MODEL SELECTION CRITERIA USING LIKELIHOOD FUNCTIONS AND OUT-OF-SAMPLE PERFORMANCE

Bailey Norwood

Peyton Ferrier

Jayson Lusk*

Paper presented at the NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management
St. Louis, Missouri, April 23-24, 2001

* Graduate research assistant (fbnorwoo@unity.ncsu.edu), graduate research assistant (pmferrie@unity.ncsu.edu), Department of Agricultural and Resource Economics, North Carolina State University and Assistant professor (JLusk@agecon.msstate.edu), Department of Agricultural Economics, Mississippi State University. Helpful comments from Eric Cartman are gratefully acknowledged.

MODEL SELECTION CRITERIA USING LIKELIHOOD FUNCTIONS AND OUT-OF-SAMPLE PERFORMANCE

Model selection is often conducted by ranking models by their out-of-sample forecast error. Such criteria only incorporate information about the expected value, whereas models usually describe the entire probability distribution. Hence, researchers may desire a criteria evaluating the performance of the entire probability distribution. Such a method is proposed and is found to increase the likelihood of selecting the true model relative to conventional model ranking techniques.

Keywords: Model selection, forecasting, heteroskedasticity.

Most empirical economic research entails the specification, estimation, and evaluation of statistical models. Once a model is specified, estimation is usually straightforward. This is because most econometric research and class lectures focus on econometric techniques of an assumed model, and there exists a general consensus among practitioners as to the best estimation techniques for most model settings. However, the literatures is not quite as clear on the best method(s) of model selection. Consequently, while estimation of a specified model is more if a science, the act of model specification is partly an art[1].

Unfortunately, results are often sensitive to the choice of functional form and/or error distribution, hereafter referred to as the “model”. Shumway and Lim provide excellent examples of how elasticities vary under alternative functional forms. They summarize the robustness problem by stating (page 275) “Attempting to narrowly bound estimates of output supply and input demand elasticity for a given category remains an exceedingly difficult task. Even using the same data, holding the point of evaluation constant, and using alternative functional forms with the same number of free parameters to be estimated, the implied elasticities can vary widely.”

Although little consensus exists regarding the best method of model selection, most agree models should be ranked by their out-of-sample performance. This is because in-sample statistics are often misleading and arbitrary. Measures of in-sample fit, like the coefficient of determination, arbitrarily prefer models with the most parameters[2].

Adjusts can be made to these measures, but ironically, those adjusts are often deemed arbitrary and hence are unpopular[3]. Hypothesis tests are often used to identify models, but numerous tests can yield conflicting results and a large amount of pre-testing may invalidate statistics from the model[4]. Furthermore, not all models are nested, and non-nested tests are famous for ambiguous conclusions. The likelihood dominance procedure provides an unambiguous model ranking, but requires all models to have an identical number of variables (Anderson et al). Locally flexible functional forms are championed for their ability to approximate any true functional form, but this approximation is only at a point. Globally flexible functional forms provide an approximation at all points, but require numerous parameters thereby impeding the data's ability to "speak."

Every measure of model performance using in-sample statistics has at least one drawback. It is difficult to find a drawback to using out-of-sample statistics like the out-of-sample-root-mean-squared error (OSRMSE) or average-out-of-sample-absolute error (AOSAE) though. Neither will arbitrarily increase of decrease as one adds more variables and yields an unambiguous ranking to any number and class of models. They are appropriate under any sample size, as even if the number of observations are small, cross-validation can be employed for a set of out-of-sample forecasts. Best of all, it has the intuitive appeal that a model—a model being a hypothesis for the process governing an economic variable—is evaluated by how it predicts in the "real world". In this paper, it is taken as given that out-of-sample statistics are the best measures of model performance. This paper then asks: What is the best out-of-sample measure for ranking models? The OSRMSE and AOSAE, though informative, have a potential drawback if the researcher is interesting in more than just the expected value.

Let $\hat{y}_{i,t}$ be an out-of-sample forecast from Model i (Model i is specified as a normal distribution) at time t and y_t be the true value. The forecast error is then $\hat{y}_{i,t} - y_t$ and is a measure of how accurate the mean equation is. Suppose the researcher also had a variance equation which could be used to forecast the variance of $\hat{y}_{i,t} - y_t$. One must wonder if a larger forecast error should necessarily penalize the model if the variance equation predicted a larger error, as both the OSRMSE and AOSAE would.

Consider two models, Models A and B, with identical mean equations but Model A models the variance as a constant and Model B models it as a function. If the mean-equation's forecasts were identical, the OSRMSE and AOSAE would rank them as equally valid models. Consider the difference of the AOSAE for two forecasts

$$(1)\quad \left\{\frac{\left|\hat{y}_{A,t} - y_t\right| + \left|\hat{y}_{A,t+1} - y_{t+1}\right|}{2}\right\} - \left\{\frac{\left|\hat{y}_{B,t} - y_t\right| + \left|\hat{y}_{B,t+1} - y_{t+1}\right|}{2}\right\} = 0\,.$$

Suppose the forecast error in time t+1 was larger than in time t and suppose the variance equation in Model B predicted that larger error. It would then seem that Model B contains more information than Model A and hence should be given a higher ranking. Recently, researchers have questioned whether model ranking criteria should account for more than just the mean equation. Dierson and Manfredo propose an innovative technique for model selection when the researcher is interested in discriminating among various mean and variance equations for a normal distribution.

THE LIKELIHOOD SCORING TECHNIQUE

Dierson and Manfredo propose a method of ranking models they call the Likelihood Scoring Technique. They note that for linear models of the form

(2) $y_t = X_t \hat{b} + e_t$ where e_t is distributed as normally with a constant variance, the term

(3) $\left(\hat{y}_t - y_t\right) \Big/ \sqrt{V\left(\hat{y}_t - y_t\right)}$ is distributed as a t-distribution with N-K degrees of freedom,

where N is the number of observations used to estimate $\hat{b}$, K is the number of parameters, and V(.) denotes variance. The variance of the forecast can be written as

(3) $V\left(\hat{y}_t - y_t\right) = V\left(X_t \hat{b} - X_t b - e_t\right) = V\left(X_t \hat{b} - e_t\right) = X_t V(\hat{b}) X_t' + V(e_t)$.

Although Dierson and Manfredo only consider the case where V(e_t) is a constant, this method easily extends to more general forms by inserting a formula instead of a constant for V(e_t). Dierson and Manfredo suggest a method of model ranking called the Likelihood Scoring Technique which entails conducting a set of T out-of-sample forecasts, calculating the following statistic

(4) $LS = \sum_{t=1}^{T} t_{N-K}\left(\frac{\hat{y}_t - y_t}{\sqrt{X_t V\left(\hat{b}\right) X_t' + V(e_t)}}\right)$ for a series of models, and ranking the models

by their Likelihood Score (LS) where the highest LS is the best model. The term $t_{T\text{-}K}$(a) means the t-distribution with N-K degrees of freedom is evaluated at the value a. This method is appealing in that, holding the forecasted variance constant, models with higher forecast errors will receive a lower ranking, but a higher forecast error does not necessarily penalize a model if its variance equation predicted a higher error.

The Likelihood Scoring Technique assumes normality of errors, but researchers may desire methods which allow different error specifications. We propose a method of model selection which uses out-of-sample forecasts but allows the user to consider various mean and variance equations *and* error distributions. It is based on the concept of a likelihood function.

THE LIKELIHOOD SCORING TECHNIQUE

Both the OSRMSE and AOSAE only incorporate information regarding each model’s expected values. The LST can only be used for normally distributed models. However, researchers often desire to compare alternative distributional assumptions and, since models are often expressed in terms of a pdf, may desire to rank models by how well each model’s pdf performs out-of-sample, i.e., the researcher may want to pick the model with the highest probability of generating a set of out-of-sample observations. A method is developed believed to achieve this. The method is outlined first and is then followed by an explanation of why the method is valid.

The OSLLF Approach To Model Selection

1) Create a set of models the researcher deems appropriate. Each model must be stated in terms of a probability density function (pdf) and all pdf's be a function of the same variable (they must all integrate over the same variable).

2) Estimate the parameters of each model using maximum likelihood.

3) Obtain a set of out-of-sample observations. In small samples use cross-validation. Using the estimated parameters from Step 2, calculate the pdf for each model at each out-of-sample observation. Denote the pdf for an out-of-sample observation y_t as $f(y_t)$.

4) Sum over $\ln\{f(y_t)\}$ for all t and pick the model with the highest $\sum_t \ln\{f(y_t)\}$ as the superior model. If the number of out-of-sample observations differ across models, choose the model with the highest average OSLLF. For small samples, one may want to employ cross validation (which is explained in the simulation description).

As an example, consider again two models; Model A where $y_i \sim N(X_{A,i}\beta_A, Z_{A,i}\alpha_A)$ and Model B where $\ln\{y_i\} \sim N(X_{B,i}\beta_B, Z_{B,i}\alpha_B)$. The first step entails specifying the pdf for each variable such that they integrate over the same variable. Let $f_A(y)$ and $f_B(y)$ be the pdf for Models A and B, respectively.

(5)
$$\begin{aligned} f_A(y_i) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{Z_{A,i}\alpha_A}}\exp\left\{-\frac{(y - X_{A,i}\beta_A)^2}{2Z_{A,i}\alpha_A}\right\} \\ f_B(y) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{Z_{B,i}\alpha_B}}\exp\left\{-\frac{(\ln\{y_i\} - X_{B,i}\beta_B)^2}{2Z_{B,i}\alpha_B}\right\}\frac{1}{y_i} \end{aligned}$$

Step two requires maximum likelihood estimation to obtain the parameter estimates $\hat{\beta}_A, \hat{\alpha}_A, \hat{\beta}_B, \hat{\alpha}_B$ by maximizing the log-likelihood functions for Models A and B, denoted LLF_A and LLF_B, respectively, using N observations.

(6)
$$\begin{aligned} LLF_A &= \sum_{i=1}^{N}\left\langle -\frac{1}{2}\ln\langle 2\pi\rangle - \frac{1}{2}\ln\langle Z_{A,i}\alpha_A\rangle + \left\{-\frac{(y - X_{A,i}\beta_A)^2}{2Z_{A,i}\alpha_A}\right\}\right\rangle \\ LLF_B &= \sum_{i=1}^{N}\left\langle -\frac{1}{2}\ln\langle 2\pi\rangle - \frac{1}{2}\ln\langle Z_{B,i}\alpha_B\rangle + \left\{-\frac{(\ln\{y_i\} - X_{B,i}\beta_B)^2}{2Z_{B,i}\alpha_B}\right\} - \ln\langle y_i\rangle\right\rangle \end{aligned}$$

Notice when estimating parameters one does not have to worry whether all likelihood functions are stated in terms of the same random variable (one does not have to worry whether the pdf's for all models integrate over the same variable). Finally, Step 3 requires a second set of observations denoted t (t is for out-of-sample observations and i is for in-sample observations). Using the estimated parameter values from Step 2, all one needs to do is plug in the out-of-sample observations into the log-likelihood functions to obtain an out-of-sample-log-likelihood function (OSLLF) for Models A and B.

$$(7)\quad \begin{aligned} OSLLF_A &= \sum_{t=1}^{t} \left\langle -\frac{1}{2}\ln\langle 2p\rangle - \frac{1}{2}\ln\langle Z_{A,t}\hat{a}_A\rangle + \left\{ -\frac{\left(y_t - X_{A,t}\hat{b}_A\right)^2}{2Z_{A,t}\hat{a}_A} \right\} \right\rangle \\ OSLLF_B &= \sum_{t=1}^{T} \left\langle -\frac{1}{2}\ln\langle 2p\rangle - \frac{1}{2}\ln\langle Z_{B,t}\hat{a}_B\rangle + \left\{ -\frac{\left(\ln\{y_t\} - X_{B,t}\hat{b}_B\right)^2}{2Z_{B,t}\hat{a}_B} \right\} - \ln\langle y_t\rangle \right\rangle \end{aligned}$$

Finally, if $OSLLF_A > OSLLF_B$ then Model A is superior. Otherwise, Model B is superior.

The procedure is simple and can be used to compare numerous types of models. It can compare various mean and variance equations of a normal distribution or any alternative distribution. The only requirement is that each model be defined as a pdf integrating over the same variable (must be a function of the same variable). However, no real justification has yet been given as to why this is a valid procedure or why each pdf must integrate over the same variable. These two issues are addressed below.

Suppose a researcher is comparing two models for a dependent variable y, where y exists in the (1,2) range. For simplicity, assume there is only one out-of-sample observation. The researcher is considering a normal and a log-normal distribution. Denote the two pdf's for y and ln(y) as f(y) and f(ln(y)), respectively. These likelihood function must be stated such that

(8) $\int f(y)dy = 1$ and

(9) $\int f(\ln(y))d\{\ln(y)\} = 1$.

Ignoring ranges of the normal distribution with very small probabilities of occurrence, in order for the probability of y taking on the values 1.01, 1.02, …, or 1 to equal one, the likelihood function f(y) must be *less* than one for every value of y. However, in order for the probability of ln(y) taking on the values ln(1.01), ln(1.02), …, or ln(1) to equal one, the likelihood function f(ln(y)) must be *greater* than zero for every value of ln(y). Hence, the likelihood function for the model assuming a log-normal distribution of y must be greater than the normal model. Simply by transforming the dependent variable one alters the ranking of models. Obviously, the OSLLF as presented so far is inadequate for dealing with all models.

A simple adjustment corrects for this though. The problem lies in the interpretation of the likelihood function. The likelihood function cannot be interpreted as a probability, it does not even have to be on the (0,1) interval. The likelihood function evaluated at any one point is meaningless--but its integral is. An out-of-sample observation of a variable y_t and its associated OSLLF value is not indicative of the probability of y_t occurring given the model specification and estimated parameters. However the integral of the OSLLF over the range y_t - σ and y_t + σ is indicative of the probability of y_t lying within this range, given the model specification and estimated parameters. Compare again the two previous models of y where one model assumes normality and the other assumes log-normality. Let $x_t = \ln(y_t)$. Though the comparison of (8) to (9) is meaningless, one can compare the integral of each distribution for a small deviations around y_t. These integrals are the probability of the out-of-sample observation y_t occurring, given each model's specification and estimated parameters.

(10) $$\int_{y_t-s}^{y_t+s} \frac{1}{\sqrt{2p}}\frac{1}{\sqrt{s_y^2}}\exp\left(\frac{-1}{2s_y^2}\left(y_t - E\{y_t\}\right)^2\right)dy$$

(11) $$\int_{\ln(y_t-s)}^{\ln(y_t+s)} \frac{1}{\sqrt{2p}}\frac{1}{\sqrt{s_x^2}}\exp\left(\frac{-1}{2s_x^2}\left(x_t - E\{x_t\}\right)^2\right)dx$$

Hence, if (10) is larger than (11), one can say a the normal distribution model is more appropriate than the log-normal. Note that since x and y are monotonic transformations, both integrals can be stated in terms of y using the method of transformations.

(12) $$\int_{y_t-s}^{y_t+s} \frac{1}{\sqrt{2p}}\frac{1}{\sqrt{s_y^2}}\exp\left(\frac{-1}{2s_y^2}\left(y - E\{y_t\}\right)^2\right)dy$$

(13) $$\int_{y_t-s}^{y_t+s} \frac{1}{\sqrt{2p}}\frac{1}{\sqrt{s_x^2}}\exp\left(\frac{-1}{2s_x^2}\left(x - E\{x_t\}\right)^2\right)\left(\frac{1}{y}\right)dy$$

Finally, we can let σ become very small and then get two likelihood functions that can meaningfully be compared.

(14) $$\frac{1}{\sqrt{2p}}\frac{1}{\sqrt{s_y^2}}\exp\left(\frac{-1}{2s_y^2}\left(y_t - E\{y_t\}\right)^2\right)$$

(15) $$\frac{1}{\sqrt{2p}}\frac{1}{\sqrt{s_x^2}}\exp\left(\frac{-1}{2s_x^2}\left(\ln(y_t) - E\{\ln(y_t)\}\right)^2\right)\left(\frac{1}{y_t}\right)$$

Notice the problem before of the likelihood function for $\ln(y_t)$ being greater than zero no longer exists because that likelihood function is divided by y_t. Hence, for likelihood functions for a series of models with different dependent variables, if the dependent variables are all a monotonic function of a single variable, say y_t, the method of transformations must be employed to convert all pdf's into pdf's for y_t (as opposed to functions of y_t).

Only after this transformation can the OSLLF be used. To get a better idea of why the OSLLF is a valid criterion, note that the value of a pdf is approximately the probability of an observation existing in a small interval divided by the range of that interval. This comes from the fact that the true probability of y_t being observed in the y_t - δ and y_t + δ interval is $F(y_t + d) - F(y_t - d) = \int_{y_t - d}^{y_t + d} f(y)dy$ where F(.) is a cumulative distribution function (cdf). So long as δ is small, f(y) will be approximately the same over the entire interval and hence can be approximated by $f(y_t)dy = f(y_t)2\delta$. The choice of δ is arbitrary, but for $f(y_t)$ to be comparable across models δ must be identical across models implying dy must be identical. If all pdf's integrate over the same variable, then the model with the highest pdf is also the model with the highest probability of generating the out-of-sample observation. Then, assuming a set of y_t's (a set of out-of-sample observations) are independent, the model with the highest product of pdf's across a set of out-of-sample observations has the highest probability of generating that sample. Therefore, the model with the highest OSLLF will be the model with the highest probability of generating the set of out-of-sample observations. Please note that one must compare log-likelihood functions across models using *out-of-sample observations*, as log-likelihood functions using *in-sample* observations will tend to arbitrarily prefer models with more parameters.

Ranking models by their OSLLF is an improvement to existing model selection criteria and is desirable for several reasons. First, it employs out-of-sample observations which are not subject to arbitrary manipulation. While the OSRMSE and the AOSAE only compare models by the performance of their expected values, the OSLLF compares models by their entire pdf. Hence, the ability of models to reflect changing variances and skewness will be reflected in the model selection. The OSLLF can be used to compare a large number of models of diverse types (a much larger number than in-sample statistics can compare). Finally, the OSLLF picks the model with the highest probability of generating the set of out-of-sample observations. In addition to its theoretical and practical justifications, the OSLLF was tested in simulations to determine its usefulness in picking the correct model relative to the OSRMSE, AOSAE, and the LST. Whether simulation results can be extended to real settings is questionable. The generality of simulations is always suspect. Regardless, simulations are the only way to compare these four model ranking criteria discussed in this paper and will provide more information regarding the best model ranking criteria than was previously available.

SIMULATION DETAILS

The goal is to make the simulations as general as possible. Therefore, a wide variety of models are used. Let the dependent variable be denoted y and the independent variables influencing y be x_1, x_2, x_3, x_4, x_5. Any x_i may affect y in a non-linear manner, thus interaction effects between x_i and x_j $\forall$ i,j are considered. The vector of possible explanatory variables are $[1, x_1, x_2, x_3, x_4, x_5, x_1^2, x_2^2, x_3^2, x_4^2, x_5^2, x_1x_2, x_1x_3, x_1x_4, x_1x_5, x_2x_3, x_2x_4, x_2x_5, x_3x_4, x_3x_5, x_4x_5]$. Let X be the matrix of $[1, x_1, x_2, x_3, x_4, x_5, x_1^2, x_2^2, x_3^2, x_4^2, x_5^2, x_1x_2, x_1x_3, x_1x_4, x_1x_5, x_2x_3, x_2x_4, x_2x_5, x_3x_4, x_3x_5, x_4x_5]$ for all observations and

X(1:j) represent a matrix with the 1st through jth columns of X. Though the value of j is chosen randomly across simulations, and it is assumed the exact value of j within any simulation *is* known by the researcher.

The true mean equation within a simulation is defined as y = $X^{(\lambda)}(1{:}j)\,\beta(1{:}j)$ for j = 6, …, 21 where $x_i^{(\lambda)}$ is the Box-Cox transformation $x_i^{(l)} = \frac{x_i^{(l)} - 1}{l}$, where each x_i is transformed using the same λ, λ can take randomly the values 0, .5, 1 and the value of λ is *not* known by the researcher within a simulation[5]. The variance equation may take on three forms; Form 1 (s=1) is a constant, Form 2 (s=2) follows $V(e) = \exp\{\alpha_0 + \alpha_1 x_1^{(\lambda)}\}^2$ and Form 3 (s=3) follows $V(e) = \exp\{\alpha_0 + \alpha_1 x_1^{(\lambda)} x_2^{(\lambda)}\}^2$. All errors are assumed to be normal. The true variance equation varies randomly across each simulation and is *not* known by the researcher within a simulation.

The goal of the simulated researcher is to use the OSRMSE, the AOSAE, the LSM, and the OSLLF in choosing which model to use. The researcher knows the mean and variance equations both take one of three forms, resulting in 9 candidate models. In the case of a constant variance, the simulated researcher employs OLS estimation such that(16) $\hat{b}_i = \left(X^{(l)}(1:j)'\,X^{(l)}(1:j)\right)^{-1} X^{(l)}(1:j)\,y$ where i denotes Models 1, 2, and 3. Models hereafter are denoted by M(λ,s) where λ is the value used in the Box-Cox transformation and s denotes which variance equation was used. Models M(λ,s = 2, 3) were estimated using a two-stage least squares estimator. First, an OLS regress identical to (16) was performed. The corresponding residuals were squared, naturally logged, and regressed against $x_1^{(\lambda)}$ and $x_1^{(\lambda)}x_2^{(\lambda)}$ for Models M(λ,s=2) and M(λ,s=3), respectively. For both models, the estimate of α_0 was changed to $\alpha_0 + 1.2704$ to maintain consistency (Greene). Then, the estimated parameters of the variance equation were used to estimate the variance for each observation and then incorporated into a Weighted-Least-Squares Estimator of the mean-equation parameters.

The estimated mean- and variance-equation parameters were then used to conduct out-of-sample forecasts. Simulations were constructed such that the sample size, the error magnitude, and the j in X(1:j) were chosen randomly to provide a wide range of settings. In each simulation the four model ranking criteria 1) OSRMSE 2) AOSAE 3) LSM 4) OSLLF where calculated and used to deem one of the 9 candidate models as "most likely the true model." The estimated true model for each criterion is then evaluated to access the likelihood of choosing the correct model.

SIMULATION RESULTS

If the model was chosen at random, the researcher would have a 1/9 = $.\overline{1}$ chance of picking the correct model. While three of the four criteria improved this probability, the improvement was small, as shown by Table 1. The OSLLF yielded the highest percentage of times picking the correct model and the LSM yielded the lowest. A series of t-tests were conducted for each pair of criteria to determine if they were significantly different and are given in Table 2. Only the LSM and the AOSAE were not significantly different from one another, concluding the OSLLF is the superior model selection criterion for the 4,000 simulations. Whether this result is extendable to other settings is questionable, but the simulations do provide some guidance whereas before there was none.

T-tests were also conducted to determine if using any criteria are better than choosing the model at random. Test results are shown in Table 3. Only the LSM was not significantly different than choosing the model at random. Although the probability of selecting the correct model using the OSRMSE, AOSAE, or the OSLLF was not greatly above 1/9, they are significantly greater implying they are useful.

Simulations were conducted for various error sizes, model types, parameter values, and sample sizes which yield somewhat general implications. However, the simulation assumed only nine possible functional forms and the researcher knew each of these nine. In reality, the number of possible models is large and researchers could never evaluate each one. Such assumptions needed to be made to conduct simulations though, and although results will not be extendable to all possible settings a researcher may face, they do provide some degree of guidance, whereas in the past there was none.

If researchers wish to evaluate various types of functional forms for the mean and variance of a process and hypothesis tests cannot be used, various model ranking criteria are available. Although in-sample statistics may provide some information regarding which model(s) are most adequate descriptions of an economic process, out-of-sample model performance will be the most reliable. Most commonly used model ranking criteria are the OSLLF and the AOSAE, however, no one has of yet evaluated which of the two are superior. The LSM has recently been proposed as a useful criteria, but again, how it compares to the OSRMSE and the AOSAE has not been evaluated. Though useful, these three measures do not explicitly evaluate the performance of the variance equation, which may be desired if data displays non-spherical errors.

The OSLLF, AOSAE, and the LSM automatically penalize models with larger forecast errors, but it was argued this penalty should not be automatic if the variance equation predicted a larger error. If researchers are specifying both a mean and variance equation, it was argued that the model ranking criteria should evaluate them simultaneously. It was shown the LSM can be extended to include a variance equation. An additional criteria, the OSLLF, was suggested as an alternative criteria. Both the LSM and the OSLLF simultaneously evaluate the mean and variance equation and hence may improve model selection. To test whether this is true, simulations were conducted.

The LSM was found to be uninformative, as picking a model at random has the same probability of picking the correct model. In order of highest to lowest performance, the best measures are the OSLLF, OSRMSE, AOSAE, and either the LSM or a random pick. The probability of picking the correct model using the OSLLF, OSRMSE, and the AOSAE are significantly higher than a random pick and each other. Though not necessarily extendable to all cases, as a rule of thumb, the OSLLF is the best method of simultaneously ranking various mean equations, variance equations, and error distributions. Simulations results suggest that, although these criteria are useful, model selection remains difficult, as the best criteria has only a small chance of picking the correct model.

FOOTNOTES

1) As a percent of total econometric lectures, the amount given to model specification is small. Opinions as to the best methods tend to vary, hence, the method of model specification is mostly chosen by the researcher's beliefs and experience and not by the replication and confirmation of research.

2) The coefficient of determination and the sum-of-squared errors can be set to one and zero, respectively, by setting the number of independent variables equal to the number of observations.

3) Such adjustments are; the adjusted coefficient of determination and information measures based on the Kullback Information Criteria, such as the Akaine Information Criteria (AIC) and the Bayesian Information Criteria (BIC). While the AIC and BIC are often employed in time-series models, they are rarely used to rank econometric models.

4) For example, if a researcher chose a model by removing variables with low significance, the remaining t-ratios are not statistics because the probability of them being significant are 100%.

REFERENCES

Anderson, D.P., T. Chaisantikulawat, A.T.K Guan, M. Kebbeh, N. Lin, and C.R. Shumway. "Choice of Functional Form for Agricultural Production Analysis." *Review of Agricultural Economics.* 18(1996):223-231.

Dierson, Matthew A. and Mark R. Manfredo, "Forecast Evaluation: A Likelihood Scoring Method," NCR-134 Conference: Applied Commodity Price Analysis, Forecasting, and Market Risk Management, April 20-21, 1998.

Green, William H. Econometric Analysis. Third Edition. Princeton Hall. 1997.

Shumway, Richard C. and Hongil Lim. "Functional Form and U.S. Agricultural Production Elasticities." *Journal of Agricultural and Resource Economics*. 18:2(1993):266-276.

TABLE 1.
PERCENT OF TIME MODEL SELECTION CRITERIA PICKED THE CORRECT MODEL

Out-of-Sample-Root-Mean-Squared Error	Average-Out-of-Sample-Absolute Error	Likelihood Scoring Method	Out-of-Sample Likelihood Function
4,000 Simulations			
14%	13%	12%	16%

TABLE 2.
T-TESTS FOR DIFFERENCES IN PERCENT OF TIMES CORRECT MODEL IS CHOSEN

	Out-of-Sample-Root-Mean-Squared Error	Average-Out-of-Sample-Absolute Error	Likelihood Scoring Method	Out-of-Sample Likelihood Function
	Test statistic for the null hypothesis that the percent of times the correct model is chosen for the criteria in column minus the criteria in row is equal to zero			
	Test Statistic[a] is $\dfrac{\overline{P}_i - \overline{P}_j}{\sqrt{\dfrac{\overline{P}_i(1-\overline{P}_i)}{T} + \dfrac{\overline{P}_j(1-\overline{P}_j)}{T} - \dfrac{2\operatorname{cov}(D_i, D_j)}{T}}}$			
Out-of-Sample-Root-Mean-Squared Error	-------	-2.71	-3.14	3.33
Average-Out-of-Sample-Absolute Error	-------	-------	-1.19	5.23
Likelihood Scoring Method	-------	-------	-------	7.05
Out-of-Sample Likelihood Function	-------	-------	-------	-------

a) D_i = 1 if criterion i picked the correct model and zero otherwise and $\overline{P}_i = \sum_{i=1}^{T} D_i \Big/ T$. Subscripts i and j refer to the criterion in the corresponding column and row, respectively. The covariance term is included because two criteria may perform better under identical settings. For instance, the LSM and OSLLF may both perform better if there is heteroskedasticity, hence they may be correlated.

TABLE 3.
T-TESTS FOR PERCENT OF TIMES CORRECT MODEL IS CHOSEN FOR EACH CRITERION IS SIGNIFICANTLY BETTER THAN CHOOSING MODEL AT RANDOM

Out-of-Sample-Root-Mean-Squared Error	Average-Out-of-Sample-Absolute Error	Likelihood Scoring Method	Out-of-Sample Likelihood Function
Test statistic for the null hypothesis that the percent of times the correct model is chosen for the criterion is significantly higher than 1/9.			
Test Statistic[a] is $\frac{\overline{P}_i - (1/9)}{\sqrt{\frac{\overline{P}_i(1-\overline{P}_i)}{T}}}$			
5.51	3.15	1.68	9.07

a) $\overline{P}_i$ is the percent of time criterion i selected the correct model.

APPENDIX A
SIMULATION DESCRIPTION

Dependent Variable: y
Independent Variables Affecting y: x_1, x_2, x_3, x_4, x_5 = x

Step 1: Generating Data on x and Parameter Values

The variables x_1, x_2, x_3, and x_5 were chosen randomly from a normal distribution with a mean of 100 and a standard deviation of 20. The unconditional distribution of x_4 is the same, however, it was set to have a correlation of .3 with x_3.

$x_i \sim N(100,20^2)$ for i = 1, 2, 3, 5.
$x_4 \sim N(100 + .3(x_3-100),\{20^2(1-.3^2)\})$

Step 2: Generate Model and Sample Size

The process of y is given by $y = X^{(\lambda)}(1:j)\beta(1:j) + e$ and is described below. The matrix X(1:j) is a matrix containing x_i's. The superscript (λ) signifies that each x_i undergoes a Box-Cox transformation. The value of lambda used for the Box-Cox transformation is chosen randomly. It may take the values .01, .5, or 1 with an equal probability. Each x_i is transformed as $x_i^{(l)} = \left(x_i^l - 1\right)/l$. Let $X^{(\lambda)} = [I_c, x_1^{(\lambda)}, \ldots, x_5^{(\lambda)}, [x_1^{(\lambda)}]^2, \ldots, [x_5^{(\lambda)}]^2, x_1^{(\lambda)}x_2^{(\lambda)}, \ldots, x_1^{(\lambda)}x_5^{(\lambda)}, x_2^{(\lambda)}x_3^{(\lambda)}, \ldots, x_4^{(\lambda)}x_5^{(\lambda)}]$ where I_c denotes a column of ones. Then, $X^{(\lambda)}(1:j)$ contains the first j columns of $X^{(\lambda)}$. For instance, $X^{(\lambda)}(1:6) = [I_c, x_1^{(\lambda)}, x_2^{(\lambda)}, x_3^{(\lambda)}, x_4^{(\lambda)}, x_5^{(\lambda)}]$ and $X^{(\lambda)}(1:21) = [I_c, x_1^{(\lambda)}, \ldots, x_5^{(\lambda)}, [x_1^{(\lambda)}]^2, \ldots, [x_5^{(\lambda)}]^2, x_1^{(\lambda)}x_2^{(\lambda)}, \ldots, x_1^{(\lambda)}x_5^{(\lambda)}, x_2^{(\lambda)}x_3^{(\lambda)}, \ldots, x_4^{(\lambda)}x_5^{(\lambda)}]$. The value of j can take on 6, …, 21; each with an equal probability. Next, the parameter vector mapping $X^{(\lambda)}(1:j)$ into the expected value of y is $\beta(1:j)$. Let $\beta = [\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4\ \beta_5\ \beta_{12}\ \beta_{22}\ \beta_{32}\ \beta_{42}\ \beta_{52}\ \beta i_{12}\ \beta i_{13}\ \beta i_{14}\ \beta i_{15}\ \beta i_{23}\ \beta_{i24}\ \beta_{i25}\ \beta_{i34}\ \beta i_{35}\ \beta_{i45}]'$ where β_{i25} denotes the parameter corresponding to the interaction term $x_1^{(\lambda)}x_5^{(\lambda)}$. The vector $\beta(1:j)$ then contains the first j rows of β.

The moments of each parameter are:
$\beta_0 \sim N(10000,100^2)$; $\beta_1 \sim N(10,3^2)$; $\beta_2 \sim N(20,6^2)$; $\beta_3 \sim N(15,4^2)$; $\beta_4 \sim N(8,2^2)$;
$\beta_5 \sim N(18,5^2)$; $\beta_{12} \sim N(.01,.005^2)$; $\beta_{22} \sim N(.001,.005^2)$; $\beta_{32} \sim N(.03,.008^2)$;
$\beta_{42} \sim N(.004,.0002^2)$; $\beta_{52} \sim N(.0005,.00004^2)$; $\beta_{i12} \sim N(.001,.005^2)$; $\beta_{i13} \sim N(.0001,.0005^2)$
$\beta_{i14} \sim N(.0005,.005^2)$; $\beta_{i15} \sim N(.0008,.0002^2)$; $\beta_{i23} \sim N(.00001,.00005^2)$;
$\beta_{i24} \sim N(.001,.0005^2)$; $\beta_{i25} \sim N(.01,.005^2)$; $\beta_{i34} \sim N(.0003,.0001^2)$;
$\beta_{i35} \sim N(.00025,.00005^2)$; $\beta_{i45} \sim N(.0025,.003^2)$

After these parameter values are simulated, each parameter is multiplied by $20^{(1-\lambda)}$ and then multiplied by one with a 50% chance and by –1 with a 50% chance. After the vector β is simulated, the expected value of y is then denoted as $X^{(\lambda)}(1:j)\beta(1:j)$.

For each mean equation there are three potential variance equations; two with and one without heteroskedasticity. A random variable s is created which equals 1, 2, or 3 with equal probability. If s = 1 the variance equals a constant. If s = 2 the variance equals $[\exp(\alpha_0 + \alpha_1 x_1^{(\lambda)})]^2$ and if s = 3 the variance equals $[\exp(\alpha_0 + \alpha_1 x_1^{(\lambda)} x_2^{(\lambda)})]^2$. First, consider the case where s = 1 and there is no heteroskedasticity. The variance is set to be a proportion of $mX^{(\lambda)}(1{:}j)\beta(1{:}j)$ where $mX^{(\lambda)}$ (1:j) is the sample mean vector of $X^{(\lambda)}$ (1:j). Let g be a random variable which may take on the values .05, .06, …, .25 with equal (1/25) probability. The variance is set to equal $[gmX^{(\lambda)}(1{:}j)\beta(1{:}j)]^2$

In the case where s = 2, the variance equation equals $[\exp(\alpha_0 + \alpha_1 x_1)]^2$. A lower bound for x_1 is $40^{(\lambda)}$ and an upper bound is $160^{(\lambda)}$. A random variable τ is created which takes on the values 1, 1.01, 1.02, …, 2 with equal (1/100) probability. The error variance is allowed to be decreasing and increasing in x_1 with equal probability. The values of α_0 and α_1 depend on whether the variance is increasing or decreasing in x_1.

Case 1: Error Variance is Increasing in x_1

In this case, the parameter vector $\alpha = [\alpha_0\ \alpha_1]$ is set such that $\exp(\alpha_0 + \alpha_1 40^{(\lambda)}) = (gmX^{(\lambda)}(1{:}j)\beta(1{:}j))^2$ and $\exp(\alpha_0 + \alpha_1 160^{(\lambda)}) = (\tau gmX^{(\lambda)}(1{:}j)\beta(1{:}j))^2$.

Case 2: Error Variance is Decreasing in x_1

In this case, the parameter vector $\alpha = [\alpha_0\ \alpha_1]$ is set such that $\exp(\alpha_0 + \alpha_1 40^{(\lambda)}) = (\tau gmX(1{:}j)\beta(1{:}j))^2$ and $\exp(\alpha_0 + \alpha_1 160^{(\lambda)}) = (gmX^{(\lambda)}(1{:}j)\beta(1{:}j))^2$.

After α has been solved for the model is complete. The model is then said to be

$y = X^{(\lambda)}(1{:}j)\beta(1{:}j) + e$ where $e \sim N(0,\exp\{\alpha_0 + \alpha_1 x_1^{(\lambda)}\})$.

If s = 3 and the variance is $\exp\{\alpha_0 + \alpha_1 x_1^{(\lambda)} x_2^{(\lambda)}\}$, the lower bound for $x_1^{(\lambda)} x_2^{(\lambda)}$ is $40^{(\lambda)}40^{(\lambda)}$ and the upper bound is $160^{(\lambda)}160^{(\lambda)}$.

Case 1: Error Variance is Increasing in x_1

In this case, the parameter vector $\alpha = [\alpha_0\ \alpha_1]$ is set such that $\exp(\alpha_0 + \alpha_1 40^{(\lambda)}40^{(\lambda)}) = (gmX^{(\lambda)}(1{:}j)\beta(1{:}j))^2$ and $\exp(\alpha_0 + \alpha_1 160^{(\lambda)}160^{(\lambda)}) = (\tau gmX^{(\lambda)}(1{:}j)\beta(1{:}j))^2$.

Case 2: Error Variance is Decreasing in x_1

In this case, the parameter vector $\alpha = [\alpha_0\ \alpha_1]$ is set such that $\exp(\alpha_0 + \alpha_1 40^{(\lambda)}40^{(\lambda)}) = (\tau gmX(1{:}j)\beta(1{:}j))^2$ and $\exp(\alpha_0 + \alpha_1 160^{(\lambda)}160^{(\lambda)}) = (gmX^{(\lambda)}(1{:}j)\beta(1{:}j))^2$.

After α has been solved for the model is complete. The model is then said to be

$y = X^{(\lambda)}(1{:}j)\beta(1{:}j) + e$ where $e \sim N(0,\exp\{\alpha_0 + \alpha_1 x_1^{(\lambda)} x_2^{(\lambda)}\})$.

The 9 candidate models are then described by the three possible values for λ and the three different variance specifications (s = 1, 2, and 3). Each model can then be denoted M(λ,s).

Step 4: Generating the Data

Data on $X^{(\lambda)}$ has already been generated. Data on y is then generating by simulating values of e from a normal distribution with a zero mean and variance as described above. The sample size is may take the values n = 30, 40, …, 120, 1000, 1500, 2000 with equal probability. The "dataset" is then the collection of y's and x_i's.

Step 5: Estimation of Candidate Models

All estimations are conducted using Weighted Least Squares, except for the candidate models with a constant variance. For the three candidate models with a constant variance, M(λ=1, 2, or 3, s=1), the estimate is $\hat{b} = \left(X^{(l)}(1:j)'X^{(l)}(1:j)\right)^{-1}X^{(l)}(1:j)y$.

For the six candidate models with heteroskedasticity a Two-Stage Weighted Least Squares is used. The first stage consists of the OLS estimation $\hat{b}_1 = \left(X^{(l)}(1:j)'X^{(l)}(1:j)\right)^{-1}X^{(l)}(1:j)$ and the corresponding residual vector $e = X\hat{b} - y$. Each element in the e vector is then squared and then its natural logarithm is taken. If the specified error variance is V(e) = exp{α_0 + $\alpha_1 x_1^{(\lambda)}$}, each squared residual is naturally logged and regressed against an intercept and $x_1^{(\lambda)}$ to obtain estimates of α_0 and α_1. The estimated standard deviation of the error term is then $\left(\exp\left\{\hat{a}_0 + 1.2704 + \hat{a}_0 x_1^{(l)}\right\}\right)^{1/2} = s$. Each observation of y and $X^{(\lambda)}$(1:j) is then divided by σ and is denoted y* and $X^{(\lambda)}$(1:j)*, respectively. If the specified error variance is V(e) = V(e) = exp{α_0 + $\alpha_1 x_1^{(\lambda)} x_2^{(\lambda)}$}, natural logarithm of each squared residual is regressed against a constant and $x_1^{(\lambda)} x_2^{(\lambda)}$ to obtain estimates of α_0 and α_1. The estimated standard deviation of the error term is then $\left(\exp\left\{\hat{a}_0 + 1.2704 + \hat{a}_0 x_1^{(l)} x_1^{(l)}\right\}\right)^{1/2} = s$. Again, each observation of y and $X^{(\lambda)}$(1:j) is then divided by σ and is denoted $y^{(\lambda)}$* and $X^{(\lambda)}$(1:j)*, respectively. Finally, the Weighted Least Square Estimate is $\hat{b} = \left(X^{(l)}(1:j)*'X^{(l)}(1:j)*\right)^{-1}X^{(l)}(1:j)*y*$.

Step 6: Calculating the Model Selection Criteria:

Out-of-sample-root-mean-squared error (OSRMSE) and average-out-of-sample-absolute error (AOSAE) are two model selection criteria used which do not take into account the variance equation. These two measures are calculated as

$$OSRMSE = \sqrt{\frac{\sum_{t=1}^{T}\left(\hat{y}_t - y_t\right)^2}{T}} \text{ and}$$

$$AOSAE = \sum_{t=1}^{T}\frac{\left|\hat{y}_t - y_t\right|}{T}$$ where t is an out-of-sample forecast, y is the true value, and $\hat{y}$ is the prediction.

How LSM and OSLLF is calculated depends on how the variance is specified. If the variance is modeled as a constant, its estimate is

$$\hat{V}(e) = \frac{\sum_{i=1}^{N} e_i^2}{N-K}$$ where K is the number of parameters estimated and i = 1, …, N denote in-sample residuals. The LSM and OSLLF measures for Models M(λ,s=1) are

$$LSM = \sum_{t=1}^{T} t_{T-K}\left(\frac{\hat{y}_t - y_t}{\sqrt{\hat{V}(e)\left(1+X_t^{(l)}\left(X^{(l)}{}'X^{(l)}\right)^{-1}X_l^{(l)}{}'\right)}}\right) \text{ and}$$

$$OSLLF = -\sum_{t=1}^{T}\left\{\frac{1}{2}\ln\left(2p\right)+\frac{1}{2}\ln\left(\hat{V}(e)\right)+\frac{(\hat{y}_t - y_t)^2}{2\hat{V}(e)}\right\}.$$

For Models M(λ,s=2) the LSM and OSLLF are calculated as

$$LSM = \sum_{t=1}^{T} t_{T-K}\left(\frac{\hat{y}_t - y_t}{\sqrt{\left(\left[\exp\left\{\hat{a}_1+\hat{a}_2 x_{1,t}^{(l)}\right\}\right]+X_t^{(l)}\left(X^{(l)}{}'{*}X^{(l)}{*}\right)^{-1}X_l^{(l)}{}'\right)}}\right)\text{and}$$

$$OSLLF = -\sum_{t=1}^{T}\left\{\frac{1}{2}\ln\left(2p\right)+\frac{1}{2}\ln\left(\left[\exp\left\{\hat{a}_0+\hat{a}_1 x_1^{(l)}\right\}\right]\right)+\frac{(\hat{y}_t - y_t)^2}{2\left[\exp\left\{\hat{a}_0+\hat{a}_1 x_{1,t}^{(l)}\right\}\right]}\right\}$$

and for Models M(λ,s=3) are

$$LSM = \sum_{t=1}^{T} t_{T-K} \left(\frac{\hat{y}_t - y_t}{\sqrt{\left(\left[\exp\left\{ \hat{a}_1 + \hat{a}_2 x_{1,t}^{(l)} x_{2,t}^{(l)} \right\} \right] + X_t^{(l)} \left(X^{(l)\prime} * X^{(l)} * \right)^{-1} X_l^{(l)\prime} \right)}} \right) \text{and}$$

$$OSLLF = -\sum_{t=1}^{T} \left\{ \frac{1}{2} \ln(2p) + \frac{1}{2} \ln\left(\left[\exp\left\{ \hat{a}_0 + \hat{a}_1 x_1^{(l)} x_2^{(l)} \right\} \right] \right) + \frac{(\hat{y}_t - y_t)^2}{2 \left[\exp\left\{ \hat{a}_0 + \hat{a}_1 x_{1,t}^{(l)} x_{2,t}^{(l)} \right\} \right]} \right\}$$

How out-of sample forecasts were conducted depends on the sample size. If the sample size is less than 60 cross-validation is used. This entails a number of different simulations, within each simulation, equal to the sample size. For instance, if the sample size is 30, the first out-of-sample forecast is conducted by estimating the parameters in the mean and variance equation using observations 2 through 30 and forecasting the first observation. Then, the parameters are estimated using observations 1 and 3 through 30 and forecasting the second. This continues until the last observation is dropped, the parameters are estimated using observations 1 through 29, and the last observation is forecasted. The parameters used in the calculation of OSRMSE, AOSAE, LSM, and OSLLF are then different for each cross-validation. If the sample is greater than 60, the first half is used to estimate the parameters and the second half employs those parameters in out-of-sample forecasts.

Step 7: Model Selection:

Within each simulation, each model selection criterion; OSRMSE, AOSAE, LSM, and OSLLF, are used to pick the superior model. The model chosen by the OSRMSE and the AOSAE is the model with the lowest OSRMSE and AOSAE. The model picked by the LSM and the OSLLF is the model with the highest LSM and OSLLF.

Step 8: Determining the Performance of Each Criteria

Simulations were conducted as described above 4,000 times. The OSRMSE, AOSAE, LSM, and OSLLF are judged by the percent of times they picked the correct model. Since there were nine possible models, a criteria is judged as informative if it picks the correct model with a frequency significantly greater than 1/9. Then, t-tests are conducted for each pair to determine if one picks the correct model at a significantly higher rate than another. How the tests were conducted is described in Tables 2 and 3.