Journal of Agricultural and Resource Economics 29(3):404–419
Copyright 2004 Western Agricultural Economics Association

Model Selection for Discrete Dependent Variables: Better Statistics for Better Steaks

F. Bailey Norwood, Jayson L. Lusk, and B. Wade Brorsen

Little research has been conducted on evaluating out-of-sample forecasts of discrete dependent variables. This study describes the large and small sample properties of two forecast evaluation techniques for discrete dependent variables: receiver-operator curves and out-of-sample log-likelihood functions. The methods are shown to provide identical model rankings in large samples and similar rankings in small samples. The likelihood function method is better at detecting forecast accuracy in small samples. By improving forecasts of fed cattle quality grades, the forecast evaluation methods are shown to increase cattle marketing revenues by $2.59/head.

Key words: discrete dependent variables, forecasting, likelihood functions, model selection, out-of-sample, quality grades, receiver-operator curves

Introduction

Model selection is perhaps the most difficult task in applied economic analysis. While economic theory assists in model formation, it rarely identifies a specific model. Researchers often turn to model selection criteria to narrow the field of candidate models. The appropriate model selection criterion is partly driven by the intended use of the model. When a model is used to predict a dependent variable, researchers will want a criterion that ranks models based on model "fit" or how well models predict movements in the dependent variable.

Many such criteria are based on in-sample statistics, such as likelihood-ratio tests and the Akaike Information Criterion (AIC). Likelihood-ratio tests can be used to assess whether additional parameters increase prediction accuracy enough to be included in the model, whereas the AIC directly measures prediction accuracy while adjusting for the number of parameters. Other criteria are based on out-of-sample criteria. In some settings, out-of-sample criteria are preferred. Neural networks, for example, are susceptible to over-fitting and require out-of-sample forecasts for validation. Other times, the choice between in-sample and out-of-sample criteria is less clear and is determined by researcher preference. For instance, Piggott (2003) placed similar weight on in-sample and out-of-sample criteria in selecting among 14 demand systems. Others place greater weight on out-of-sample than in-sample criteria. Kastens and Brester (1996) argue that

---

F. Bailey Norwood is assistant professor, Department of Agricultural Economics, Oklahoma State University; Jayson L. Lusk is associate professor, Department of Agricultural Economics, Purdue University; and B. Wade Brorsen is Regents Professor and Jean & Patsy Neustadt Chair, Department of Agricultural Economics, Oklahoma State University.

Review coordinated by David Aadland.

economic restrictions should be incorporated in demand systems, despite the fact that they are rejected in-sample, because they improve out-of-sample forecasts.

Comparing forecasts between models is relatively straightforward when the forecasted variable is continuous. Typically, the model with lowest mean squared forecast error is preferred. Hypothesis tests such as the AGS test (Ashley, Granger, and Schmalensee, 1980) and a recently developed test by Ashley (1998) can be used to discern whether forecast errors from competing models are significantly different. How one should compare forecasts of discrete variables has received less attention.

Despite the lack of work in this area, economists are faced with a plethora of problems associated with discrete variables. Examples include problems dealing with sample selection bias (Heckman, 1979), technology adoption (Roberts, English, and Larson, 2002), predicting turning points (Dorfman, 1998), consumer choice (Loureiro and Hine, 2002), and willingness to pay (Loomis, Bair, and Gonzalez-Caban, 2002; Haener, Boxall, and Adamowicz, 2001). Clearly, researchers are in need of methods to evaluate the forecasting performance of models with discrete dependent variables. Moreover, as methods susceptible to over-fitting, such as neural networks, are increasingly applied to discrete dependent variables, forecast evaluation will become a necessary component of model selection.

Forecasting discrete dependent variables is more difficult than continuous variables. For instance, suppose we are interested in forecasting a variable G, which can only take the values zero or one. Standard logit and probit models forecast the probability G equals one. Although a higher probability indicates a greater likelihood G will equal one, it is not clear what threshold this probability should exceed before officially forecasting "$G = 1$." Often a threshold of 0.5 is used, but this choice is only desirable if the cost of misclassifying a "$G = 1$" is equal to the cost of misclassifying a "$G = 0$."[1] The optimal threshold depends on the cost (benefit) of an incorrect (correct) forecast of "$G = 1$" and "$G = 0$," which requires specifying a loss function. Because the threshold choice is problem-dependent, and forecasting performance may differ with small changes in the threshold, general methods of model selection should not depend on a specific threshold.

Moreover, since a forecast using a threshold will be either zero or one, the mean squared error criterion will assign a confident correct forecast (such as a forecasted probability of 0.99) a score equal to a less-confident, but nevertheless correct, forecast (such as a forecasted probability of 0.51). This article compares and contrasts two methods for evaluating forecasts of discrete dependent variables over all or many possible thresholds. The first is borrowed from the medical profession, and is referred to as receiver-operator curves (ROCs). The second method entails ranking models by likelihood function values observed at out-of-sample observations. We refer to this approach as the out-of-sample log-likelihood function (OSLLF) approach.

After outlining the two methods, we introduce the concept of divergent distributions, which is the source of forecast accuracy for discrete dependent variables. The greater the divergence, the greater the forecast accuracy. We then show that ROCs and OSLLFs are both measures of divergence, and prove that both criteria will provide an identical model ranking and will choose the best model in large samples. Next, simulations are

[1] For example, in cancer detection where $G = 1$ indicates cancer and $G = 0$ indicates no cancer, a lower threshold than 0.5 would be used. This is because the cost of inaccurately predicting "no cancer" can be deadly for the patient, while the cost of inaccurately predicting "cancer" is smaller.

used to determine which criterion performs best in small samples. ROCs are useful because they allow visual inspection of forecast performance and are absolute measures of forecast ability, while OSLLFs only provide relative measures of forecast accuracy. However, if the task is to choose between two models, simulations reveal a slight preference for the OSLLF criterion. Finally, the model selection criteria are applied to a problem recently posed by Lusk et al. (2003) involving the prediction of cattle quality grades.

Forecasting Discrete Dependent Variables

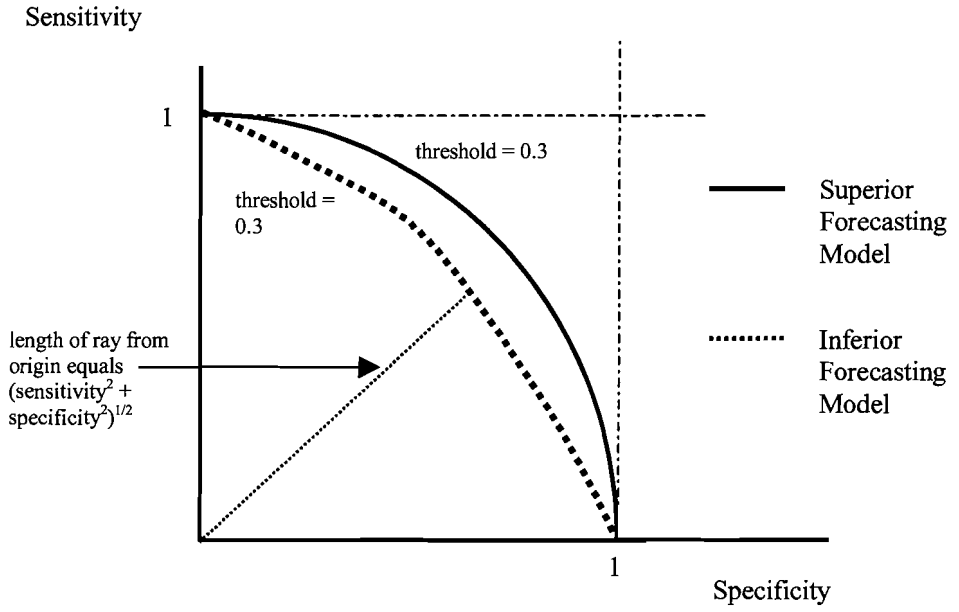
Suppose the variable of interest, G, can only take on the values zero or one. Most models do not output the forecasts "$G = 1$" or "$G = 0$," but instead output the probability that G will equal one. The researcher must then specify a threshold to officially forecast "$G = 1$." As noted previously, this threshold is problem-specific. Rather than rank models at one particular threshold, many in the medical profession have elected to rank models based on their forecasting ability at all threshold values.

Model performance is often measured by the frequency of observations where "$G = 1$" is correctly forecasted, or $P(\hat{G} = 1 \mid G = 1)$. This measure is referred to as the *sensitivity* of the model. Sensitivity alone is an incomplete picture of model performance, because if the mean of G is high, even a näive model that always predicts "$G = 1$" will obtain a high sensitivity score. However, this näive forecast will rank low on the *specificity* scale, which is the frequency of forecasts where "$G = 0$" is correctly predicted, or $P(\hat{G} = 0 \mid G = 0)$. When a low threshold is used, models will achieve a high sensitivity but a low specificity score. A high threshold implies low sensitivity but high specificity (Hsieh and Turnbull, 1996). To avoid the threshold-dependency problem, one can deem Model A superior in forecasting ability to Model B if it has a higher sensitivity and specificity at every threshold value.

Receiver-operator curves (ROCs) measure forecast accuracy of discrete dependent variables. ROCs are attained by calculating the sensitivity (percentage of correct "$G = 1$" forecasts) and specificity (percentage of correct "$G = 0$" forecasts) for each possible threshold. An ROC is then a plot of sensitivity on the y-axis against specificity on the x-axis for all thresholds. The ROC will have a negative slope, will be nonnegative, and the formula for the area underneath the ROC has an upper-bound value of $\sqrt{2}$. An illustration is given in figure 1, where one model's ROC clearly dominates another. The process of picking Model A over Model B if A's ROC always lies above B's ROC is referred to in this paper as the ROC dominance (ROCD) criterion.[2]

In some instances ROCs will cross, leading to an ambiguous model ranking using the ROCD criterion. In these cases, to attain an unambiguous ranking, the model with the largest area underneath its ROC can be chosen. This area is obtained by performing integration of the distance from the origin to each point on the ROC over all thresholds, as demonstrated in figure 1. This is referred to as the generalized ROC (GROC) criterion (Reiser and Faraggi, 1997). Recent advances have made ROCs easier to use, as they can be estimated as smooth curves directly from data using maximum likelihood (Hsieh and Turnbull, 1996; Blume, 2002), and statistical tests are available for distinguishing significant differences in ROCs (Reiser and Faraggi, 1997; Venkatraman and Begg, 1996).

[2] This term is chosen by the authors, as no unique name for this approach is offered in the literature.



Note: Sensitivity is the percentage of correct $G = 1$ forecasts, and specificity is the percentage of correct $G = 0$ forecasts, given a particular threshold. A superior forecasting model will have a higher sensitivity for every value of specificity, and vice versa. The model whose ROC lies completely above another is deemed the superior model. Consider the two models at a threshold of 0.3. At this threshold, the superior model has a higher percentage of correct $G = 1$ and $G = 0$ forecasts. Thus, at that threshold, it is a better model. If the curves cross, one can pick the model with the largest area underneath the ROC. The area can be measured by the integral of the ray drawn above over all threshold values.

Figure 1. Illustration of receiver-operator curves (ROCs)

The term "curve" is actually deceiving, since the functions generating ROCs are not necessarily continuous. Let $\hat{P}_t$ be the predicted probability $G_t = 1$, where t refers to an out-of-sample forecast. Also, let c be the threshold where we predict $G_t = 1$ when $\hat{P}_t \geq c$. The point on the ROC when $c = 0.5$ is represented by:

$$\left\{ \left[\frac{\sum_t (1 - G_t) I\left[\hat{P}_t < 0.5\right]}{\sum_t (1 - G_t)} \right], \left[\frac{\sum_t G_t I\left[\hat{P}_t \geq 0.5\right]}{\sum_t G_t} \right] \right\},$$

where $I[\cdot]$ is an indicator equaling one if its argument is true, and zero if false. For continuous ROCs, the area underneath the ROC can be calculated as:

$$GROC = \int_0^1 \sqrt{\left[\frac{\sum_t (1 - G_t) I\left[\hat{P}_t < c\right]}{\sum_t (1 - G_t)} \right]^2 + \left[\frac{\sum_t G_t I\left[\hat{P}_t \geq c\right]}{\sum_t G_t} \right]^2} \, dc. \tag{1}$$

ROCs are not necessarily continuous. Imagine a model that perfectly predicts whether a variable will take the value zero or one, regardless of the threshold. All points on this ROC will lie at the point (1, 1). However, the absence of a continuous curve does not prohibit integration of (1), nor does it preclude (1) from being a measure of forecast accuracy. Integration of (1) for this perfect model yields a value of $\sqrt{2}$ and is the highest possible GROC value.

One advantage of ROCs is that they allow visual inspection of forecast performance. Moreover, since the value of the GROC criterion given in (1) must lie between zero and $\sqrt{2}$, the measure $GROC/\sqrt{2}$ is similar to the coefficient of determination in that it lies between zero and one. The GROC criterion is an absolute measure of performance, allowing one to compare forecast performance across different data and models.

A second potentially useful criterion for judging forecast performance of discrete dependent variables is based on the Kullback-Leibler Information Criterion, which selects models closest to the true data-generating process (Stone, 1977; Shao, 1993). This criterion selects the model with the highest log-likelihood function observed at out-of-sample observations.[3] Originally, this was referred to as cross-validation, but over time "cross-validation" has taken on numerous definitions. For clarity, we refer to this approach as the out-of-sample log-likelihood function (OSLLF) approach. In a forthcoming study, Norwood, Roberts, and Lusk illustrate the usefulness of OSLLFs in selecting yield distributions, and Norwood, Ferrier, and Lusk (2001) report that the OSLLF has been found to select true models with a higher frequency than many competing criteria.

The OSLLF criterion may be especially desirable in the discrete variable case because it can rate forecasting ability without requiring the specification of a threshold. For variables that can only take the values zero or one, the OSLLF is calculated as:

$$OSLLF = \sum_{t=1}^{T} (1 - G_t)\ln\left[1 - \hat{P}_t\right] + \sum_{t=1}^{T} G_t \ln\left[\hat{P}_t\right]. \tag{2}$$

Evaluating forecasts using log-likelihood functions preserves information on a model's confidence which would be lost when using mean squared error. For example, one could forecast "$G = 1$" whenever $\hat{P}_t > 0.5$ and evaluate the mean squared error. However, this assigns a correct forecast of $\hat{P}_t = 0.51$ the same score as a correct forecast of $\hat{P}_t = 0.99$, when the second forecast should be scored higher. The OSLLF criterion accounts for differing levels of model confidence by giving the first forecast a score of ln(0.51) and the second forecast a higher score of ln(0.99). Contrary to the ROCs, an OSLLF does not provide a visual representation of forecast accuracy and is not an absolute measure of performance. The OSLLF values from different data cannot be compared. However, evidence is provided below that the OSLLF criterion is a better measure of relative performance between models using the same data.

The following section shows that the predictive power of a model with a discrete dependent variable depends on how $\hat{P}_t$ behaves when the dependent variable is one and when it is zero. A concept of divergent distributions is introduced, where divergence is a measure of the distance between the distributions of $\hat{P}_t$ when the dependent variable is one and when it is zero. Predictive power is shown to be directly related to the degree of divergence. We then demonstrate that the ROCD, GROC, and OSLLF criteria are all measures of divergence with similar statistical properties.

---

[3] "Closeness" here is defined as the logarithm of a candidate model's likelihood function value minus the logarithm of the true model's likelihood function value. The Kullback-Leibler Information Criterion states that models with higher expected log-likelihood function values contain greater information. Models are often estimated by maximizing a log-likelihood function. If in-sample observations are used, the likelihood function will be higher than its expected value due to the fact that some of the observations are used for parameter estimation (Akaike, 1972; Sawa, 1978). To correct for this bias, one can provide a penalty that reduces the in-sample likelihood function value according to the number of parameters, or employ out-of-sample observations, where no penalty is needed.

Divergent Distributions, Receiver-Operator Curves, and Log-Likelihood Functions

When forecasting whether a variable G_t will equal zero or one, an index is generally used where a higher index value indicates a greater probability $G_t = 1$. Conversely, a lower index value suggests a greater probability $G_t = 0$. This index at observation t is denoted by $\hat{P}_t$ and is assumed to lie between zero and one. In economics, the index is usually generated from a model such as a logit model. In the medical profession, the index is often the direct measurement of a medical test, such as a cholesterol level.

If a model has any predictive ability, the value of $\hat{P}_t$ will tend to be larger when $G_t = 1$ than when $G_t = 0$. For example, if $G_t = 1$, the average value of $\hat{P}_t$ should be higher than when $G_t = 0$, i.e., $E(\hat{P}_t \mid G_t = 1) > E(\hat{P}_t \mid G_t = 0)$. Let $f_0(\hat{P}_t)$ be the probability distribution of $\hat{P}_t$ when $G_t = 0$, and $f_1(\hat{P}_t)$ be the distribution when $G_t = 1$. If f_0 and f_1 are identical, the model has no predictive power. Moreover, models where f_0 and f_1 are further apart will have more predictive ability. Hereafter, the distance between f_0 and f_1 is referred to as "divergence," where greater divergence implies greater distance.

Figure 2 illustrates divergence for two hypothetical models. The distributions are close together for Model B, indicating little divergence. In this case, Model B provides very little information on the true value of G_t. At a threshold of 0.5, where one forecasts "$G_t = 1$" if $\hat{P}_t > 0.5$, an incorrect forecast is almost as likely as a correct forecast. This is little improvement over a coin toss. Conversely, due to the large divergence for Model A, at a threshold of 0.5 all forecasts will be correct. The predictive power of a model stems directly from the degree of divergence between the distributions of $f_0(\hat{P}_t)$ and $f_1(\hat{P}_t)$.

At any particular threshold c, model sensitivity is described by

$$1 - F_1(c) = \int_c^1 f_1(\hat{P}_t)\, d\hat{P}_t.$$

This is the frequency $\hat{P}_t$ will exceed c when $G = 1$, and thus describes the frequency of correct "$G = 1$" forecasts at threshold c. Similarly,

$$F_0(c) = \int_0^c f_0(\hat{P}_t)\, d\hat{P}_t$$

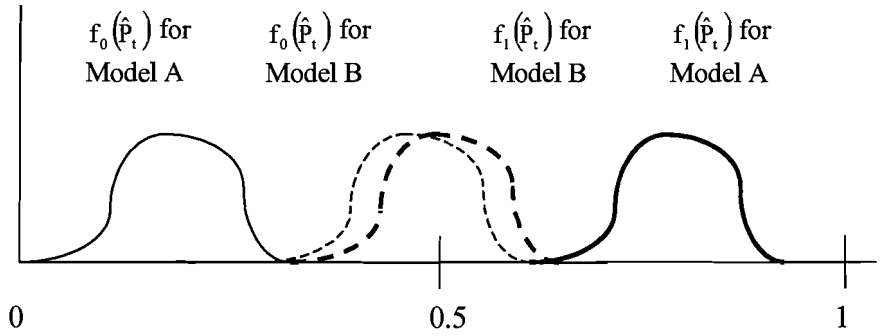
is the model specificity, which details the frequency of correct "$G = 0$" forecasts at threshold c. By definition, the true ROC is the set of points $\{F_0(c), 1 - F_1(c)\}$ for all values of c.

The GROC and the OSLLF criteria are measures of divergence. To demonstrate this, first consider the true GROC criterion value shown in (3):

$$\int_0^1 \sqrt{[F_0(c)]^2 + [1 - F_1(c)]^2}\, dc. \tag{3}$$

Greater divergence can be defined as a simultaneous increase in the value of $F_0(c)$ and a decrease in the value of $F_1(c)\,\forall\, c$. This essentially truncates $F_0(c)$ toward zero and $F_1(c)$ toward one. It is obvious that this would increase the value of (3), implying (3) measures divergence. The OSLLF also measures divergence. The expected value of the OSLLF can be written as:[4]

[4] The variable G_t is not viewed as a random variable here, because we are holding the set of observations used for forecasting constant. Instead, we are evaluating the statistical properties of a single model's forecasting ability at a fixed set of observations.



Note: $\hat{P}_t$ is the predicted probability G_t will equal one. The predicted probability G_t will equal zero is then $1 - \hat{P}_t$. The term $f_0(\hat{P}_t)$ is the probability distribution of $\hat{P}_t$ when $G_t = 0$, and $f_1(\hat{P}_t)$ is the probability distribution of $\hat{P}_t$ when $G_t = 1$.

Figure 2. Degree of divergence for two hypothetical models

$$(4) \qquad E[LLF] = \sum_t (1 - G_t) \int_0^1 \ln(1 - \hat{P}_t) f_0(\hat{P}_t)\, d\hat{P}_t + \sum_t G_t \int_0^1 \ln(\hat{P}_t) f_1(\hat{P}_t)\, d\hat{P}_t.$$

Truncation of $f_0(\hat{P}_t)$ toward zero can be achieved by decreasing the endpoint over which it is integrated by ε, while requiring it to still integrate to one.[5] Truncation of $f_1(\hat{P}_t)$ is obtained by increasing the beginning point over which it is integrated by ε, also requiring it integrate to one. Consider the partial effect this truncation has on the expected OSLLF value:

$$(5) \qquad \frac{dE[LLF]}{d\varepsilon} = \sum_t G_t \left(\frac{-\ln(\varepsilon) f_1(\varepsilon)}{\left[1 - \int_0^{\varepsilon} f_1(\hat{P}_t)\, d\hat{P}_t \right]} \right) + \sum_t G_t \left(\frac{\left[\int_{\varepsilon}^1 \ln(\hat{P}_t) f_1(\hat{P}_t)\, d\hat{P}_t \right] f_1(\varepsilon)}{\left[1 - \int_0^{\varepsilon} f_1(\hat{P}_t)\, d\hat{P}_t \right]^2} \right)$$
$$- \sum_t (1 - G_t) \left(\frac{\ln(\varepsilon) f_0(1 - \varepsilon)}{\left[1 - \int_{1-\varepsilon}^1 f_0(\hat{P}_t)\, d\hat{P}_t \right]} \right)$$
$$+ \sum_t (1 - G_t) \left(\frac{\left[\int_0^{1-\varepsilon} \ln(1 - \hat{P}_t) f_0(\hat{P}_t)\, d\hat{P}_t \right] f_0(1 - \varepsilon)}{\left[1 - \int_{1-\varepsilon}^1 f_0(\hat{P}_t)\, d\hat{P}_t \right]^2} \right).$$

Since ε lies in the (0, 1) range, ln(ε) will always be negative, making (5) positive—proving that greater divergence increases the expected OSLLF value.

This implies ROCs and OSLLFs are both measures of divergence. It does not imply they are equally desirable criteria. Next, we demonstrate that under a plausible assumption, the ROCD, GROC, and the OSLLF criteria will asymptotically provide identical model rankings. This assumption is referred to as the dual-divergence assumption.

---

[5] That is, if $f(X)$ is a probability density function with the support (0, 1), the integral $\int_0^1 f(X)dX$ must equal one. If $f(X)$ is truncated from below at η, the new integral will only equal one if it is multiplied by the constant $1 - \int_\eta^1 f(X)dX$, i.e.,

$$\int_0^{\eta} f(X)dX \,/\, 1 - \int_\eta^1 f(X)dX = 1.$$

When comparing two models in large samples, the dual-divergence assumption states that one model will always exhibit greater divergence than the other. Let the superscript i on the term $F_0^i(c)$ refer to Model i. The dual-divergence assumption requires that if Model A exhibits greater divergence when $G = 0$ ($F_0^A(c) > F_0^B(c) \forall c$), then Model A must also exhibit greater divergence when $G = 1$ ($F_1^A(c) < F_1^B(c) \forall c$). If the assumption does not hold, then Model A could exhibit greater divergence when $G = 0$ but less divergence when $G = 1$ compared to Model B, and it would be unclear which model displays greater total divergence.

Consider again the example of predicting quality grades in cattle. Suppose *Days on Feed* is the only variable determining whether a steer graded Choice. Further, suppose a steer grades Choice whenever *Days on Feed* $\geq$ 100. Assume that *Days on Feed* is measured with error. One cannot say with 100% certainty whether a steer will grade Choice given the measured *Days on Feed*, but instead must express the probability of grading Choice. A logit model estimating whether a steer grades Choice as a function of *Days on Feed* will specify $\hat{P}_t$ as a continuous function in the (0,1) interval. The function $f_0(\hat{P}_t)$ will contain mass over a series of points closer to zero, and the function $f_1(\hat{P}_t)$ will contain mass over points closer to one. The dual-divergence assumption requires that if the measurement error increases, $F_0(\hat{P}_t)$ decreases and $F_1(\hat{P}_t)$ increases at every $\hat{P}_t$. Both distributions move closer together.

Now, suppose *Days on Feed* can be measured perfectly. In this case, one can use the indicator function $\hat{P}_t = I[\textit{Days on Feed} \geq 100]$ to generate perfect forecasts. The functions $f_0(\hat{P}_t)$ and $f_1(\hat{P}_t)$ will now be centered with all their mass at zero and one, respectively. Divergence increases for both distributions $f_0(\hat{P}_t)$ and $f_1(\hat{P}_t)$ when moving from the approximating statistical model to the true deterministic model.

We believe this provides an accurate depiction of what happens when a model is replaced with another that better represents reality. The new model contains more information, and divergence increases for both $f_0(\hat{P}_t)$ and $f_1(\hat{P}_t)$. At the very least, this provides us with a useful metaphor for characterizing models with more or less information. This metaphor is utilized in the dual-divergence assumption.

Large Sample Properties

When calculating empirical ROCs, the empirical distributions $\hat{F}_0(\hat{P}_t)$ and $\hat{F}_1(\hat{P}_t)$ are used to calculate (3). Asymptotically, $\hat{F}_0(\hat{P}_t)$ and $\hat{F}_1(\hat{P}_t)$ will converge to $F_0(\hat{P}_t)$ and $F_1(\hat{P}_t)$ by definition. Consider Models A and B. The dual-divergent assumption implies that one model, say Model A, will display greater divergence and the two conditions in (6) will hold:

$$F_1^A(c) < F_1^B(c) \forall c \quad \text{and} \quad F_0^A(c) > F_0^B(c) \forall c. \tag{6}$$

From equation (6), Model A's ROC will always lie above Model B's ROC in large samples, and will be chosen under both the ROCD and the GROC criterion. Note that (6) implies:[6]

---

[6] Equation (6) uses the fact that, so long as Y is nonnegative and has an expected value less than infinity, $E(Y) = \int_0^\infty (1 - F(Y))dY$, where $F(Y)$ is the cumulative distribution function. This can be proven by integrating $\int_0^\infty Y dF(Y)$ using integration by parts.

$$(7)\quad \int_0^1 \left[1 - F_1^A(\hat{P}_t)\right] d\hat{P}_t = E^A(\hat{P}_t \mid G_t = 1) > E^B(\hat{P}_t \mid G_t = 1) = \int_0^1 \left[1 - F_1^B(\hat{P}_t)\right] d\hat{P}_t$$

and

$$\int_0^1 \left[1 - F_0^A(\hat{P}_t)\right] d\hat{P}_t = E^A(\hat{P}_t \mid G_t = 0) < E^B(\hat{P}_t \mid G_t = 0) = \int_0^1 \left[1 - F_0^B(\hat{P}_t)\right] d\hat{P}_t,$$

which states that the expected value of $\hat{P}_t$ is larger for Model A than Model B when $G_t = 1$, and is smaller for Model A when $G_t = 0$.

It is now proven that, asymptotically, Model A will be ranked higher using the OSLLF criterion as well. The difference in OSLLF values between Models A and B is specified as:

$$(8)\quad OSLLF_A - OSLLF_B = \sum_{t=1}^{T} G_t \left[\ln(\hat{P}_t^A) - \ln(\hat{P}_t^B)\right] + \sum_{t=1}^{T} (1 - G_t)\left[\ln(1 - \hat{P}_t^A) - \ln(1 - \hat{P}_t^B)\right].$$

According to Slutsky's theorem, (8) converges in probability to:

$$(9)\quad OSLLF_A - OSLLF_B = \sum_{t=1}^{T} G_t \left[\ln\left(E^A(\hat{P}_t^A \mid G_t = 1)\right) - \ln\left(E^B(\hat{P}_t^B \mid G_t = 1)\right)\right] + \sum_{t=1}^{T} (1 - G_t)\left[\ln\left(1 - E^A(\hat{P}_t^A \mid G_t = 0)\right) - \ln\left(1 - E^B(\hat{P}_t^B \mid G_t = 0)\right)\right].$$

Using the result from (7), the expression in (9) is unambiguously greater than zero, and Model A will asymptotically obtain a higher OSLLF value than Model B. This proves that, asymptotically, all three criteria will choose the same model.

Small Sample Properties

In small samples, or if the dual-divergence assumption does not hold, ROCs may cross. The ROCD criterion will then yield an ambiguous model ranking. In these cases, although the GROC and OSLLF criteria will provide an unambiguous ranking, they may not agree on the preferred model. This begs the question which of the two criteria is "better." We address this question using a simulation. Refer to figure 2, where divergence is illustrated for hypothetical Models A and B. It is obvious that Model A exhibits much greater divergence. In the simulation used, Model A is assumed to exhibit greater divergence than Model B, but the difference in divergence is very small. Thus, although Model A is truly superior, in a single small sample, Model B could easily appear superior and be chosen by the GROC and/or the OSLLF criteria. Using simulations, we calculate the percentage of times Model B is incorrectly chosen using each criteria. The method with the lowest percentage of incorrect choices is deemed a better detector of divergence.

The distributions $f_0^A(c)$, $f_1^A(c)$, $f_0^B(c)$, and $f_1^B(c)$ are assumed to be normal distributions truncated between zero and one. The means of $f_0^A(c)$ and $f_1^A(c)$ before truncation are assumed to be 0.3 and 0.7, while the means for $f_0^B(c)$ and $f_1^B(c)$ are 0.32 and 0.68, respectively. The standard deviation for all distributions before truncation is

0.1.[7] By this choice of parameters, Model A has greater divergence, but due to their similarities, Model B may be chosen in small samples. Since Model A exhibits greater divergence, it is said to be superior. In repeated samples it will provide better forecasts. The true frequency at which $G_t = 1$ is set to 0.7, and the sample size is 50. At each simulation, values of G_t are randomly chosen. If $G_t = 0$, values of $\hat{P}_t$ are randomly drawn from the distribution $f_0^A(\hat{P}_t)$ for Model A and $f_0^B(\hat{P}_t)$ for Model B. If $G_t = 1$, the values of $\hat{P}_t$ are randomly drawn from the distribution $f_1^A(\hat{P}_t)$ for Model A and $f_1^B(\hat{P}_t)$ for Model B. The random draws are then used to calculate the OSLLF value in (2). The area underneath the ROC is measured by the integral given in (1).

The preferred model at each simulation is the one with the largest OSLLF or GROC value. After 1,000 simulations, the OSLLF criterion chose the inferior model in 17% of simulations with a standard error of 0.0118, while the percentage for the GROC criterion was 23% with a standard error of 0.0133.[8] Although the criteria performed similarly, the simulations suggest the OSLLF criterion is slightly better at detecting divergence. This finding was robust across alternative means, standard deviations, and expected values for G_t. Table 1 presents simulation results under different parameter assumptions. The OSLLF outperformed the GROC regardless of the simulation setting.

In the next section, the two criteria are used to study a problem posed by Lusk et al. (2003) where a marketing strategy for fed cattle entailed forecasting whether cattle will grade Choice or better. Lusk et al. only considered one model for predicting Choice. In the discussion below, this model is compared against several other forms to determine if better forecasting models exist. The marketing simulation in Lusk et al. is repeated with a better forecasting model to estimate the monetary value of the ROC and the OSLLF criteria.

Forecasting Fed Cattle Quality Grades

A larger number of animals are being marketed on an individual basis, referred to as "selling on a grid," where they receive premiums and discounts for individual carcass and quality characteristics. Schroeder and Graff (2000) illustrated the economic value of producers accurately knowing their cattle quality and marketing them accordingly. Unfortunately, cattle quality is not known until after slaughter, and producers must use forecasts of quality characteristics to determine the optimal marketing strategy. Koontz et al. (2000) found that profits could be enhanced by forecasting quality grades and sorting animals according to optimal marketing dates. A number of observable factors, such as the number of days on feed, placement weight, genetics, etc., can be used to forecast cattle quality at slaughter. In addition to these measures, recent research has illustrated the ability of ultrasound measurements of ribeye area, backfat, and marbling to improve forecasts of cattle quality (Lusk et al., 2003).

This analysis seeks to determine whether the aforementioned model selection criteria can be used to identify superior forecasting models of cattle quality. We apply the model selection techniques to the data used by Lusk et al. (2003), who focused on the predictive

[7] Random draws from the truncated normal are performed using the acceptance-rejection method. Random numbers are generated from the normal distribution with the specified mean and standard deviation, but are only accepted if they lie between zero and one.

[8] It is worth noting that if the sample size is increased to 500, both percentages fall below 1%.

Table 1. Simulation Results for Small Sample Properties of OSLLF and GROC

Mean of				Standard	Frequency	Sample	% of Time Inferior Model Is Chosen by	
$f_0^A(c)$	$f_1^A(c)$	$f_0^B(c)$	$f_1^B(c)$	Deviation[a]	$G_t = 1$	Size[b]	OSLLF	GROC
0.30	0.70	0.32	0.68	0.1	0.7	30	28%	31%
0.30	0.70	0.32	0.68	0.1	0.7	100	6%	19%
0.30	0.70	0.32	0.68	0.1	0.3	30	13%	38%
0.30	0.70	0.32	0.68	0.1	0.3	100	8%	21%
0.30	0.70	0.32	0.68	0.2	0.7	30	42%	45%
0.30	0.70	0.32	0.68	0.2	0.7	100	23%	25%
0.20	0.80	0.32	0.68	0.1	0.7	30	0%	3%
0.20	0.80	0.32	0.68	0.1	0.7	100	0%	1%
0.30	0.70	0.305	0.695	0.1	0.7	30	39%	48%
0.30	0.70	0.305	0.695	0.1	0.7	100	35%	47%

Notes: $f_0^A(c)$ is the distribution of $\hat{P}_t$, the predicted probability $G_t = 1$, for Model A given that $G_t = 0$. Similarly, $f_1^B(c)$ is the distribution of $\hat{P}_t$ for Model B given that $G_t = 1$. Thus, the superscript refers to the model and the subscript indicates the true value of G_t. All simulations employed 1,000 iterations.

[a] The standard deviation for $f_0^A(c)$, $f_1^A(c)$, $f_0^B(c)$, and $f_1^B(c)$.

[b] Sample size refers to the number of forecasts used to calculate the out-of-sample log-likelihood function (OSLLF) and generalized receiver-operator curve (GROC) values.

power of ultrasound data. The primary determinant of profitability on a grid is whether an animal grades Choice or higher (hereafter, Choice). Lusk et al. used a logit model to predict whether an animal will grade Choice based on the several variables mentioned, including ultrasound measures. The authors demonstrated that predictions from the logit model incorporating ultrasound data could enhance revenue by $4.16/head over models which ignored ultrasound information. Lusk et al. also showed that if the model forecasts were 100% accurate, revenue would increase by $21.35/head.

The latter result exemplifies the potential economic value in determining better forecasting models. In the following, we seek to determine whether the model selection criteria can be used to identify models with superior forecasting ability, which in turn would result in greater economic value associated with ultrasound technology.

Let $G = 1$ if the quality grade is Choice or better, and $G = 0$ otherwise. In addition to the logit model used in Lusk et al., a probit model and neural network model are also used to estimate the probability $G = 1$. Moreover, different combinations of explanatory variables are evaluated for the logit and probit models. The probability of achieving a Choice or better grade was stated as a function of ribeye area (*REA*), backfat (*BF*), and marbling (*MAR*), each measured using ultrasound. Other attributes not measured by ultrasound are days on feed (*DOF*), placement weight (*PLWT*), and a dummy variable indicating whether the sire or dam was an Angus (*ANGUS*).

Lusk et al. evaluated the two sets of variables using a logit model. One form uses ultrasound variables and the other form does not:

(10) Variable Set 1: Probability $(G = 1) = f(DOF, PLWT, ANGUS)$;

(11) Variable Set 2: Probability $(G = 1) = f(REA, BF, MAR, DOF, PLWT, ANGUS)$.

Alternative specifications are also developed by letting $f(\cdot)$ be a logit or probit model or a neural network. For the logit and probit models, the following additional explanatory variables are considered:

(12) Variable Set 3: Probability $(G = 1) = f(REA, BF, MAR, DOF, PLWT, ANGUS, REA^2, MAR^2)$;

(13) Variable Set 4: Probability $(G = 1) = f(REA, BF, MAR, DOF, PLWT, ANGUS, REA*MAR)$;

(14) Variable Set 5: Probability $(G = 1) = f(REA, BF, MAR, DOF, PLWT, ANGUS, REA^2, MAR^2, REA*MAR)$.

This provides a total of 11 models. Estimation of probit and logit models was accomplished using standard maximum-likelihood procedures in MATLAB. The neural network model was a multilayer perceptron network with two hidden layers, which can be written as:

(15) Probability $(G = 1) =$

$$\hat{P}_t = F\left[\sum_{j=1}^{2} W_j f_j\left(w_{j,0} + w_{j,1}REA + w_{j,2}BF + w_{j,3}MAR + w_{j,4}DOF + w_{j,5}PLWT + w_{j,6}ANGUS\right) + W_0\right],$$

where W_j and $w_{j,i}$ denote parameters to be estimated, f_j is a symmetric logistic function, and F is a logistic function. The weights were estimated by maximizing the binomial log-likelihood function with a weight decay term as follows:

(16)
$$\max \sum_{t=1}^{T}(1 - G_t)\ln[1 - \hat{P}_t] + \sum_{t=1}^{T} G_t \ln[\hat{P}_t] - \lambda\left[W_0^2 + W_1^2 + W_2^2 + \sum_{j=1}^{2}\sum_{i=0}^{6} w_{j,i}^2\right].$$

In (16), λ is a weight decay coefficient used to prohibit the network from over-fitting the data, and is set equal to 0.005 (Chavarriaga, 2001). The weight decay term is not included when calculating the OSLLF value. The estimation, performed in MATLAB, used 100 different starting values with the nonlinear constraint $0.05 \leq \hat{P}_t \leq 0.95$.[9]

A total of 162 observations are available for estimation and forecasting. The forecasts are accomplished using grouped cross-validation where, for each validation, 27 observations are left out of the estimation and used for forecasting.[10] This procedure follows Zhang's (1991) suggestion that there be at least five validation groups. For the 162 forecasts, the OSLLF and the GROC values are calculated for each model and reported in table 2.

---

[9] Without this constraint, neural networks tend to set $\hat{P}_t$ equal to zero or one at one or more observations, which are outside the domain of the log-likelihood function.

[10] To illustrate, at the first iteration, the first 27 observations are removed from the 162 total observations. The model is estimated using observations 28–162, and is then used to forecast observations 1–27. At the second iteration, the model is estimated using observations 1–27 and 55–162, and is then used to forecast observations 28–54.

Table 2. Fed Cattle Quality Grade Forecast Evaluation Results

Model[a]	Average Out-of-Sample Log-Likelihood Function (OSLLF) Value[b] / [Rank][c]		Generalized Receiver-Operator Curve (GROC) Measure[d] / [Rank][c]	
Logit Using:				
Variable Set 1 (*logit1*)	-0.6692	[11]	0.9403	[9]
Variable Set 2 (*logit2*)	-0.6137	[5]	0.9485	[4]
Variable Set 3 (*logit3*)	**-0.5955**	**[1]**	**0.9527**	**[1]**
Variable Set 4 (*logit4*)	-0.6178	[6]	0.9470	[5]
Variable Set 5 (*logit5*)	-0.5972	[2]	0.9498	[2]
Probit Using:				
Variable Set 1 (*probit1*)	-0.6691	[10]	0.9400	[10]
Variable Set 2 (*probit2*)	-0.6239	[7]	0.9453	[6]
Variable Set 3 (*probit3*)	-0.6028	[3]	0.9495	[3]
Variable Set 4 (*probit4*)	-0.6301	[8]	0.9360	[11]
Variable Set 5 (*probit5*)	-0.6104	[4]	0.9433	[7]
Neural Network (*neural*)	-0.6308	[9]	0.9419	[8]

[a] Variable Set 1 is given by text equation (10), and Variable Set 5 is given by text equation (14).
[b] The OSLLF value divided by 162 forecasts.
[c] Numbers in brackets are the model rankings for each criteria. A rank of one indicates the best model, while a rank of 11 is the worst model.
[d] The GROC measure was calculated as (1) and is not divided by $\sqrt{2}$.

Model Selection Results

After the grouped cross-validation was used to obtain 162 out-of-sample forecasts, each model was ranked according to the OSLLF and the GROC criteria. As shown in table 2, both criteria agreed on the three highest ranked models and chose the logit model using variable set 3 (*logit3*) as the best forecaster. Models without ultrasound data (*logit1* and *probit1*) and the neural network (*neural*) performed poorly. In addition to comparing criteria in table 2, models can also be compared by plotting the ROCs, as shown in figure 3. The ROCs for *logit3* and *neural* exhibit ROC dominance over *logit1*, illustrating the contribution of ultrasound data to forecasts. Although *logit3* does not ROC-dominate *neural*, its ROC lies above that of *neural* most of the time.

In the Lusk et al. analysis, in-sample predictions from *logit2* were compared with in-sample predictions from *logit1* to estimate the returns from ultrasound data. Here, we are interested in determining how much returns might increase if ultrasound data were used in conjunction with a better forecasting model. To accomplish this, the cattle marketing simulation in Lusk et al. was repeated; however, instead of using in-sample predictions, we focus on out-of-sample predictions as would be the case in actual cattle marketing decisions. The simulation involved using forecasted quality characteristics to determine whether an animal should be marketed on a live weight, dressed weight, or grid basis. By measuring the increase in revenues using *logit3* instead of *logit2*, we can estimate the value of model selection criteria in cattle marketing decisions.

Due to space considerations and a desire to avoid redundancy, most simulation details are deferred to the Lusk et al. (2003) article. Lusk et al. used *logit2* to predict quality grades, as demonstrated by equation (1) in their article. We replaced *logit2* with the model that was ranked highest according to the OSLLF and GROC criteria—*logit3*.

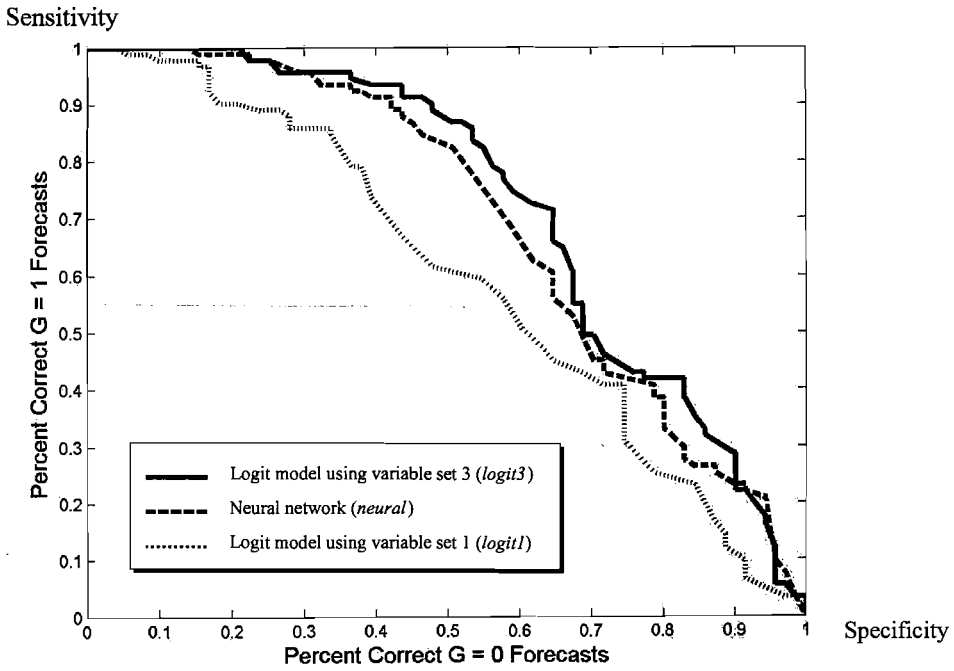


Figure 3. Receiver-operator curves for selected logit and neural network models

After replacing *logit2* with *logit3*, the exact simulation described in Lusk et al. was repeated to observe how marketing revenues change.

Simulation results indicate that the average revenue obtained using marketing methods based on predictions from *logit2* was $861.59/head, which is $2.59/head lower than the average revenue obtained using marketing methods based on predictions from *logit3* ($864.18/head). The marginal cost of using model selection criteria is relatively inexpensive. Thus, the $2.59/head benefit from model selection criteria is quite large, especially in comparison to the $4.16/head value of ultrasound technology reported by Lusk et al.

Discussion

This study is motivated by the frequent use of discrete variable models in economic analysis and the importance of forecast evaluation. Research on how one should evaluate forecasts of discrete dependent variables is rare, especially in the agricultural economics literature. This paper evaluates two methods for ranking forecasts of discrete dependent variables: receiver-operator curves (ROCs) and out-of-sample log-likelihood functions (OSLLFs). Both criteria are shown to be statistically valid measures of forecast performance, and share similar large and small sample properties. The theoretical prediction that the model selection criteria will frequently choose the same model is verified by an empirical analysis of cattle grades.

The theoretical and empirical examples here assume a single variable which takes on the values zero or one. The ROC and OSLLF criteria are easily extendible to multiple

dependent variables, such as multiple recreational site choice. In these cases, there will be a separate receiver-operator curve for each dependent variable. The OSLLF is more easily implemented by specifying a multivariate likelihood function. A multivariate function also incorporates information on error correlations across dependent variables, which should reap efficiency gains similar to those in seemingly unrelated regressions. This across-equation information is not present in the generalized ROC (GROC) criterion Given that simulations reveal a slight preference for the OSLLF criterion and it is easier to calculate, we recommend using the OSLLF for relative model comparisons when the dependent variable can take on multiple discrete outcomes.

[*Received December 2003; final revision received June 2004.*]

References

Akaike, H. "Information Theory and an Extension of the Maximum Likelihood Principle." *Proceedings of the Second International Symposium on Information Theory*, ed., N. Petrov and F. Csadki, pp. 267–281. Budapest: Akademiaai Kiado, 1972.

Ashley, R. "A New Technique for Postsample Model Selection and Validation." *J. Econ. Dynamics and Control* 22(1998):647–665.

Ashley, R., C. W. J. Granger, and R. Schmalensee. "Advertising and Aggregate Consumption: An Analysis of Causality." *Econometrica* 48(July 1980):1149–1167.

Blume, J. D. "Estimation and Covariate Adjustment of Smooth ROC Curves." Working paper, Center for Statistical Sciences, Brown University, August 2002.

Chavarriaga, R. "Modern Approaches to Neural Network Theory: Supervised Learning Algorithms." Ecole Polytechnique Federale de Lausanne, 19 February 2001. Online. Available at http://diwww.epfl.ch/~rchavarr/docs/ann_report.pdf.

Dorfman, J. H. "Bayesian Composite Qualitative Forecasting: Hog Prices Again." *Amer. J. Agr. Econ.* 80,3(August 1998):543–551.

Haener, M. K., P. C. Boxall, and W. L. Adamowicz. "Modeling Recreation Site Choice: Do Hypothetical Choices Reflect Actual Behavior?" *Amer. J. Agr. Econ.* 83,3(August 2001):629–642.

Heckman, J. J. "Sample Selection Bias as a Specification Error." *Econometrica* 47,1(January 1979): 153–161.

Hsieh, F., and B. W. Turnbull. "Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve." *Annals of Statistics* 24,1(February 1996):25–40.

Kastens, T. L., and G. W. Brester. "Model Selection and Forecasting Ability of Theory-Constrained Food Demand Systems." *Amer. J. Agr. Econ.* 78,2(1996):301–312.

Koontz, S. R., D. L. Hoag, J. L. Walker, and J. R. Brethour. "Returns to Market Timing and Sorting of Fed Cattle." *Proceedings of the 2000 NCR-134 Conference on Applied Price Analysis, Forecasting, and Market Risk Management*. Chicago, IL, April 2000. Online. Available at http://agecon.lib.umn.edu/.

Loomis, J. B., L. S. Bair, and A. Gonzalez-Caban. "Language Related Differences in a Contingent Valuation Study: English versus Spanish." *Amer. J. Agr. Econ.* 84,4(November 2002):1091–1102.

Loureiro, M. L., and S. Hine. "Discovering Niche Markets: A Comparison of Consumer Willingness to Pay for Local (Colorado Grown), Organic, and GMO-Free Products." *J. Agr. and Appl. Econ.* 34,3 (December 2002):477–487.

Lusk, J. L., R. Little, A. Williams, J. Anderson, and B. McKinley. "Utilizing Ultrasound Technology to Improve Livestock Marketing Decisions." *Rev. Agr. Econ.* 25,1(Spring/Summer 2003):203–217.

Norwood, B., P. Ferrier, and J. Lusk. "Model Selection Using Likelihood Functions and Out-of-Sample Performance." *Proceedings of the NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*, 2001. Online. Available at http://agecon.lib.umn.edu/.

Norwood, B., M. Roberts, and J. Lusk. "Ranking Crop Yield Models Using Out-of-Sample Likelihood Functions." *Amer. J. Agr. Econ.* 86,4(November 2004):1302–1343.

Piggott, N. E. "The Nested PIGLOG Model." *Amer. J. Agr. Econ.* 85,1(February 2003):1–15.

Reiser, B., and D. Faraggi. "Confidence Intervals for the Generalized ROC Criterion." *Biometrics* 53(June 1997):644–652.

Roberts, R. K., B. C. English, and J. A. Larson. "Factors Affecting the Location of Precision Farming Technology Adoption in Tennessee." *J. Extension* 40,1(February 2002). Online. Available at http://www.joe.org/joe/2002february/rb3.html.

Sawa, T. "Information Criteria for Discriminating Among Alternative Regression Models." *Econometrica* 46(1978):1273–1291.

Schroeder, T. C., and J. L. Graff. "Estimated Value of Increased Pricing Accuracy for Fed Cattle." *Rev. Agr. Econ.* 22(Spring/Summer 2000):89–101.

Shao, J. "Linear Model Selection by Cross-Validation." *J. Amer. Statis. Assoc.* 88, no. 422 (1993): 486–494.

Stone, M. "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion." *J. Royal Statis. Society, Series B* (Methodological) 39,1(1977):44–47.

Venkatraman, E. S., and C. B. Begg. "A Distribution-Free Procedure for Comparing Receiver Operator Characteristic Curves from a Paired Experiment." *Biometrika* 83,4(1996):835–848.

Zhang, P. "On the Distributional Properties of Model Selection Criteria." *J. Amer. Statis. Assoc.* 87, no. 419 (1991):732–737.