high.

These results come with many caveats. First, our Monte Carlo analyses are based on small sample sizes ($N = 75$). We need to confirm whether these results hold as sample size increases. Second, our Monte Carlo analyses are based on a single set of coefficient values for relevant variables. We need to confirm the robustness of our results for alternative coefficient values. Third, our Monte Carlo analyses assume that the explanatory variables are orthogonal, and that the error term in the DGP is spherical. Fourth, our experimental design assumes that the true model is one of the candidate models. Fifth, alternative performance measures, such as reliability of hypothesis testing

as opposed to MSE of estimated coefficients, could result in different conclusions regarding MSC performance.

That being said, the main contribution of our study is that it highlights the value of information criteria in general – and the *SICC* in particular – as an effective model selection criterion. Recent research in economics has emphasized the utility of General-to-Specific modeling and Bayesian Model Averaging (Hendry and Krolzig, 2005; Sala-i-Martin et al., 2004). Our results suggest that the *SICC* may prove superior in many circumstances. This would be good news from a computational perspective. Because information criteria use the sum of squared residuals from linear regressions, it is now relatively easy to search over very large numbers of models in order to find a single best model, or set of models. For example, Reed (2008) appends a standard procedure in SAS that relies on the "leaps and bounds" algorithm developed by Furnival and Wilson (1974). He is able to sort through all possible combinations of 60 variables – a total of approximately $10^{18}$ models – to find a best *AIC* and best *SIC* model. The associated computer program requires about an hour to run on a standard desktop computer. It is hoped that this study stimulates further research into these topics.

## REFERENCES


Burnham, Kenneth P. and David R. Anderson. "Multimodel Inference: Understanding the AIC and BIC in Model Selection." Sociological Methods & Research Vol. 33, No. 2 (November 2004): 261-304..

Campos, Julia, David F. Hendry, and Hans-Martin Krolzig. "Consistent Model Selection by Automatic GETS Approach." Oxford Bulletin of Economics and Statistics Vol. 65 (2003): 803-819.

Fernández, Carmen, Eduardo Ley, and Mark F. J. Steel. "Model Uncertainty in Cross-Country Growth Regressions." Journal of Applied Econometrics Vol. 16 (2001): 563-576.

Furnival, G. M. and Robert W. J. Wilson. "Regression By Leaps and Bounds." Technometrics Vol. 16 (1974): 499-511.

Granger, Clive and Allan Timmermann. "Data Mining with Local Model Specification Uncertainty: A Discussion of Hoover and Perez." Econometrics Journal Vol. 2 (1999): 220-225.

Hendry, David F. and Hans-Martin Krolzig, "The Properties of Automatic GETS Modelling." The Economic Journal, Vol. 115 (March 2005): C32-C61.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. "Bayesian Model Averaging: A Tutorial." Statistical Science Vol. 14, No. 4 (1999): 382-417.

Hoover, Kevin D. and Stephen J. Perez, "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search." Econometrics Journal Vol. 2 (1999): 167-191.

Hoover, Kevin D. and Stephen J. Perez, "Truth and Robustness in Cross-country Growth Regressions." Oxford Bulletin of Economics and Statistics Vol. 66, No. 5 (2004): 765-798.

Hurvich, C. M. and C. L. Tsai. "Regression and Time Series Model Selection in Small Samples." Biometrika, Vol. 76 (1989): 297-307.

Kuha, Jouni. "AIC and BIC: Comparisons of Assumptions and Performance." Sociological Methods & Research Vol. 33, No. 2 (November 2004): 188-229.

McQuarrie, Allan D. "A Small-sample Correction for the Schwarz SIC Model Selection Criterion." Statistics & Probability Letters Vol. 44 (1999): 79-86.

McQuarrie, Allan D. R. and Chih-Ling Tsai. Regression and Time Series Model Selection. Singapore: World Scientific Publishing Co. Pte. Ltd., 1998.

Poskitt, D. S., and A. R. Tremayne. "Determining a Portfolio of Time Series Models." Biometrika Vol. 74, No. 1 (1987): 125-137.

Reed, W. Robert. "The Determinants of U.S. State Economic Growth: A Less Extreme Bounds Analysis." Economic Inquiry (2008): forthcoming.

Sala-i-Martin, Xavier, Gernot Doppelhofer, and Ronald I. Miller, "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach." American Economic Review Vol. 94, No. 4 (2004): 813-835.

Sugiura, N. "Further Analysis of the Data By Akaike's Information Criterion and the Finite Corrections." Communications in Statistics – Theory and Methods, Vol. 7 (1978): 13-26.

Weakliem, David L. "Introduction to the Special Issue on Model Selection." Sociological Methods & Research Vol. 33, No. 2 (November 2004): 188-229.

**TABLE 1**
**Description of Model Selection Criteria**

| | | | |
|---|---|---|---|
| ***Information Criterion (IC) Models:*** | | | |
| ***1)*** | ***AIC*** | $AIC = ln\left(\hat{\sigma}^2\right) + \frac{2\left(\tilde{K}+2\right)}{N}$ | $\hat{\beta}_k$ is the estimate of $\beta_k$ in the model with the minimum *IC* value. If $X_k$ does not appear in that model, $\hat{\beta}_k = 0$. NOTE: $\hat{\sigma}^2$ is the maximum likelihood estimate of the variance of the error term; $\tilde{K}$ is the number of coefficients in the model excluding the intercept; and *N* is the number of observations. |
| ***2)*** | ***AICC*** | $AICC = ln\left(\hat{\sigma}^2\right) + \frac{\left(N+\tilde{K}+1\right)}{\left(N-\tilde{K}-3\right)}$ | |
| ***3)*** | ***SIC*** | $SIC = ln\left(\hat{\sigma}^2\right) + \frac{\left(\tilde{K}+1\right)\cdot ln\left(N\right)}{N}$ | |
| ***4)*** | ***SICC*** | $SICC = ln\left(\hat{\sigma}^2\right) + \frac{\left(\tilde{K}+1\right)\cdot ln\left(N\right)}{\left(N-\tilde{K}-3\right)}$ | |
| ***General-to-Specific Modelling:*** | | | |
| ***13)*** | ***BWStepwise(5%)*** | $\hat{\beta}_k$ is the estimate of $\beta_k$ in the regression model selected through the following iterative process: Step One: Estimate the full model with all variables. Step Two: Exclude the variable with smallest *t*-statistic. Step Three: Continue the process until all variables have a *t*-statistic greater than or equal to the 5% critical value (two-tailed test). | |
| ***(14)*** | ***Autometrics(1%)*** | $\hat{\beta}_k$ is the estimate of $\beta_k$ in the model chosen by the "Autometrics" program in PCGive 12 (1% criterion). If $X_k$ does not appear in that model, $\hat{\beta}_k = 0$. | |
| ***Portfolio Models:*** | | | |
| ***6)*** | ***AIC < 2*** | $\hat{\beta}_k$ is the average value of $\beta_k$ estimates from the portfolio of models that lie within a distance $\Re = 2$ of the respective minimum *IC* model, where $\Re_m = exp\left[-\frac{1}{2}\left(IC_{min} - IC_m\right)\right]$, $IC_{min}$ is the minimum *IC* value among all $2^L$ models, and $IC_m$ is the value of the respective *IC* in model *m*, $m=1,2,\ldots,2^L$. If $X_k$ does not appear in any of the portfolio models, $\hat{\beta}_k = 0$. | |
| ***7)*** | ***AICC < 2*** | | |
| ***8)*** | ***SIC < 2*** | | |
| ***9)*** | ***SICC < 2*** | | |

| | |
|---|---|
| ***Portfolio Models (continued):*** | |
| ***10) AIC < $\sqrt{10}$*** | Same as above, except $\Re = \sqrt{10}$ . |
| ***11) AICC < $\sqrt{10}$*** | |
| ***12) SIC < $\sqrt{10}$*** | |
| ***13) SICC < $\sqrt{10}$*** | |
| ***Bayesian Model Averaging:*** | |
| ***14) LL_Weighted*** | $\hat{\beta}_k$ is the weighted average value of $\beta_k$ estimates over all $2^L$ models, where model weights are determined according to $\omega_m = \frac{\ell_m}{\sum_{m=1}^{2^L} \ell_m}$, $m=1,2,\ldots,2^L$, and $\ell$ is the maximized value of the log likelihood function for model $m$. For the $2^{L-1}$ models where $X_k$ does not appear in any of the portfolio models, $\hat{\beta}_k = 0$. |
| ***15) LL_Weighted( $\hat{\beta}_k \neq 0$ )*** | $\hat{\beta}_k$ is the weighted average value of $\beta_k$ estimates over the $2^{L-1}$ models where $X_k$ is included in the regression equation. Model weights are determined according to $\omega_m = \frac{\ell_m}{\sum_{m \in \{\text{models that contain the variable } X_k\}} \ell_m}$. |

**TABLE 2**
**Comparison of MSC Performance**

| *MSC* | *MSE* | | | *MAD* | | |
|---|---|---|---|---|---|---|
| | *Mean* | *Minimum* | *Maximum* | *Mean* | *Minimum* | *Maximum* |
| ***AIC*** | 7.22 | 2.60 | 13.67 | 7.13 | 2.60 | 13.07 |
| ***AIC < 2*** | 8.68 | 4.97 | 11.35 | 8.67 | 4.90 | 11.30 |
| ***AIC < $\sqrt{10}$*** | 9.10 | 4.97 | 12.93 | 9.07 | 4.90 | 12.80 |
| ***AICC*** | 6.81 | 2.60 | 13.00 | 6.84 | 2.40 | 13.87 |
| ***AICC < 2*** | 8.71 | 4.97 | 11.30 | 8.67 | 4.90 | 11.30 |
| ***AICC < $\sqrt{10}$*** | 9.10 | 4.97 | 12.93 | 9.07 | 4.90 | 12.80 |
| ***GETS*** | 6.44 | 2.50 | 13.60 | 6.44 | 2.40 | 13.73 |
| ***LL_Weighted*** | 6.53 | 1.47 | 9.40 | 6.86 | 1.87 | 10.00 |
| ***LL_Weighted( $\hat{\beta}_k \neq 0$ )*** | 9.70 | 3.00 | 15.00 | 9.48 | 3.20 | 15.00 |
| ***SIC*** | 6.24 | 2.00 | 14.20 | 6.27 | 2.00 | 14.10 |
| ***SIC < 2*** | 8.70 | 4.97 | 11.30 | 8.68 | 4.90 | 11.30 |
| ***SIC < $\sqrt{10}$*** | 9.10 | 4.97 | 12.93 | 9.07 | 4.90 | 12.80 |
| ***SICC*** | 5.91 | 1.00 | 15.00 | 6.01 | 1.00 | 15.00 |
| ***SICC < 2*** | 8.67 | 4.97 | 11.50 | 8.65 | 4.90 | 11.30 |
| ***SICC < $\sqrt{10}$*** | 9.10 | 4.97 | 12.93 | 9.07 | 4.90 | 12.80 |

**TABLE 3**
**Experimental Results for the Case $R^2 = 70\%$, $L = 5$**

| | *AIC* | *AICC* | *GETS* | *LL_Weighted* | *SIC* | *SICC* |
|---|---|---|---|---|---|---|
| ***K=2:*** | | | | | | |
| ***1*** | 5 | 4 | 3 | 7 | 2 | 1 |
| ***2*** | 5 | 4 | 3 | 7 | 2 | 1 |
| ***3*** | 5 | 4 | 3 | 6 | 2 | 1 |
| ***4*** | 5 | 4 | 3 | 6 | 2 | 1 |
| ***5*** | 5 | 4 | 3 | 6 | 2 | 1 |
| ***Average(Irrelevant)*** | *5.0* | *4.0* | *3.0* | *6.0* | *2.0* | *1.0* |
| ***Average(Relevant)*** | *5.0* | *4.0* | *3.0* | *7.0* | *2.0* | *1.0* |
| ***Average(Overall)*** | *5.0* | *4.0* | *3.0* | *6.4* | *2.0* | *1.0* |
| ***K=3:*** | | | | | | |
| ***1*** | 2 | 1 | 5 | 4 | 7 | 6 |
| ***2*** | 4 | 5 | 3 | 7 | 2 | 1 |
| ***3*** | 5 | 4 | 3 | 7 | 2 | 1 |
| ***4*** | 5 | 4 | 3 | 6 | 2 | 1 |
| ***5*** | 5 | 4 | 3 | 6 | 2 | 1 |
| ***Average(Irrelevant)*** | *5.0* | *4.0* | *3.0* | *6.0* | *2.0* | *1.0* |
| ***Average(Relevant)*** | *3.7* | *3.3* | *3.7* | *6.0* | *3.7* | *2.7* |
| ***Average(Overall)*** | *4.2* | *3.6* | *3.4* | *6.0* | *3.0* | *2.0* |
| ***K=4:*** | | | | | | |
| ***1*** | 3 | 2 | 4 | 6 | 5 | 7 |
| ***2*** | 1 | 3 | 6 | 4 | 5 | 7 |
| ***3*** | 1 | 2 | 6 | 4 | 5 | 7 |
| ***4*** | 3 | 2 | 6 | 4 | 5 | 7 |
| ***5*** | 5 | 4 | 3 | 6 | 2 | 1 |
| ***Average(Irrelevant)*** | *5.0* | *4.0* | *3.0* | *6.0* | *2.0* | *1.0* |
| ***Average(Relevant)*** | *2.0* | *2.3* | *5.5* | *4.5* | *5.0* | *7.0* |
| ***Average(Overall)*** | *2.6* | *2.6* | *5.0* | *4.8* | *4.4* | *5.8* |
| ***K=5:*** | | | | | | |
| ***1*** | 2 | 3 | 13 | 4 | 14 | 15 |
| ***2*** | 2 | 3 | 5 | 4 | 6 | 15 |
| ***3*** | 2 | 3 | 5 | 4 | 6 | 7 |
| ***4*** | 2 | 3 | 5 | 4 | 6 | 15 |
| ***5*** | 9 | 10 | 12 | 13 | 14 | 15 |
| ***Average(Overall)*** | *3.4* | *4.4* | *8.0* | *5.8* | *9.2* | *13.4* |

**TABLE 4**
**MSC Performance as a Function of $R^2$, $K/L$, and $K$**

| | *Coefficient* | *t-Stat* | *p-value* |
|---|---|---|---|
| ***Dep. Variable = AIC, R-squared = 0.429*** | | | |
| ***Constant*** | 8.202 | 16.91 | 0.000 |
| ***K/L*** | 3.480 | 6.06 | 0.000 |
| $R^2$ | -0.047 | -6.30 | 0.000 |
| ***Dep. Variable = AICC, R-squared = 0.517*** | | | |
| ***Constant*** | 7.362 | 13.40 | 0.000 |
| ***K/L*** | 5.219 | 7.83 | 0.000 |
| $R^2$ | -0.056 | -6.69 | 0.000 |
| ***Dep. Variable = GETS, R-squared = 0.522*** | | | |
| ***Constant*** | 6.283 | 8.98 | 0.000 |
| ***K/L*** | 7.090 | 9.02 | 0.000 |
| $R^2$ | -0.061 | -5.61 | 0.000 |
| ***Dep. Variable = LL Weighted, R-squared = 0.205*** | | | |
| ***Constant*** | 4.444 | 10.77 | 0.000 |
| ***K/L*** | 0.031 | 0.07 | 0.947 |
| $R^2$ | 0.034 | 5.01 | 0.000 |
| ***Dep. Variable = SIC, R-squared = 0.496*** | | | |
| ***Constant*** | 4.888 | 6.22 | 0.000 |
| ***K/L*** | 8.287 | 9.16 | 0.000 |
| $R^2$ | -0.052 | -4.23 | 0.000 |
| ***Dep. Variable = SICC, R-squared = 0.481*** | | | |
| ***Constant*** | 3.531 | 3.89 | 0.000 |
| ***K/L*** | 10.004 | 9.45 | 0.000 |
| $R^2$ | -0.051 | -3.41 | 0.001 |

NOTE: All equations are estimated using OLS. Each regression uses 120 observations, with each observation corresponding to a single experiment. The dependent variable is the average rank of the respective MSC in a given experiment. Standard errors are robust to heteroscedasticity.

**TABLE 5**
**Comparison of AIC Portfolio Models as a Function of *K* and $R^2$**

| | ***R2 = 90%*** | | | ***R2 = 35%*** | | |
|---|---|---|---|---|---|---|
| | ***AIC*** | ***AIC < 2*** | ***AIC < $\sqrt{10}$*** | ***AIC*** | ***AIC < 2*** | ***AIC < $\sqrt{10}$*** |
| ***K=1*** | | | | | | |
| ***1*** | 13 | 7.5 | 7.5 | 13 | 4.5 | 4.5 |
| ***2*** | 5 | 9 | 13.5 | 5 | 10.5 | 10.5 |
| ***3*** | 5 | 11 | 13.5 | 5 | 11.5 | 11.5 |
| ***4*** | 5 | 9 | 13.5 | 5 | 10.5 | 10.5 |
| ***5*** | 5 | 11 | 13.5 | 5 | 11.5 | 11.5 |
| ***Average*** | *6.6* | *9.5* | *12.3* | *6.6* | *9.7* | *9.7* |
| ***K=2*** | | | | | | |
| ***1*** | 5 | 15 | 9.5 | 10 | 4.5 | 4.5 |
| ***2*** | 5 | 10 | 13.5 | 11 | 4.5 | 4.5 |
| ***3*** | 5 | 8 | 13.5 | 5 | 11.5 | 11.5 |
| ***4*** | 5 | 8 | 13.5 | 5 | 10.5 | 10.5 |
| ***5*** | 5 | 11 | 8.5 | 5 | 10.5 | 10.5 |
| ***Average*** | *5* | *10.4* | *11.7* | *7.2* | *8.3* | *8.3* |
| ***K=3*** | | | | | | |
| ***1*** | 4 | 8 | 13.5 | 11 | 5.5 | 5.5 |
| ***2*** | 4 | 8 | 13.5 | 11 | 5.5 | 5.5 |
| ***3*** | 5 | 8 | 13.5 | 11 | 5.5 | 5.5 |
| ***4*** | 5 | 8 | 13.5 | 5 | 10.5 | 10.5 |
| ***5*** | 5 | 12 | 8.5 | 5 | 10.5 | 10.5 |
| ***Average*** | *4.6* | *8.8* | *12.5* | *8.6* | *7.5* | *7.5* |
| ***K=4*** | | | | | | |
| ***1*** | 2 | 8 | 13.5 | 11 | 5.5 | 5.5 |
| ***2*** | 4 | 15 | 10.5 | 11 | 4.5 | 4.5 |
| ***3*** | 4 | 8 | 13.5 | 11 | 6.5 | 6.5 |
| ***4*** | 5 | 11 | 13.5 | 11 | 4.5 | 4.5 |
| ***5*** | 5 | 12 | 8.5 | 5 | 10.5 | 10.5 |
| ***Average*** | *4* | *10.8* | *11.9* | *9.8* | *6.3* | *6.3* |
| ***K=5*** | | | | | | |
| ***1*** | 3 | 8 | 12.5 | 11 | 4.5 | 4.5 |
| ***2*** | 3 | 8 | 13.5 | 11 | 5.5 | 5.5 |
| ***3*** | 3 | 8 | 13.5 | 11 | 6.5 | 6.5 |
| ***4*** | 4 | 8 | 13.5 | 11 | 4.5 | 4.5 |
| ***5*** | 11 | 8 | 3 | 11 | 4.5 | 4.5 |
| ***Average*** | *4.8* | *8* | *11.2* | *11* | *5.1* | *5.1* |

**TABLE 6**
**Performance of Portfolio Models**

| | *Over All Experiments* | *Over All Experiments Where (i) $R^2$ = 35% Or 50% and (ii) K/L > 0.50* |
|---|---|---|
| ***AIC Portfolio Models Better Than AIC Models*** | 37/120 = 30.8% | 25/32 |
| ***AICC Portfolio Models Better Than AICC Models*** | 39/120 = 32.5% | 25/32 = 78.1% |
| ***SIC Portfolio Models Better Than SIC Models*** | 41/120 = 34.2% | 26/32 = 81.2% |
| ***SICC Portfolio Models Better Than SICC Models*** | 38/120 = 31.7% | 23/32 = 71.9% |
| ***A Portfolio Model Dominates All Other Models*** | 21/120 = 17.5% | 15/32 = 46.9% |

**TABLE 7**
**Comparison of Alternative Log-Likelihood Weighted Models**

**A. An Illustrative Example: The Case of *L = 10*, $R^2 = 50\%$, and *K = 4***

| | *LL_Weighted(All)* | *LL_Weighted($\hat{\beta} > 0$)* |
|---|---|---|
| ***1*** | 10 | 1 |
| ***2*** | 10 | 9 |
| ***3*** | 2 | 1 |
| ***4*** | 10 | 9 |
| ***5*** | 6 | 15 |
| ***6*** | 6 | 15 |
| ***7*** | 5 | 15 |
| ***8*** | 6 | 15 |
| ***9*** | 6 | 7 |
| ***10*** | 6 | 15 |
| ***Average*** | *6.7* | *10.2* |

**B. Percentage of Individual Experiments where *LL_Weighted($\hat{\beta} \neq 0$)* Dominates *LL_Weighted(All)* for the Included Variables**

| | $R^2$=90% | $R^2$=70% | $R^2$=50% | $R^2$=35% | *Sum* |
|---|---|---|---|---|---|
| ***L = 5*** | $^{14}/_{15} = 93.3\%$ | $^{15}/_{15} = 100.0\%$ | $^{15}/_{15} = 100.0\%$ | $^{15}/_{15} = 100.0\%$ | $^{59}/_{60} = 98.3\%$ |
| ***L = 10*** | $^{54}/_{55} = 98.2\%$ | $^{55}/_{55} = 100.0\%$ | $^{34}/_{55} = 61.8\%$ | $^{1}/_{55} = 1.8\%$ | $^{144}/_{220} = 65.5\%$ |
| ***L = 15*** | $^{120}/_{120} = 100.0\%$ | $^{120}/_{120} = 100.0\%$ | $^{114}/_{120} = 95.0\%$ | $^{4}/_{120} = 3.3\%$ | $^{358}/_{480} = 74.6\%$ |
| ***Sum*** | $^{188}/_{190} = 98.9\%$ | $^{190}/_{190} = 100.0\%$ | $^{163}/_{190} = 85.8\%$ | $^{20}/_{190} = 10.5\%$ | $^{561}/_{760} = 73.4\%$ |

**FIGURE 1**
**MSC Performance as a Function of *K*, $R^2$ and *L***

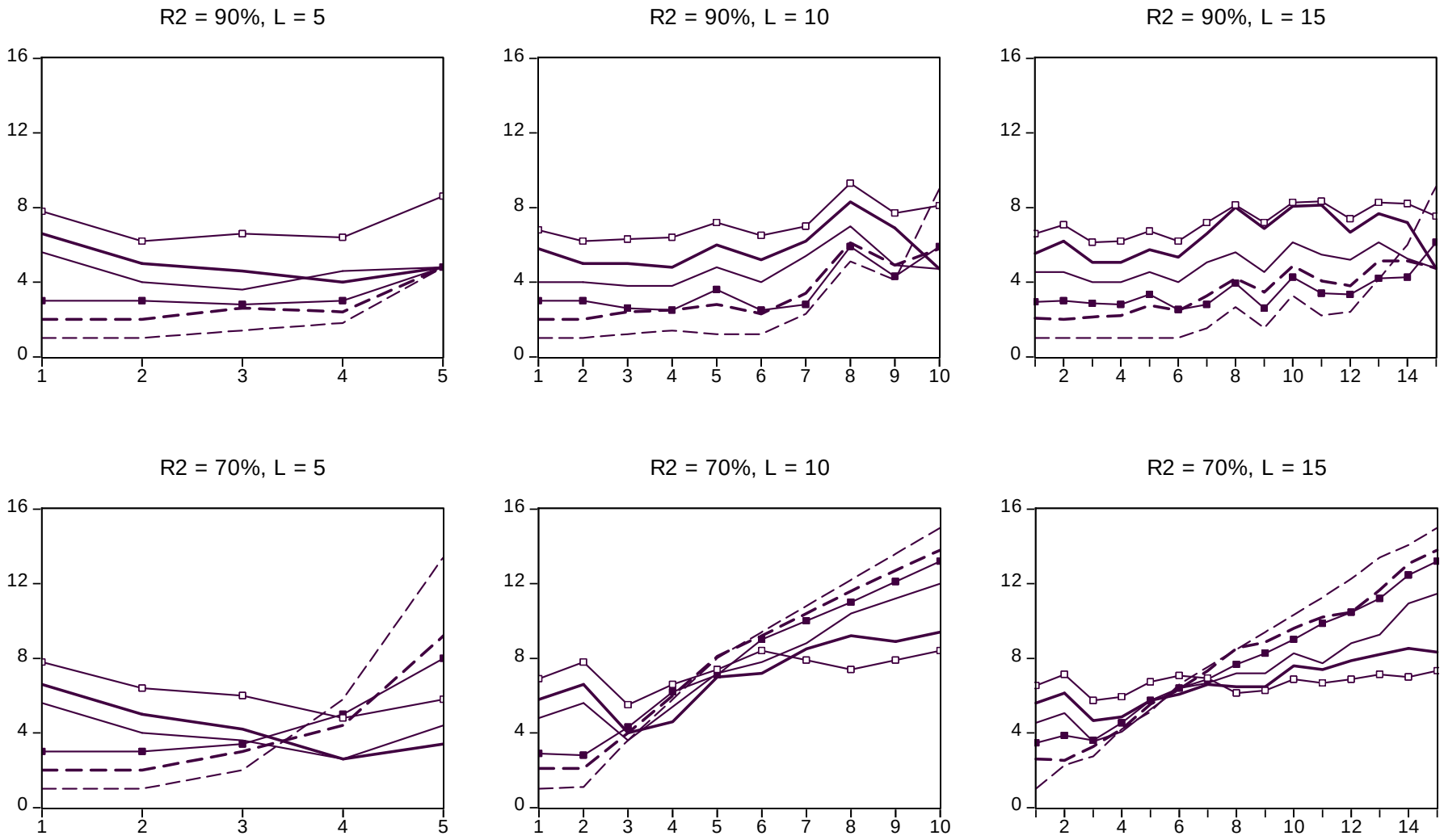

**FIGURE 1 (continued)**
**MSC Performance as a Function of *K*, $R^2$ and *L***