Applications of penalized binary choice estimators with improved predictive fit

D. J. Miller[1] & W.-H. Liu[2]
[1] *Department of Economics, University of Missouri, USA*
[2] *National Defense Management College, National Defense University, Taiwan, Republic of China*

Abstract

This paper presents applications of penalized ML estimators for binary choice problems. The penalty is based on an information theoretic measure of predictive fit for binary choice outcomes, and the resulting penalized ML estimators are asymptotically equivalent to the associated ML estimators but may have a better in-sample and out-of-sample predictive fit in finite samples. The proposed methods are demonstrated with a set of Monte Carlo experiments and two examples from the applied finance literature.
Keywords: binary choice, information theory, penalized ML, prediction.

1 Introduction

The sampling properties of the maximum likelihood (ML) estimators for binary choice problems are well established. Much of the existing research has focused on the properties of estimators for the response coefficients, which is important for model selection and estimating the marginal effects of the explanatory variables. However, the use of fitted models to predict choices made by agents outside the current sample is very important in practice but has attracted less attention from researchers. In some cases, the ML estimators may exhibit poor in-sample and out-of-sample predictive performance, especially when the sample size is small. Although several useful predictive goodness-of-fit measures have been proposed, there are no standard remedies for poor in-sample or out-of-sample predictive fit.

As noted by Train [1], there is a conceptual problem with measuring the in-sample predictive fit – the predicted choice probabilities are defined with respect to the relative frequency of choices in repeated samples and do not indicate the actual

WIT Transactions on Modelling and Simulation, Vol 43, © 2006 WIT Press
www.witpress.com, ISSN 1743-355X (on-line)
doi:10.2495/CF060201

probability that a respondent takes a particular action. Consequently, researchers should focus on the out-of-sample (rather than in-sample) predictive fit of an estimated binary choice model. Accordingly, Miller [2] derives a penalized ML estimator with improved out-of-sample predictive fit by adding a measure of in-sample predictive fit to the log-likelihood function. The purpose of this paper is to compare the ML and penalized ML estimators using examples from applied financial research.

2 ML and penalized ML binary choice estimators

2.1 ML Estimation of the binary choice model

For $i = 1, \ldots, n$ independent agents, we observe $Y_i = 1$ if agent i takes a particular action and $Y_i = 0$ otherwise. The binary decision process is represented by a latent utility model, $Y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$, where Y_i^* is the unobserved net utility associated with taking the action, $\mathbf{x}_i$ is a k-vector of individual–specific explanatory variables, $\mathbf{x}_i\boldsymbol{\beta}$ is the conditional mean component of Y_i^* that is common to all agents with characteristics $\mathbf{x}_i$, and ε_i is the mean-zero idiosyncratic error component of latent utility. The agent takes the action ($Y_i = 1$) if their net utility is positive ($Y_i^* > 0$), and the conditional probability that the agent takes the action is

$$\Pr\left[Y_i = 1 \mid \mathbf{x}_i\right] = \Pr\left[Y_i^* > 0 \mid \mathbf{x}_i\right] = \Pr\left[\varepsilon_i > -\mathbf{x}_i\boldsymbol{\beta} \mid \mathbf{x}_i\right] = F_\varepsilon\left(\mathbf{x}_i\boldsymbol{\beta}\right) \quad (1)$$

where the last equality follows if the latent error distribution is symmetric about zero. The two most commonly used model specifications for F_ε are the Normal $(0, \sigma^2)$ CDF (normit or probit model) and the Logistic$(0, \sigma)$ CDF (logit model). The response coefficients $\boldsymbol{\beta}$ are only defined up to scale, and the parameters are commonly identified under the normalization $\sigma = 1$.

Given probability model F_ε, the log-likelihood function is

$$\ell\left(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{x}\right) = \sum_{i=1}^{n} Y_i \ \ln\left[F_\varepsilon\left(\mathbf{x}_i\boldsymbol{\beta}\right)\right] + \sum_{i=1}^{n} \left(1 - Y_i\right) \ \ln\left[1 - F_\varepsilon\left(\mathbf{x}_i\boldsymbol{\beta}\right)\right] \quad (2)$$

The associated necessary conditions for the ML estimator of $\boldsymbol{\beta}$ are

$$\frac{\partial \ell\left(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{x}\right)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i' \left[\frac{Y_i \ f_\varepsilon\left(\mathbf{x}_i\boldsymbol{\beta}\right)}{F_\varepsilon\left(\mathbf{x}_i\boldsymbol{\beta}\right)} - \frac{\left(1 - Y_i\right) \ f_\varepsilon\left(\mathbf{x}_i\boldsymbol{\beta}\right)}{1 - F_\varepsilon\left(\mathbf{x}_i\boldsymbol{\beta}\right)}\right] = \mathbf{0} \quad (3)$$

where $f_\varepsilon\left(\mathbf{x}_i\boldsymbol{\beta}\right)$ is the PDF for the latent error process evaluated at $\mathbf{x}_i\boldsymbol{\beta}$. In general, the ML estimation problem does not have a closed-form (explicit) solution for the estimator of $\boldsymbol{\beta}$ (denoted $\widehat{\boldsymbol{\beta}}$), and numerical optimization tools must be used to compute the ML estimates for a given sample.

Under standard regularity conditions, the ML estimator is $\sqrt{n}$-consistent such that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ as $n \to \infty$ where $\boldsymbol{\beta}_0$ is the true parameter vector (up to arbitrary

scale). The ML estimators are also asymptotically normal as $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Delta}_0^{-1}\boldsymbol{\Xi}_0\boldsymbol{\Delta}_0^{-1}\right)$ where

$$\boldsymbol{\Delta}_0 \equiv \lim_{n\to\infty} \mathrm{E}\left[-n^{-1}\left.\frac{\partial^2\ell(\boldsymbol{\beta};\mathbf{Y},\mathbf{x})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] \tag{4}$$

$$\boldsymbol{\Xi}_0 \equiv \lim_{n\to\infty} \mathrm{E}\left[n^{-1}\left.\frac{\partial\ell(\boldsymbol{\beta};\mathbf{Y},\mathbf{x})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\left.\frac{\partial\ell(\boldsymbol{\beta};\mathbf{Y},\mathbf{x})}{\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] \tag{5}$$

If the binary choice model is correctly specified, the information matrix equality holds such that $\boldsymbol{\Xi}_0 = -\boldsymbol{\Delta}_0$ and the ML estimator is asymptotically efficient where $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Delta}_0^{-1}\right)$.

The predicted values for each Y_i in a fitted binary choice model are derived from the estimated choice probabilities under the step function

$$\widehat{Y}_i = \begin{cases} 0 & \text{if} \quad F_\varepsilon\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\right) < 0.5 \\ 1 & \text{if} \quad F_\varepsilon\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\right) \geq 0.5 \end{cases} \tag{6}$$

where $F_\varepsilon\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}\right)$ is the estimated choice probability conditional on $\mathbf{x}_i$. The standard diagnostic tool for describing in-sample predictive fit of a binary choice model is the prediction success table (see Maddala [3])

Actual Outcomes	Predicted Outcomes $\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$
$Y_i = 1$	φ_{11}	φ_{10}
$Y_i = 0$	φ_{01}	φ_{00}

Although prediction success tables are typically reported as counts of correct or incorrect predictions, the rows of the tables in this study are stated as the conditional frequency of predicted outcomes given the actual outcomes

$$\varphi_{00} = \frac{\sum_{i=1}^n (1-Y_i)(1-\widehat{Y}_i)}{n_0} \quad \text{and} \quad \varphi_{01} = \frac{\sum_{i=1}^n (1-Y_i)\widehat{Y}_i}{n_0} \tag{7}$$

$$\varphi_{10} = \frac{\sum_{i=1}^n Y_i(1-\widehat{Y}_i)}{n_1} \quad \text{and} \quad \varphi_{11} = \frac{\sum_{i=1}^n Y_i\widehat{Y}_i}{n_1} \tag{8}$$

where $n_0 = \sum_{i=1}^n (1-Y_i)$ is the number of observed zeroes, $n_1 = \sum_{i=1}^n Y_i$ is the number of observed ones, and $n_0 + n_1 = n$.

2.2 An information theoretic measure of predictive fit

To form a penalty function for predictive fit, Miller [2] considers the case of ideal in-sample predictive success for which the predicted outcomes $\widehat{Y}_i$ match the observed outcomes Y_i for all i. The ideal conditional outcomes for the prediction success table are denoted $\varphi^0_{00} = \varphi^0_{11} = 1$ and $\varphi^0_{01} = \varphi^0_{10} = 0$. Further, the goodness of in-sample predictive fit for an estimated model relative to the ideal case is measured as the difference between the estimated conditional distributions $\boldsymbol{\varphi}_j \equiv (\varphi_{j0}, \varphi_{j1})$ and the ideal distributions $\boldsymbol{\varphi}^0_j \equiv \left(\varphi^0_{j0}, \varphi^0_{j1}\right)$ for $j = 0, 1$. From information theory, one plausible measure of this difference is the Kullback–Leibler cross-entropy or directed divergence functional (see Kullback and Leibler [4])

$$I\left(\boldsymbol{\varphi}^0_j, \boldsymbol{\varphi}_j\right) = \varphi^0_{j0} \ln\left(\frac{\varphi^0_{j0}}{\varphi_{j0}}\right) + \varphi^0_{j1} \ln\left(\frac{\varphi^0_{j1}}{\varphi_{j1}}\right) = -\ln\left(\varphi_{jj}\right) \geq 0 \tag{9}$$

for each j. Under this divergence criterion, $I\left(\boldsymbol{\varphi}^0_j, \boldsymbol{\varphi}_j\right) = 0$ if the estimated conditional distributions coincide with the ideal case, $\varphi_{00} = \varphi_{11} = 1$ (i.e., zero predictive divergence). Otherwise, $I\left(\boldsymbol{\varphi}^0_j, \boldsymbol{\varphi}_j\right)$ increases as the observed and ideal cases diverge (i.e., there are more prediction errors).

Further, to make the penalty function suitable for estimation purposes, Miller [2] replaces the step function in eqn. (6) with a smooth approximation, $g\left(z, \theta\right) : [0, 1] \rightarrow [0, 1]$, that is continuously differentiable when θ is finite, monotonically increasing, and converge to the step function as $\theta \rightarrow \infty$. The associated approximation to the elements of the prediction success table are formed by replacing $\widehat{Y}_i$ with $g\left(F_\varepsilon(\mathbf{x}_i\boldsymbol{\beta}), \theta\right)$ in eqns. (7) and (8) above, and the approximated elements in the table are denoted φ^a_{jh}. The approximated predictive divergence functional is

$$I\left(\boldsymbol{\varphi}^0_j, \boldsymbol{\varphi}^a_j\right) = -\ln\left(\varphi^a_{jj}\right) \geq 0 \tag{10}$$

for each j. The properties of the penalized ML estimator hold for any $g(z, \theta)$ that satisfies these conditions, and the empirical examples presented in the next section are based on the scaled hyperbolic tangent function

$$g\left(z, \theta\right) = \frac{1 + \tanh(\theta(z - 0.5))}{2} \tag{11}$$

2.3 Sampling properties of the penalized ML estimator

Formally, the penalized ML objective function is

$$M(\boldsymbol{\beta}, \eta) = \ell\left(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{x}\right) + \eta \sum_{j=0}^{1} \ln\left(\varphi^a_{jj}\right) \tag{12}$$

and the penalized ML estimator is denoted $\widetilde{\boldsymbol{\beta}}_\eta$. The parameter $\eta \geq 0$ controls the trade-off between the log-likelihood and the predictive fit of the estimated binary

choice model. As η increases, predictive fit becomes more important in the estimation problem, and the penalized ML estimates are more strongly adjusted. The necessary conditions are

$$\frac{\partial \ell\left(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{x}\right)}{\partial \boldsymbol{\beta}} + \frac{\eta}{n_1 \varphi_{11}^a} \sum_{i=1}^{n} \frac{\partial g_i}{\partial F_i} \frac{\partial F_i}{\partial \boldsymbol{\beta}} Y_i - \frac{\eta}{n_0 \varphi_{00}^a} \sum_{i=1}^{n} \frac{\partial g_i}{\partial F_i} \frac{\partial F_i}{\partial \boldsymbol{\beta}} (1 - Y_i) = \mathbf{0} \quad (13)$$

where $g_i \equiv g\left(F_\varepsilon(\mathbf{x}_i \boldsymbol{\beta}), \theta\right)$ and $F_i \equiv F_\varepsilon(\mathbf{x}_i \boldsymbol{\beta})$. Note that eqn. (13) reduces to the standard ML necessary condition in eqn. (3) when $\eta = 0$. For $\eta > 0$, the necessary conditions for the penalized ML estimation problem may be numerically solved for $\widetilde{\boldsymbol{\beta}}_\eta$.

The necessary conditions stated in eqn. (13) may also be used to prove the following claims about the large-sample properties of $\widetilde{\boldsymbol{\beta}}_\eta$ for finite $\eta \geq 0$:

- **Proposition 1**: $\widetilde{\boldsymbol{\beta}}_\eta$ is $\sqrt{n}$-consistent such that $\widetilde{\boldsymbol{\beta}}_\eta \xrightarrow{p} \boldsymbol{\beta}^0$.
- **Proposition 2**: $\widetilde{\boldsymbol{\beta}}_\eta$ is asymptotically equivalent to $\widehat{\boldsymbol{\beta}}$.

Formal proofs are based on the differences in stochastic order of the terms in eqn. (13) where the log-likelihood term is $O_p(n)$ and the penalty terms are $O_p(1)$ (assuming $n_1/n = O(1)$). Thus, the penalty terms have smaller stochastic order than the log-likelihood component and do not affect the first-order asymptotic properties of the ML estimator.

2.4 Predictive properties of the penalized ML estimator

In small samples, the penalty in eqn. (12) only adjusts the estimated binary choice probabilities that are *local* or limited to a small neighborhood about the 0.5 threshold in the smoothed step function, $g(z, \theta)$. The penalized ML procedure is also *adaptive* and only corrects some of the ML prediction errors without inducing other in-sample prediction errors. To prove that the method may improve predictive fit, Miller [2] provides the following existence theorem:

- **Proposition 3**: There exists some $\eta > 0$ such that $\widetilde{\boldsymbol{\beta}}_\eta$ has weakly smaller approximated in-sample predictive divergence than $\widehat{\boldsymbol{\beta}}$.

He also demonstrates the locally adaptive character of $\widetilde{\boldsymbol{\beta}}_\eta$ by showing that the fitted binary choice probabilities are increased if $Y_i = 1$ and (i) η increases (predictive fit becomes more important), (ii) $F_\varepsilon(\mathbf{x}_i \widetilde{\boldsymbol{\beta}}_\eta)$ is closer to 0.5 (observations closer to the threshold are better candidates for adjustment), (iii) n_1 decreases (smaller samples require stronger adjustment), and (iv) φ_{11}^a decreases (less favorable predictive success for observations of $Y_i = 1$ require stronger adjustment). Finally, Miller [2] shows how to use a cross-validation (CV) estimator of the penalty weight parameter η. The value of η selected under the CV criterion is denoted $\widetilde{\eta}$ and is $O_p\left(n^{1/3}\right)$ such that $\widetilde{\boldsymbol{\beta}}_{\widetilde{\eta}}$ has the same first-order asymptotic properties as $\widetilde{\boldsymbol{\beta}}_\eta$ stated in Propositions 1 and 2.

3 Examples

In this section, two examples from the applied finance literature are used to illustrate the performance of the penalized ML logit estimator (with alternative values of $\eta > 0$) relative to ML logit ($\eta = 0$). Other plausible estimators are the ML probit estimator as well as semiparametric estimators such as the maximum score estimator introduced by Manski [5, 6] and the smoothed maximum score estimator developed by Horowitz [7]. Although the maximum score estimators are expected to have good predictive fit because the objective functions are the count of correctly predicted $Y_i = 1$ outcomes, the ML logit estimator has the best predictive fit among these traditional alternatives.

3.1 Example 1: mortgage data

The first example is based on data from Dhillon, Shilling, and Sirmans [8]. The dependent variable represents the decision of a mortgage applicant to accept a fixed rate or adjustable rate mortgage (ARM) (i.e., $Y_i = 1$ if ARM), and the data include $n = 78$ observations ($n_0 = 32$ and $n_1 = 46$). The set of explanatory variables includes the fixed interest rate, the difference between the fixed and variable rates, the Treasury yield spread, the ratio of points paid on adjustable versus fixed rate mortgages, the ratio of maturities on adjustable versus fixed rate mortgages, and the net worth of the applicant. The predictive success table for the fitted ML logit model is presented in the upper left corner of table 1. Although n is relatively small, the ML logit model provides reasonably good predictive fit for the fixed rate cases (83% correct) and the ARM cases (72% correct). The prediction success results for the optimal penalized ML estimator are stated in the lower left corner of table 1. Under $\widetilde{\eta} = 11$, the prediction success rates increase to over 93% for the fixed rate case and over 81% for the ARM case. The prediction success tables for other values of η are also presented in table 1, and the fitted penalized ML model achieves perfect predictive fit as η increases above 100.

To illustrate the locally adaptive character of the penalized ML estimator, the fitted ML logit (solid line) and penalized ML logit choice probabilities (circles) are presented in figure 1. The observations are the ordered ML logit predictions $F_\varepsilon(\mathbf{x}_i\widehat{\boldsymbol{\beta}})$ so that outcomes below the 0.5 threshold are $\widehat{Y}_i = 0$ and outcomes above the line are $\widehat{Y}_i = 1$. The penalized ML logit predicted values (circles) are vertically shifted away from the solid line to reflect the locally adaptive changes in the ML logit probabilities. Note that the adjustments are small in cases with strong predictions (i.e., $F_\varepsilon(\mathbf{x}_i\widehat{\boldsymbol{\beta}}) < 0.2$ or $F_\varepsilon(\mathbf{x}_i\widehat{\boldsymbol{\beta}}) > 0.8$), and most of the adjustments to the ML logit outcomes are restricted to outcomes in a neighborhood of 0.5. In the figure, the five observations marked with 'plus' symbols were initially predicted as $\widehat{Y}_i = 0$ under the ML logit model but were corrected to $\widetilde{Y}_i = 1$ under the penalized ML procedure. Further, the three 'minus' cases were initially predicted as $\widehat{Y}_i = 1$ but were corrected under the penalized ML logit model. These eight corrected predictions account for the gain in predictive fit reported in table 1 (0.8261 + 5/46

Table 1: Prediction success tables for Examples 1 and 2.

	Example 1: Mortgage Data			Example 2: Credit Data		
	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	η	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	η
$Y_i = 1$	0.8261	0.1739	0	0.9029	0.0971	0
$Y_i = 0$	0.2812	0.7188		0.6300	0.3700	
$Y_i = 1$	0.9348	0.0652	25	0.9943	0.0057	200
$Y_i = 0$	0.1250	0.8750		0.2267	0.7733	
$Y_i = 1$	0.9348	0.0652	75	1.0000	0.0000	500
$Y_i = 0$	0.0312	0.9688		0.1233	0.8767	
$Y_i = 1$	1.0000	0.0000	101	1.0000	0.0000	3223
$Y_i = 0$	0.0000	1.0000		0.0000	1.0000	
$Y_i = 1$	0.9348	0.0652	$\widetilde{\eta} = 11$	0.9771	0.0229	$\widetilde{\eta} = 88$
$Y_i = 0$	0.1875	0.8125		0.3100	0.6900	
	$n_1 = 46$	$n_0 = 32$		$n_1 = 700$	$n_0 = 300$	

= 0.9348 for $Y_i = 1$ and 0.7188 + 3/32 = 0.8125 for $Y_i = 0$). Also, note that there are four observations among these outcomes that were correctly predicted and were not adjusted due to the adaptive character of the penalized ML estimator.

3.2 Example 2: credit data

Credit scoring models are used to predict the potential success or failure of a borrower to repay a loan given the type of loan and information about the borrower's credit history. Hand and Henley [9] note that lenders increasingly rely on statistical decision tools for credit scoring due to the large increase in loan applications and the limited number of experienced credit analysts. Fahrmeir and Tutz [10] provide a set of credit scores assigned by experienced loan analysts to $n = 1,000$ (with $n_1 = 700$ and $n_0 = 300$) individual loan applicants in southern Germany. The dependent variable is the credit risk of a loan applicant ($Y_i = 1$ for a good credit risk), and the explanatory variables include an indicator of the applicant's relationship with the lender, the level of the applicant's checking balance, the loan duration, the applicant's credit history, the type of loan (private versus professional), and an indicator of the applicant's employment status. The predictive success table for the fitted ML logit model appears in the upper right corner of table 1, and the predictive fit is relatively good for good-risk applicants (i.e., $Y_i = 1$) but is quite poor for the poor-risk cases. The predictive success table for the optimal penalized ML logit estimator appears in the lower right corner of table 1, and the predictive fit in both categories is improved relative to ML logit. The results for other values

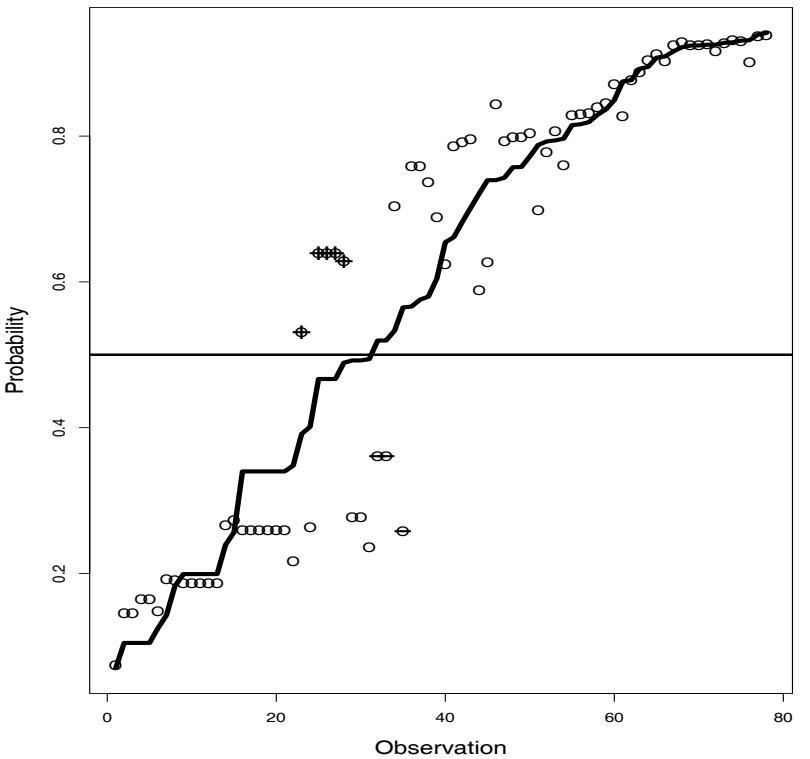

Figure 1: ML and optimal penalized ML logit predictions, Example 1.

of η are also reported in table 1, and the penalized ML logit estimator achieves perfect predictive fit for $\eta \geq 3,223$.

3.3 Out-of-sample predictive performance

Although Henley and Hand [11] show that the ML logit estimator is among the most accurate methods for predicting poor credit risks, lenders may achieve additional gains if they can further reduce the potentially large costs of making poor loans. To examine the predictive performance of the ML logit and penalized ML logit estimators, a bootstrap procedure is used to estimate the expected in-sample and out-of-sample predictive success tables given the data for Example 2. For each of $m = 5,000$ replications, $\bar{n} < n$ elements are drawn at random from the $n = 1,000$ observations, and the ML logit and optimal penalized ML logit parameter estimates are computed from the remaining $n - \bar{n}$ observations. The specified levels of the out-of-sample observation counts, $\bar{n} \in \{100, 150, 200, 250\}$, repre-

Table 2: In-sample and out-of-sample predictive success for Example 2.

Optimal Penalized ML Logit Estimator					
	In-Sample		Out-of-Sample		
	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	$\bar{n}$
$Y_i = 1$	0.9825	0.0175	0.6457	0.3543	100
$Y_i = 0$	0.3038	0.6962	0.2444	0.7556	
$Y_i = 1$	0.9837	0.0163	0.7119	0.2881	200
$Y_i = 0$	0.2996	0.7004	0.3182	0.6808	
$Y_i = 1$	0.9848	0.0152	0.7972	0.2028	400
$Y_i = 0$	0.2939	0.7061	0.4228	0.5772	
$Y_i = 1$	0.9861	0.0139	0.8749	0.1251	600
$Y_i = 0$	0.2876	0.7124	0.6023	0.3977	
Maximum Likelihood Logit Estimator					
	In-Sample		Out-of-Sample		
	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	$\widehat{Y}_i = 1$	$\widehat{Y}_i = 0$	$\bar{n}$
$Y_i = 1$	0.9097	0.0903	0.9057	0.0943	100
$Y_i = 0$	0.6410	0.3590	0.6488	0.3512	
$Y_i = 1$	0.9100	0.0900	0.9048	0.0952	200
$Y_i = 0$	0.6380	0.3620	0.6450	0.3550	
$Y_i = 1$	0.9098	0.0902	0.9052	0.0948	400
$Y_i = 0$	0.6356	0.3644	0.6387	0.3613	
$Y_i = 1$	0.9099	0.0901	0.8988	0.1012	600
$Y_i = 0$	0.6331	0.3669	0.6323	0.3677	

sent 10%, 20%, 40%, and 60% of the total observations in the data set. For each $\bar{n}$ and simulation trial $j = 1, \ldots, m$, the fitted ML logit and penalized ML logit models are used to predict the $n - \bar{n}$ in-sample and $\bar{n}$ out-of-sample bootstrap observations. The in-sample and out-of-sample prediction success tables are computed for each bootstrap trial, and the expected values of the tables are estimated by the sample averages of the replicated predictive success tables.

The bootstrap simulation results are reported in table 2. The in-sample and out-of-sample results for the ML logit estimator are quite close to the prediction success tables reported in table 1. For the optimal penalized ML logit estimator, the in-sample predictive success results are also quite comparable to the outcomes reported in table 1. As expected, the out-of-sample predictive fit is not as good for the good-risk category ($Y_i = 1$), and the ML logit estimator has better predic-

tive success. However, as noted above, the key decision error to avoid is offering a loan to a poor credit risk. For the poor-risk case ($Y_i = 0$), the optimal penalized ML logit estimator exhibits uniformly better predictive success, especially as the amount of in-sample information used to form the out-of-sample predictions increases relative to $\bar{n}$. In particular, the prediction success rate for poor credit risks is more than double the rate achieved by ML logit when $\bar{n}/n$ is only 10%. Given that the credit databases available for in-sample model estimation may be very large relative to the number of credit applications, the bootstrap evidence suggests that penalized ML logit may have significant advantages relative to ML logit in reducing the costs of extending credit to risky borrowers.

References

[1] Train, K., *Discrete Choice Methods with Simulation*. Cambridge University Press: New York, 2003.
[2] Miller, D., Penalized ML estimators of binary choice models with improved predictive fit. working paper, University of Missouri, 2006.
[3] Maddala, G.S., *Limited-Dependent and Qualitative Variables in Economics*. Cambridge University Press: New York, 1991.
[4] Kullback, S. & Leibler, R., On information and sufficiency. *Annals of Mathematical Statistics*, **22**, pp. 79–86, 1951.
[5] Manski, C., Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, **3**, pp. 205–28, 1975.
[6] Manski, C., Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics*, **27**, pp. 313–34, 1985.
[7] Horowitz, J., A smoothed maximum score estimator for the binary response model. *Econometrica*, **60**, pp. 505–31, 1992.
[8] Dhillon, U., Shilling, J. & Sirmans, C., Choosing between fixed and adjustable rage mortgages: a note. *Journal of Money, Credit, and Banking*, **19**, pp. 260–7, 1987.
[9] Hand, D. & Henley, W., Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, **160**, pp. 523–41, 1997.
[10] Fahrmeir, L. & Tutz, G., *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag: New York, 1994.
[11] Henley, W. & Hand, D., A k-nearest-neighbor classifier for assessing consumer credit risk. *Statistician*, **45**, pp. 77–95, 1996.