Uncertain Times and Early Predictions of Bank Failure

Cullen F. Goenner*

Professor of Economics and Finance

University of North Dakota

293 Centennial Drive

Grand Forks, North Dakota

Phone: (701) 775-3852

Email: cullen.goenner@und.edu

JEL Classifications: G17, G21, G28

Key words: bank failure, Bayesian model averaging, prediction, banking crisis.

* I would like to thank the Editor (Richard Warr), an anonymous referee, Margie Tieslau, Ronald Gilbert, Dennis Petruska, Bob DeYoung, and seminar participants at the 2013 Missouri Valley Economic Association annual meeting in Kansas City for helpful comments.

Uncertain Times and Early Predictions of Bank Failure

Abstract

The Great Financial Crisis shows that bank failure in the United States, while rare, is a concern during uncertain times. Our interest is in the ability to predict future failures at the start of a crisis, when the recent past has few events on which to base our inferences. We show that policymakers using estimates based on the S & L crisis would identify in early 2009 that 2.0% of banks were in critical condition and 7.0% were unhealthy. This is comparable to the 1.7% of banks that failed within a year and the 3.9% of banks that would fail during the crisis.

1. Introduction

When Lehman Brothers filed for bankruptcy on September 15 of 2008, it became quite clear to everyone that the financial sector was again in crisis and that commercial banks in the United States and elsewhere were at a heightened risk of failure. The magnitude of the risk to banks was unclear at the time, as the previous fifteen years had seen, on average, less than eight failures a year in the United States and two years (2005, 2006) with none. Traditional early warning models of bank failure rely on the recent pattern of previous failures to base their predictions. With few recent failures to draw from, it would seem natural to look at a past period of crisis to guide US policymakers' predictions in late 2008 of the banks that would fail subsequent to the Great Financial Crisis. Previous research (Cole, Cornyn, and Gunther, 1995; Cole and Gunther, 1998) shows that the statistical models used by the Federal Reserve were quite accurate in predicting failures during the last major banking crisis of 1985 – 1992, referred to as the Savings and Loans (S & L) crisis. These models' predictions are based on the financial conditions of banks that are captured in their call reports and reflect measures of banks' capital adequacy, asset quality, management quality, earnings, liquidity, and sensitivity to market risk (CAMELS). Cole, Cornyn, and Gunther (1995) and Cole and Gunther (1998) find using model estimates based on data from an earlier year during the S & L crisis enables them to accurately

predict failures out-of-sample later during the same crisis. We consider whether failure patterns established during the S & L crisis are also useful in predicting failures at the start of the Great Financial Crisis.

Several studies (Cleary and Hebb, 2016; Cole and White, 2012; DeYoung and Torna, 2013; Jin, Kangaretnam, and Lobo, 2011; Ng and Roychowdhury, 2014) examine bank failures during the Great Financial Crisis. Similar to Cole, Cornyn, and Gunther (1995) and Cole and Gunther (1998), these studies use data exclusive to a particular crisis period for both their modeling and evaluation purposes. The contributions of these studies provide policymakers important insights into the factors that influence failure during the financial crisis. For example, Jin, Kangaretnam, and Lobo (2011) find that a bank's choice of auditor plays a role in failures during the crisis, as does treatment of loan loss reserves as regulatory capital (Ng and Roychowdhury, 2014). DeYoung and Torna (2013) find that a banks' exposure to non-traditional banking activities (insurance underwriting, securitization, investment banking, and venture capital) puts a bank at a higher risk of failure during the Great Financial Crisis. The issue from the policymaker's perspective is that the usefulness of these models' estimates is retrospective. In other words, they are useful in helping to understand the conditions that contribute to failure after the crisis period examined has passed, while their relevance to failures in a future crisis is yet unclear.

This paper follows the observation made by Cole and White (2012) that the same financial conditions influencing bank failures in 2009 also affected failures during the 1980s (Cole and Gunther, 1998; Lane, Looney, and Wansley, 1986; Thomson, 1992; Whalen, 1991). Our focus differs from Cole and White (2012) and others in that we explicitly use the previous failure experience of banks during the S & L crisis to build a prediction model, which is applied

to data observed by policymakers at the start of 2009 for out-of-sample predictions of failures in the period 2009-2015. In a sense, we attempt to identify initial conditions that serve as common risk factors of bank failures across different crises episodes, which allow for the creation of risk scores at the start of a subsequent crisis. If these risk scores are accurate, then policymakers in early 2009 would have a means of assessing risks to banks early at the start of the crisis. It would also suggest lessons learned from the Great Financial Crisis may help predict risk exposure during the next crisis.

We assess the risk of bank failure using both logit and survival models. Whereas others (Cole and Wu, 2010; Mayes and Stremmel, 2014) focus on the relative accuracy of the two models, we focus on the different information that each model provides policymakers at the start of a crisis. The logit model's estimates using data from the S & L crisis can correctly identify 106 of the 129 (82%) banks that actually fail within a year using year-end financial data from 2008 and a cutoff established in the earlier period. Our Cox proportional hazards model uses estimates from the S & L crisis and year-end data from only 2008 to predict risk scores and the survival experience of banks throughout the period 2009-2015. We show that our model's prediction of risk of failure is as accurate later in the crisis period as it is early in the crisis period, which indicates banks' initial financial conditions are good for assessing the risk to banks throughout a crisis period. We measure accuracy with a time dependent version of the area under the Receiver Operating Characteristic (ROC) curve (Heagerty and Zheng, 2005). Policymakers, using our Cox estimates and classification of risk scores, would identify in early 2009 that 2.0% of banks were in critical condition, 7.0% were unhealthy, and 91% were healthy. This is comparable to the 1.7% of banks that failed within a year, the 3.9% of banks that would fail later during the crisis, and the 94.3% that remained healthy throughout.

Our paper contributes to the literature, as we demonstrate models of bank failure are subject to the uncertainty of which variables to include in the model's specification. Cole, Cornyn, and Gunther (1995, p. 6) note that the Federal Reserve identifies approximately thirty financial variables most likely to affect the probability of bank failure.[1] From this list, Cole, Cornyn, and Gunther (1995) use stepwise selection to determine the subset of these variables relevant to failures during the S & L crisis. Bank failures, even during crises, are relatively rare events, and in such cases where there are many potential risk factors, predictions based on a single model specification are likely sensitive to variable selection (Volinsky, Madigan, Raftery, and Kronmal, 1997). We use techniques of Bayesian model averaging (BMA) to base our inferences on estimates that explicitly account for our uncertainty of the model's specification by averaging over the estimates from several different specifications. By accounting for model uncertainty, we improve, relative to stepwise selection, our out-of-sample predictions of the logit and Cox models.

2. Empirical models of bank failure

Policymakers prefer to avoid financial crises, so when they occur, researchers often focus their attention on a postmortem dissection to identify the conditions that caused the crisis. Studies of bank failure follow this pattern, as researchers try to uncover the unique characteristics of banks that fail during a crisis. Chen, Chidambaran, Imerman and Sopranzetti (2014), for example, undertake a detailed analysis of the debt maturity structure of Lehman Brothers to understand the failure of the investment bank in 2008. They observe that in 2008, Lehman was increasingly rolling their short-term debt forward, which would create liquidity

[1] However, Lane, Looney, and Wansley (1986, p. 516) note there is little agreement by regulators on which of these factors are most important.

problems if they were unable to refinance. Based on the firm's liability structure, the authors use changes in the firm's market equity value and volatility to predict at the end of March 2008 that Lehman was 73% likely to default by year-end 2008, relative to a 33% chance at the beginning of the year. Chen, Chidambaran, Imerman and Sopranzetti's (2014) results indicate fundamentals (i.e., high leverage, over reliance on short-term funding, and insufficient collateral) were the causes of Lehman's failure.

History shows the Great Financial Crisis and rise of bank failures will most likely not be the last. The triggers of the next crisis are also likely to differ from past episodes and are difficult to predict. What remains unclear, until a subsequent crisis, is whether the lessons learned and the patterns of failures based on fundamentals observed during the Great Financial Crisis can help inform decision making at the start of the next crisis. We explore this question by examining whether a prediction model based on the pattern of failures observed during the S & L crisis can predict failures during the Great Financial Crisis, when conditioning on banks' financial conditions at year-end 2008. The purpose of this analysis is to determine what policymakers, at the start of the crisis, could have inferred about the future risk of banks failing from the data available and from model estimates based on a past crisis.

In general, there are two empirical approaches to modeling bank failure.[2] The first approach models bank failure as a binary outcome, i.e., whether a bank fails in a subsequent

[2] Methods using machine learning (e.g., support vector machines and neural networks) are another approach to modeling bank failure that is being developed in the literature (Le and Viviani, 2018; Gogas, Papadimitriou, and Agrapetidou, 2018). Le and Viviani's (2018, 24) use of a neural network improves the accuracy of predicting US bank failure over a traditional logit regression model, though the difference is "not very big". The area under the receiver operating characteristic curve (ROC) is 79.6% for their logit model and 81.9% for their neural network. The logit model though outperforms the predictions of the support vector machine, where the area under the ROC is 71.5%.

period conditioning on a set of observables at a given point in time. These models use a cross-section of banks and estimate the probability of bank failure using either probit (Cole and Gunther, 1998; Cole, Cornyn, and Gunther, 1995; Wheelock, 1992) or logistic regression (Arena, 2008; Cole and White, 2012; Ng and Rowchowdary, 2014). In their early warning model, Cole and Gunther (1998) use bank financial data from year-end 1985 to predict whether a bank fails in the two year period 1986(Q2) – 1988(Q1). Their results indicate that lower capital ratios, declining asset quality, lower earnings, and less liquidity play significant roles in failures during the S & L crisis period.

The predictions from a binary response model are probabilities, $\hat{p}_i$. Therefore, one must adopt a classification rule, *c*, to distinguish between banks predicted to fail $(\hat{p}_i > c)$ from those that are not $(\hat{p}_i \leq c)$. We measure accuracy by the sensitivity and specificity of the model's classifications. Sensitivity measures the fraction of banks predicted to fail $(\hat{p}_i > c \mid y_i = 1)$ among the subset of banks that actually fail during the period examined, whereas specificity measures the fraction of banks predicted not to fail $(\hat{p}_i \leq c \mid y_i = 0)$ among the subset that do not fail. One may also use error rates to consider the model's classifications. A type – 1 error consists of a bank predicted not to fail which fails, whereas a type – 2 error consists of a bank that is predicted to fail but does not fail. Type-1 errors are more costly to banks, as they may result in costs from failures that could have been potentially avoided if predicted, while type – 2 errors entail diverting limited supervisory resources to a healthy bank. Mathematically the type – 1 error rate is equal to one minus the sensitivity, and the type – 2 error rate equals one minus the specificity. For a given model specification a tradeoff exists between the two types of errors,

i.e., reducing the cutoff to classify a failing bank reduces type – 1 errors at the expense of more type – 2 errors, whereas raising the cutoff has the opposite effect.

Cole and Gunther (1998) find that their probit model estimates from 1985, when applied to data from year-end 1987, result in out-of-sample failure predictions with a type-1 error rate of 9.8% for a type – 2 error rate of 10%. Using estimates from 1987 and data from 1989 the type – 1 error rate is 7.9% for a type-2 error rate of 10%. A plot of type-1 versus type – 2 error rates, across the range of classification cutoffs, allows for more general and visual comparison of the model's classifications. For example, Cole and Gunther (1998) show that their model specification estimated with financial data from call reports has lower type – 1 errors than a specification estimated with a bank's composite CAMELS rating over the range of type – 2 errors.[3] A closely related measure of performance is the receiver operating characteristic curve (ROC), which plots the tradeoff between the model's sensitivity relative to type – 2 errors. The area under the ROC curve (AUC) is a statistic that summarizes the model's predictive accuracy over the range of cutoffs and measures the probability a randomly selected failed bank has a higher predicted risk score than a bank that does not fail (control). A value equal to one indicates the model's predictions are able to completely discriminate between failed and non-failed banks, whereas a value equal to 0.5 indicates the predictions are no better than pure chance.

Logistic and probit regression models of bank failure have also been applied using panel data (Betz, Oprica, Peltonen, and Sarlin, 2014; Jin, Kangaretnam, and Lobo, 2011; Mayes and Stremel, 2014). An important aspect of bank failures is that once a bank fails, it cannot possibly

[3] CAMELS ratings assigned by regulators to banks during examinations are not made publicly available.

fail again. The econometric issue this poses when using panel data with failures is observations are no longer independent of each other over time, which results (Shumway, 2001) in biased and inconsistent estimates. Shumway (2001) shows this issue can be alleviated by correcting the standard errors of the multi-period logit model to account for the lack of independence, and that the resulting specification is equivalent to a discrete time hazard model. DeYoung and Torna (2013) and Cole and Wu (2010) use this approach in their analyses of bank failure.

An alternative, which we use, is to directly model the time to bank failure using a survival model. The dependent variable in a survival model is the time to failure, T_i, which is the difference in time (days, years, etc.) between when a bank becomes at risk of failure and when it either fails, or if it does not fail, the end of the study period. Observations in the latter case are said to be censored. Studies of bank failure (Brown and Dinc, 2005; Lane, Looney, and Wansley, 1986; Mayes and Stremel, 2014; Ng and Rowchowdary, 2014; Whalen, 1991; Wheelock and Wilson, 1995, 2000) using a survival model typically specify a Cox proportional hazards model.[4] Estimates from the model allow one to specify the hazard function at time *t*, which represents the rate of failure at a point in time, conditioning on failure having yet to occur. This differs from the logit model that considers the proportion of failures within a time period. The hazard of failure $h(t, x, \beta) = h_0(t)e^{x\beta}$ is a function of two terms.[5] The baseline hazard, $h_0(t)$, characterizes how the hazard of bank failure changes relative to time at risk, while $e^{x\beta}$ is a feature of the Cox model that characterizes how the hazard of bank failure depends on the control variables. The model is said to be semi-parametric in the sense that the baseline hazard's

[4] Arena (2008) instead uses a fully parametric Weibull regression survival model.

[5] The notation for the hazard model used here is similar to Hosmer, Lemeshow, and May (2008).

$h_0(t)$ dependence on time is unspecified and the coefficients enter the model linearly. The effects of the covariates on survival are evaluated with a transform of the coefficients using the hazard ratio (equation 1):

$$HR\left(t, x_1, x_0\right) = \frac{h_0(t)e^{x_1\beta}}{h_0(t)e^{x_0\beta}} = \frac{e^{x_1\beta}}{e^{x_0\beta}} = e^{(x_1 - x_0)\beta} \quad (1)$$

The hazard ratio (HR) is a measure of relative risk that does not depend on time and reflects the change in a ratio of rates from a given change in covariate *x*'s values, ($x_1 - x_0$).

Using the hazard function, one is also able to specify a survival function $S\left(t, x, \beta\right) = S_0(t)e^{x\beta}$, where the baseline survivor function $S_0(t)$ is a function of the cumulative hazard function $S_0(t) = e^{-H_0(t)}$. The survival function represents the probability of observing a survival time greater than time *t* and is mathematically related to the probability of observing a failure time less than *t*, i.e., $S\left(t, x, \beta\right) = 1 - F\left(t, x, \beta\right)$. One way to interpret the linear prediction of $x\beta$ in the survival function is as a risk score, $M_i = x_i\hat{\beta}$, such that banks with similar risk scores experience similar survival rates at different points in time.

We use a time dependent measure of sensitivity and specificity introduced by Heagerty and Zheng (2005) to compare the survival model's accuracy to that of the binary response model.[6] At any point in time, *t*, we observe the subset of banks still at risk that fail in the current period $(T_i = t)$, which are referred to as incident cases, and those that remain alive into the future $(T_i > t)$, i.e., dynamic cases. Our survival model predicts a bank will fail at time *t* if their

[6] The time dependent ROC and AUC analyses are conducted using the R-package risksetROC (Heagerty and Saha-Chaudhuri, 2012).

estimated risk score is greater than *c*, a user defined cutoff. Heagerty and Zheng (2005) define the incident sensitivity of the model's predictions as the proportion of banks that fail at time *t* (incident cases) with risk scores greater than *c*, i.e., $P(M_i > c \mid T_i = t)$. Dynamic specificity (Heagerty and Zheng, 2005) refers to the proportion of banks that remain alive at time *t* (dynamic cases), with risk scores less than or equal to the cutoff, i.e., $P(M_i \leq c \mid T_i > t)$. Incident sensitivity and dynamic specificity are time dependent in the sense that their values vary by follow up time *t*. Using the incident sensitivity and dynamic specificity measures, one is then able to generate ROC curves for different follow up times that span the range of cutoff scores and calculate measures of the AUC that vary with follow up time. A time dependent AUC(t) value allows for evaluating how the predictive performance of the model changes with time, which is important here, as our interest is in developing a model that uses only initial conditions to achieve both good short-term and long-term predictive performance during a crisis period.

Experience shows that resolving bank failures during a crisis can take years after an initial decline in financial condition. Shumway (2001, p. 102) notes that when sampling periods are long, controlling for time at risk is important, which is the case when trying to assess the risk of bank failure during a crisis. Therefore, we use estimates of the static logit model to predict failures in the short-run, i.e., within a year of our observed initial conditions. We rely on estimates from the Cox proportional hazards model that account for time at risk to model a bank's risk of failure throughout the entire crisis period. Our Cox model though does not use time-varying covariates, i.e., explanatory variables that change over time, as we condition our model's predictions only on the initial conditions that are observed at the start of a crisis.

The benefit of estimating the model using only initial conditions (i.e., time-fixed covariates) is we can generate predictions of the estimated survival profiles of banks based on

their different risk scores. This allows policymakers to assess, at the start of a crisis, the variation of risk exposure in future bank failures and across the banking system. With time-varying covariates the survivor function is no longer interpretable in this manner and one loses, more generally, the ability for such predictions, as the survival time in the future then depends on covariates that are not yet observed. In addition, we observe that our initial conditions are equally well suited to the prediction of failures early in the crisis as they are for later follow up times based on little variation of the time dependent AUC(t) over time.

3. Bayesian model averaging

Our analysis uses Bayesian model averaging to incorporate our uncertainty as to which dependent variables we should include in our model's specification. Rather than base inference on a single model specification, the estimates from Bayesian model averaging (BMA) use a weighted average of estimates from several specifications, with weights determined by the posterior support each receives from the data. BMA offers a theoretically appealing method of accounting for uncertainty in model specification as it avoids potential issues with "p-hacking" (Harvey, 2017), where researchers choose to report only the estimates from the model specifications that support their hypotheses. The result of which Harvey (2017) notes tends towards biased findings that are not robust under subsequent testing. By bringing a Bayesian approach to data mining, it has been shown (Raftery, Madigan, and Volinsky, 1995; Volinsky, Madigan, Raftery, and Kronmal, 1997) that BMA offers better predictive performance than other methods, such as stepwise techniques. We apply BMA to both logistic and Cox proportional hazards models and the discussion below generalizes to either type of model with differences between the two as noted.

For our purpose, a model specification is defined by a linear combination of variables. Both the Cox and logistic models are non-linear, though the variables enter the models in a linear fashion. The set of model specifications of interest here include each of the different linear combinations of variables that are identified as potentially relevant to predicting bank failure. Identifying *p* variables of interest implies there are then $K = 2^p$ different model specifications to consider. The weight given to each of the *K* different models' estimates are determined by the specification's posterior model probability. The posterior distribution of our model parameters (β) given the data (D = y, X) is equal to

$$P(\beta \mid D) = \sum_{k=1}^{K} P(\beta \mid M_k, D) P(M_k \mid D) \tag{2}$$

The first component of equation 2, $P(\beta \mid M_k, D)$, represents the posterior distribution of estimates from the different model specifications. Volinsky, Madigan, Raftery, and Kronmal (1997) show this distribution can be approximated by applying maximum likelihood to the *K* different models in the case of logit and Cox models. The second component, $P(M_k \mid D)$, is the posterior model probability, i.e., weights, which represents for a given model specification the posterior likelihood the specification is the true model that generates the data. The sum of the weights across models is equal to one. By Bayes' rule and the law of total probability the posterior model probability is

$$P(M_k \mid D) = \frac{P(D \mid M_k) P(M_k)}{\sum_{l=1}^{K} P(D \mid M_l) P(M_l)} \tag{3}$$

where $P(D|M_k)$ is the likelihood and $P(M_k)$ is the prior probability that model M_k is the true model. We assume a uniform prior such that each of our models under consideration is as

equally likely to be the true model as another.[7] If, for example, one had strong information that the next crisis would be driven by exposure to non-guaranteed private student loan debt, then we could incorporate this into our prior for inclusion of the variable in the model.[8] A uniform prior though is reasonable without strong prior beliefs. The PMP then simplifies to become

$$P(M_k \mid D) = \frac{P(D \mid M_k)}{\sum_{l=1}^{K} P(D \mid M_l)} \tag{4}$$

The integrated likelihood, also referred to as the marginal likelihood, is found by integrating over parameter vector β_k

$$P(D \mid M_k) = \int_{\beta_k} P(D \mid \beta_k, M_k) P(\beta_k \mid M_k) d\beta_k \tag{5}$$

where β_k is a vector of parameters, $P(D|\beta_k, M_k)$ is the likelihood, and $P(\beta_k|M_k)$ is the prior density of β_k under model M_k. Volinsky, Madigan, Raftery, and Kronmal (1997) suggest the integral in equation 5 can be approximated using the Laplace method by using a function of Schwarz's (1978) Bayesian information criterion.

$$\begin{aligned} P(D \mid M_k) &\approx \exp\left(-\frac{1}{2} BIC_k^{'}\right) \\ BIC_k^{'} &= BIC_k - BIC_0 = -LRT + p_k \log(N) \end{aligned} \tag{6}$$

BIC_k and BIC_0 are the values of the Bayesian information criterion for model specification k and the null model (constant only), respectively. The difference of which is equal to the likelihood

---

[7] Fernandez, Ley, and Steel (2001) note the assumption is common when there is not strong prior information to suggest otherwise. Raferty (1995) finds the assumption on priors has little impact on the posterior distribution.

[8] Sheila Bair, former Chairman of the FDIC (2006-2011), has been vocal (Nasiripour, 2016) in noting the next crisis may be driven by student loans, due to similarities between the current market for student loans and the pre-crisis mortgage market, whereby easy access to credit has led to higher tuition prices and over extended borrowers. Unfortunately, call reports of commercial banks aggregate federally guaranteed and private student loan debt with other consumer loans.

ratio statistic subtracted from the number of parameters in model *k*, p_k, multiplied by the natural log of the number of observations. For the logistic model, the number of observations in equation 6, *N*, is equal to the sample size (Raftery, 1995). In the case of the Cox model, one could use the number of units under observation, the total time of all units under observation, or the number of events, i.e., failures. The latter choice, which we use below, is preferred by Raftery, Madigan, and Volinsky (1995).

To test hypotheses under Bayesian model averaging one uses Bayes factors. A Bayes factor allows us to compare the evidence in favor of one hypothesis relative to another. Consider two hypotheses, H_0 and H_1, where we have prior beliefs as to their validity given by $P(H_0)$ and $P(H_1)$. Using Bayes rule, it can be shown that the odds of observing the null relative to the alternative are given by:

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_0)}{P(D \mid H_1)} \frac{P(H_1)}{P(H_0)} \tag{7}$$

The posterior odds of the null hypothesis being true is equal to the Bayes factor multiplied by the prior odds in favor of the null. If each hypothesis is a priori equally likely, then the posterior odds of the null is simply equal to the Bayes factor. A Bayes factor is an odds ratio of probabilities that can be converted into the probability that the null hypothesis is true.[9] Therefore, a Bayes factor of 20 is interpreted as the null hypothesis being 20 times more likely than the alternative, which corresponds to a 95% probability of the null being true and a 5% probability in favor of the alternative.

[9] The odds ratio equals = $\Omega(H_0 \mid D) = \frac{P(H_0 \mid D)}{1 - P(H_0 \mid D)}$ therefore $P(H_0 \mid D) = \frac{\Omega(H_0 \mid D)}{1 + \Omega(H_0 \mid D)}$

Bayes factors differ fundamentally in interpretation from p-values found in the Neyman-Pearson approach to statistics. A p-value measures the probability of observing an outcome in the data more extreme than what is assumed under the null hypothesis, and in a sense represents $P(D \mid H_0)$. A p-value of 0.05 indicates the null hypothesis is rejected 5% of the time, when it is in fact true, and yet it does not tell us the probability the null hypothesis is true $P(H_0 \mid D)$. An example follows by Edwards, Lindman, and Savage (1963, pp. 221-222) that highlights this distinction. The frequentist perspective is that by assuming the null hypothesis is true, the *t*-statistic from a two-tailed t-test, with many degrees of freedom, will exceed 1.96 2.5% of the time. Similarly, 0.5% of the time the value will exceed 2.58, which implies that 2% of the time the statistic will lie between 1.96 and 2.58, when the null hypothesis tested is true. It might appear the data strongly favor the alternative if the observed test statistic lies within this interval. Consider an alternative, when the null is false the statistic lies uniformly between the values -20 and 20. In this example the value of t lies between the values of 1.96 and 2.58 with probability 1.55%. Given the alternative, the data favor the null.

Bayes factors are a nice alternative to p-values, which can overstate the evidence against the null in large samples (Greene, 1997; Harvey, 2017; Leamer, 1978). Specification searches, such as stepwise regression, can further magnify this problem, resulting in statistically significant relations in simulated data that do not exist. Freedman (1983) demonstrates this point using data where the 100 observations in each of his 51 variables (y, $X_1 \ldots X_{50}$) are white noise, i.e., independent draws from the standard normal distribution. An exploratory regression is first run on all 50 explanatory variables. After removing the variables that are not significant at the 25% level, the model is re-run. Freedman (1983) finds that 6 of the 15 variables used in his second pass are significant at the 5% level, when by construction there is no relation between the data.

Repeating this same exercise for a different random sample, Raftery, Madigan, and Hoeting (1997) observe 10 of 18 variables in the second stage were significant at the 5% level, and 15 variables were selected using stepwise regression, with 10 statistically significant at the 5% level. Applying BMA to this simulated data, Raftery, Madigan, and Hoeting (1997) observe that the only specification averaged over was the null model, i.e., a specification containing only an intercept, which by construction is the model that generated the data. Basing our inferences on Bayes factors reduces the number of false positives that Harvey (2017) warns of and we are left with a better understanding of the underlying relations in our data and an explicit accounting for our uncertainty in the model's specification.

We apply BMA to the logit and Cox models using the R-package BMA (Raftery, Hoeting, Volinksy, Painter, and Yeung, 2018). To increase the speed of estimation, the routine narrows down the number of models to average over by eliminating specifications that receive little support from the data. These are specifications where the odds are more than twenty to one in favor of another model specification being the true model. Excluding these specifications has little impact on our inferences given the low weight each would receive if included. The routine provides estimates of the posterior means and standard errors of the coefficients, which can be easily compared to their single equation MLE counterparts. For each variable, the routine calculates the posterior probability that the coefficient is non-zero $P(\beta_k \neq 0 \mid D)$, referred to as the posterior effect probability (PEP). Raftery (1995) considers the statistical evidence of an effect to be weak, positive, strong, and very strong according to a commonly used rule of thumb based on Bayes factors of 1, 3, 20, and 150, which correspond to PEP values of .5, .75, .95, and .99 on the probability scale, respectively.

4. Data

We draw the control variables we use from the December Report of Condition and Income (call report) data provided by individual banks to the Federal Reserve, Federal Deposit Insurance Corporation, the Comptroller of the Currency, and are distributed by the Federal Financial Institutions Examination Council. Our sample includes commercial banks, state-chartered banks, and cooperative banks. We rely on theory and previous research to identify the list of financial variables most likely related to bank failure. Also guiding this choice is our desire to focus on the failure experience of banks during the S & L crisis for basing predictions of failures during the Great Financial Crisis. The twenty-five variables we use include the year-end call report items considered by the Federal Reserve for use in their Financial Institutions Monitoring System (FIMS) model (Cole, Cornyn, and Gunther, 1995) to predict banks' risk of failure.[10] This implies there are more than 33.5 million (2^{25}) different specifications one could consider. As Cole, Cornyn, and Gunther (1995) note, these variables were selected based on the Fed's review of the literature and their use in examination reports. Further, these measures were strong predictors of failures during the S & L crisis. Table 1 lists the measures, and an online appendix includes the complete definitions. Other than the banks' age and size, we scale each of the control variables by total assets. End-of-year (December) call report data for an institution become available six hours after their submission. Their submission is due no more than 30 calendar days after the report date which makes the December data available as of February 1 the following year.[11] To ensure policymakers are able to observe year-end data prior to failures

[10] The FDIC uses a similar variable in their statistical CAMELS off-site rating (SCOR) model to predict changes in CAMELS ratings (Collier, Forbush, Nuxoll, and O'Keefe, 2003). The series, RCFD1406, loans past due 30-89 days and still accruing interest is included in both the government's FIMS and SCOR models. The series is not used here as it is confidential for the period 1984-1990.

[11] See the call report instructions available on the FFIEC website for further discussion of the current reporting guidelines. https://www.ffiec.gov

occurring, we analyze failures beginning in February of the year following the availability of bank financial data.

[Insert Table 1 about here]

The FDIC's list of failed banks identifies whether a bank fails and the date of failure. At year-end 1984 there were 14,024 banks in our sample, and among these banks, 1,100 would fail (7.9%) during the period 2/1/1985 - 2/1/1994. Table 1 reports and tests the difference in mean financial conditions and other characteristics between failed and non-failed banks. Banks that failed during the S & L crisis have characteristics that are significantly different than their counterparts in 1984. They were generally younger banks that had weaker performing loans, less liquidity, a higher reliance on jumbo CDs and brokered deposits among their liabilities, and higher asset concentration in commercial real estate and C & I loans. Interestingly, failed and non-failed banks had similar levels of equity relative to total assets (8.9%), which reflects the overall weakness of the banking system subsequent to years of rising interest rates. We find no significant differences in federal funds purchased or sold, volatile liability expenses, and the shares of assets in either consumer or agriculture loans.

As of year-end 2008, there were 7,441 banks in our sample of which 422 failed (5.7%) in the period 2/1/2009 – 2/1/2015. We find similar generalizations in the Great Financial Crisis as to the S & L crisis. The failing banks in the Great Financial Crisis were younger, had weaker performing loans, less liquidity, and a higher reliance on jumbo CDs and brokered deposits among their liabilities. They also had a higher share of commercial real estate loans. Failed banks during the Great Financial Crisis also had lower shares of consumer and agriculture loans. However, failed banks in the Great Financial Crisis were significantly less capitalized (8.0%)

than their counterparts (11.5%) and were even less capitalized than banks during the S & L crisis (8.9%).

A few other generalizations appear between the two crisis periods. Banks in the later crisis period were less reliant on core deposits, as their share among total assets decreased among non-failed (failed) banks from 15.8% to 10.8% (16.8% to 6.3%). The reduction in demand deposits, along with reductions in reserve requirements over time, likely contributed to the decrease in cash held by banks. This reduction in cash was also part of a more general trend of banks holding fewer liquid assets – securities held as a share of assets declined from 28.4% to 20.3% (15.8% to 11.1%) among non-failed (failed) banks. There were also notable changes to the composition of banks' loan portfolios over time. The shares of C & I, consumer, and agriculture loans decreased, whereas the shares of commercial real estate and non-commercial real estate (omitted category) increased.

5. Results and discussion

5.1. Logit models with rolling year-ahead predictions

The goal of our analysis is to examine whether the bank failure experience observed during the S & L crisis can predict at year-end 2008 the failures observed during the Great Financial Crisis. Our first analysis uses logistic regression to try and identify those banks that are at imminent risk of failure at year-end 2008. Similar to the Federal Reserve's FIMS model, we use a cross-section of banks' financial data at year-end to estimate whether failure occurs in the following year. We use two approaches to select variables for the model's specification for comparison purposes. The first approach uses Bayesian model averaging and the latter, similar

to the FIMS model specification (Cole, Cornyn, and Gunther, 1995) applies stepwise selection, which we determine on the basis of Akaike's information criterion.[12]

For each of the years 1984-1991, we estimate a rolling prediction model of bank failures in the year ahead with the two variable selection approaches. That is, we use year-end data from 1984 to predict whether failure occurs between 2/1/1985 and 2/1/1986, and then use 1985 year-end data to predict failures between 2/1/1986 and 2/1/1987, and so on. Results (we report these in the appendix) from the BMA analysis reveal there exists uncertainty as to the model's true specification with the number of specifications averaged over ranging from 4 (1984) to 37 (1985), with an average of 18 specifications.[13] Even when uncertainty is minimal, e.g. 1984, the two approaches suggest that different factors are important to predicting failures. The model chosen by stepwise selection using the 1984 data we report in Table 2 includes these several measures that are not averaged over in the BMA model: federal funds sold (p-value 0.020), provisions for loan losses (p-value 0.085), and securities (p-value 0.002). It also includes age (p-value 0.003) and cash (p-value 0.026), which receive little support from the data under BMA. We observe that p-values from a single model specification tend to overstate the evidence of an effect in the presence of model uncertainty. Based on our BMA estimates across the years (see online appendix), we find evidence against there being an effect from federal funds purchased, volatile liability expense, and consumer loans, as each has a posterior effect probability (PEP)

[12] An alternative to stepwise selection to find the best single model specification is the use of a biased regression, such as the lasso approach. Lasso (Tibshirani, 1996) is similar to ridge estimation, but imposes different restrictions where many of the coefficient estimates are instead shrunk to equal zero, i.e., a subset of variables are used in the predictions. This adds the parsimony found in stepwise models with the potential of improved prediction over linear models. Lasso and ridge methods are typically reserved for high-dimensional data (thousands of variables) and result in highly biased coefficient estimates, which make their interpretation difficult (Goeman, 2010).

[13] An online appendix contains the logit model estimates from BMA and stepwise selection for each of the years 1985-1992.

less than 5%. The measures of cash, charge-offs, insider loans, and age each have PEPs less than 50%, which indicate they do not receive support for influencing failures.

[Insert Table 2 about here]

Our interest is primarily on the models' out-of-sample predictive performance. We therefore focus less on the potentially different marginal effects implied by the two approaches, i.e., differences in the coefficient estimates and their standard errors and are less concerned with issues of multicollinearity in our candidate regressors. Despite the observed differences in the specifications across the two approaches, the out-of-sample type – 1 versus type – 2 error rates are quite similar throughout the S & L crisis period (1984-1992). To be clear, the out-of-sample predictions we report in Table 3 for prediction year 1985 are determined using the model estimates from 1984, along with financial data from year-end 1985 to predict failure outcomes between 2/1/1986 and 2/1/1987. For a given type – 2 error rate of 10% there is a type – 1 error rate that ranges over time between 1.7% and 23.6% for BMA and ranges between 3.1% and 22.0% for the stepwise model (see Table 3). BMA has a lower type – 1 error rate in four of the eight rolling predictions and is equal in another year, relative to the stepwise model. The model estimated using 1987 year-end data has the lowest prediction error. In the left panel of Figure 1, we plot the errors from this specification and observe there is little definitive difference in the predictive discrimination of the two variable selection methods across the range of type – 2 errors.

[Insert Table 3 and Figure 1 about here]

An issue with rolling predictions described above is that they rely on failures in the previous period to make predictions in the next. That is, our model requires failure episodes to estimate the model and in the years (1993-2008) following the S & L crisis there were few bank

failures in any given year to update the models' estimates. Therefore, the right panel of Figure 1 shows how well the prediction models' estimates from the S & L crisis years are able to predict failures between 2/1/2009 and 2/1/2010 based on year-end data from 2008, i.e., early into the subsequent crisis. We find that the accuracy of the models diminishes for predictions in the Great Financial Crisis relative to predictions during the S & L crisis, which is not surprising given the predictions are based on estimates from more than seventeen years in the past, rather than the previous year. However, the comparison we show in Table 3 of out-of-sample type –1 and type – 2 errors at the start of the Great Financial Crisis period reveals that the estimates of BMA models outperform stepwise regression in each year, other than 1985 for a given type – 2 error rate of 10%, and by a wide margin in some years, (7.8% in 1987). As we show in the right panel of Figure 1, the difference in the two models' predictive discrimination is also generalizable by the lower type – 1 errors over a wide range of type-2 errors for predictions in 2008, based on the 1987 BMA model's estimates.

These results suggest BMA estimates from an earlier crisis period outperform stepwise estimates when applied to predicting whether banks will fail in a subsequent crisis period. Policymakers could have used the estimates from the S & L crisis to create an early warning prediction during the most recent crisis. Using the estimates and a cutoff from the 1987 model, which provides the lowest type – 1 error rate for a type – 2 error rate of 10% during the S & L crisis, the model at year-end 2008 identifies 458 banks as likely to fail in the next year and correctly classifies 106 of the 129 (82%) banks that actually fail in 2009.[14] The limitation of the

[14] The corresponding type-2 error is equal to 4.8%. The error rates we report in Table 3 differ as they are based on outcomes and thus cutoffs that are not observed at year-end 2008, which are only known to policymakers at year-end 2009.

logit model is that it is not well suited for predicting failures through time given only a set of initial conditions. For this we turn to a survival model to account for time at risk and the slow resolution of failures during a crisis.

5.2. Cox model of bank failure

We use bank financial data from year-end 2008 and a Cox proportional hazards model to predict banks' time to failure during the period (2/1/2009-2/1/2015). We follow the banks through 2/1/2015 (after the crisis itself) to account for the often slow resolution of failures following financial deterioration. Similar to the estimation of our logit models, we use the failure experience observed during the S & L crisis to build an estimation model that will be applied out-of-sample at the start of the Great Financial Crisis. The Cox model is estimated using bank financial data that is available year-end 1984 to predict this set of banks' times to failure through the period 2/1/1985 – 2/1/1994. We limit the estimation of our model to the use of these "initial" conditions, which allows for our predictions to be based only on the information available to policymakers at the start of the crisis. Banks that do not fail during the period examined are said to be censored, as we do not observe their time to failure. Censoring may occur due to merger, a change in charter and hence reporting requirements, or voluntary closure. Banks are also said to be censored if they remain in the sample through the end of the period examined without failing. During the S & L crisis period (2/1/1985 – 2/1/1994) 92% of observations are censored, and in the out-of-sample period (2/1/2009-2/1/2015) 94% are censored. The high level of censoring observed in the data adds an extra layer of uncertainty to estimation of the Cox model, which is not present in the logit model.

Estimates of the Cox model using Bayesian model averaging and stepwise selection appear in Table 4. The BMA results indicate that 23 specifications were averaged over and that

the specification with the highest posterior model probability has a 30% likelihood of being the true model that generates the data. The model specification chosen by stepwise selection was not included in the set of models averaged over and includes a number of variables that did not receive support under BMA. The stepwise model includes cash (p-value 0.015), non-interest expense (p-value 0.004), dividends (p-value 0.077) , and federal funds sold (p-value 0.003) as a share of total assets; yet, each measure is less than 25% likely to have an effect under BMA based on their posterior effect probability. In the case of dividends, where the posterior effect probability reported is less than 5%, we would conclude there is evidence against an effect. Our estimates, consistent with Harvey's (2017) warning, again indicate that p-values can quite dramatically overstate the evidence of an effect relative to their Bayesian counterparts (posterior effect probabilities).

[Insert Table 4 about here]

The left panel of Figure 2 shows in-sample predictions of the Cox model are quite similar across the two variable selection approaches. The figure depicts the relation between type – 2 and type – 1 errors based on Heagerty and Zheng's (2005) definitions of incident sensitivity and dynamic specificity measured a year (365 days) at risk, i.e., 2/1/1986. For a type – 2 error of 10% (dynamic specificity is 90%), we find a corresponding type – 1 error rate of 47% (incident sensitivity is 53%) from our BMA and stepwise models. The right panel of figure 2 displays the area under the ROC curve over time to measure our models' accuracy. The AUC is equal to 0.82 for both BMA and stepwise models at the one-year mark, which indicates banks that fail at one year at risk are 82% more likely to have a risk score higher than their counterparts that survive. We find that the two models' risk scores which are based on their initial financial conditions also provide good long-term predictive power. This is evident as the AUC remains above 0.80

through time, which indicates banks' initial conditions are as well suited to discriminate between failures and non-failures early into a crisis as they are later on.

[Insert Figure 2 about here]

Our interest is in the ability of our models' estimates to predict the failure experience of banks during the Great Financial Crisis, i.e., out-of-sample. We apply our Cox model estimates based on the S & L crisis and data from year-end 2008 to predict failures in the period (2/1/2009-2/1/2015). We rely only on information, model estimates and data, available to policymakers at the start of the crisis. The predictive accuracy of the BMA model is similar out-of-sample to the in-sample predictions. For example, the AUC is greater than 0.78 during the entire time banks are at risk (Figure 2). The stepwise model performs substantially worse out-of-sample and in relation to the BMA model. A comparison of type – 1 and type – 2 errors at 365 days show for any type – 2 error that BMA has a lower type – 1 error. The AUC for the stepwise model is approximately equal to 0.66 throughout the time period examined. These results demonstrate that the initial conditions of banks at a start of a crisis and estimation models from a previous episode are useful for failure predictions during a crisis. We also find that Bayesian model averaging improves our out-of-sample predictions.

To test the sensitivity of our prediction model to the choice of initial conditions, we re-estimate the model with year-end 1985 financial data (banks become at risk of failing starting year-end 1985). The BMA estimates (we report these in an appendix) indicate 24 models were averaged over, and the best specification was 25% likely to be the true model. Predictive accuracy of both the stepwise and BMA estimates are similar in-sample – a dynamic false positive rate of 10% coincides to an incident sensitivity of 59% for each model. We find that BMA outperforms stepwise estimation out-of-sample, producing lower type – 1 errors (higher

sensitivity) for a given type – 2 error rate. The type – 1 error of our BMA estimates is 49% compared to 61% from stepwise selection. Our estimates further reveal the predictive accuracy is similar over time. The AUC ranges between 0.84 and 0.80 for our BMA estimates and between 0.76 and 0.74 for stepwise. It appears our Cox model's predictions from using BMA are not particularly sensitive to our choice of using estimates from the model estimated with 1984 data.

Policymakers can use the Cox model's estimates to draw inferences about the predicted survival experiences of banks during subsequent crises. We compare the experiences of a representative healthy, unhealthy, and critically ill bank, where healthy banks are defined as having a risk score equal to the average value of banks that did not fail during the S & L crisis, unhealthy banks have a score equal to the average of those who failed more than a year later, and critically ill are those who failed within the first year.[15] The risk score is equal to the linear portion of the proportional hazard and is useful (Hosmer, Lemeshow, and May, 2008) in comparing the survival experiences of different representative banks.[16] In-sample, during the S & L crisis, the BMA model's estimated risk scores classify 3.5% of the population of banks as critical, 8.1% as unhealthy, and 88.4% as healthy at year-end 1984. For this same period, we observe 0.8% of banks failed within a year, 7.1% banks would fail more than a year into the crisis, and 92.1% remained healthy throughout. The model provides an accurate overall assessment of risk to the banking system, based only on the information found in banks'

[15] Whalen (1991) uses a similar comparison of banks to describe the survivor functions in-sample. Lane (1984) also compares the survival experience of failed and non-failed banks.

[16] We provide a figure that plots the survival experience over time for each of the representative risk scores in the online appendix.

conditions at the start of the crisis. If one applies the same classification criterion from 1984 to banks' out-of-sample in 2008, we predict that 2.0% of banks are in critical condition, 7.0% are unhealthy, and 91% are healthy as of year-end 2008.[17] We compare this to the 1.7% of banks that failed within a year, the 3.9% of banks that would fail later during the crisis, and the 94.3% that remained healthy. The difference in the model's performance out-of-sample, with respect to the classification of unhealthy banks, i.e., prediction of more failures than occurred, is likely the result of the unprecedented interventions taken by the government to stabilize the banking system. Therefore, the model's predictions should be viewed as an "early warning" of what may occur prior to any intervention.

5.3. Predictions for the next crisis

Our results above suggest that when the next banking crisis strikes, policymakers should look to the failure experience of the most recent financial crisis to guide their predictions of bank failures. Table 5 reports the Bayesian model averaging estimates of the logit and Cox models applied to data from the Great Financial Crisis. The logit model is estimated using year-end data from 2008 to predict whether banks fail in 2009. For a type – 2 error rate of 10%, the corresponding type – 1 error rate of the logit model is equal to 4.6% and the area under the ROC (AUC) is equal to 0.98. Of the 25 variables considered, only 4 receive strong support (posterior effect probability >= 95%) for inclusion in the model specification, which include nonaccrual loans, equity, net income, and brokered deposits. However, several variables that received very strong support for inclusion in 1984 were not averaged over in 2008.These variables include

[17] The risk score is estimated relative to the typical bank for each period examined with the classification cutoffs based on the risk scores and failure experience observed during the S & L crisis.

loans past due 90 or more days, jumbo CDs, bank size, and the shares of C & I and agricultural loans. The earnings measure and brokered deposits indicator were added to the 2008 model's specification.

The Cox model is applied to the most recent crisis to estimate the time to bank failure between 2/1/2009 and 2/1/2015 when controlling for year-end 2008 bank data. We find there is a great deal of uncertainty in the true model's specification as 363 different models are averaged over by BMA and the best model specification is only 1.4%, likely to be the true model.[18] Each of the candidate control variables, other than federal funds purchased, is averaged over in the 2008 BMA model. Eleven of the variables receive strong support from the data - these variables include the four variables strongly supported in the 2008 logit model (nonaccrual loans, equity, net income, and brokered deposits) and loans past due, foreclosed real estate, securities, jumbo CDs, cash, demand deposits, and the share of consumer loans. Measures that received strong support for inclusion in the 2008 Cox model and not the 1984 model included equity, cash, deposits, brokered deposits, and the share of consumer loans. The shares of C& I, commercial real estate, and agriculture loans, along with age, size and share of insider loans received very strong support for inclusion in 1984 but had minimal posterior effect probabilities in 2008. We find for a dynamic type - 2 error rate of 10%, an incident sensitivity of 71% (type – 1 error of 29%) at 365 days at risk. The AUC(t) at 365 days is equal to 0.93 and remains above 0.87 throughout the 6 years the banks are at risk. The estimates can also be used to update the risk scores of banks that are healthy, unhealthy, and in critical condition – these risk scores, measured relative to the mean, are equal to -0.220, 3.083, and 4.957, respectively. When the next crisis

[18] Uncertainty is expected to be higher for the Cox model due to the challenge of predicting time to failure in the presence of heavy censoring (94%).

strikes, policymakers can use the Cox model's coefficients from Table 5, along with the most recent bank financial data to create predicted risk scores and apply the cutoffs above to determine the range of risks to bank failure.

6. Conclusions

Prior to the Great Financial Crisis, one might have viewed the S & L crisis of the 1980s as a "one-off", given the otherwise rarity of bank failures in the United States following the Great Depression of the 1930s. The bank failures that accompanied the crisis and great recession remind us that the banking system is always at risk. Therefore, policymakers need to be able to understand banks' exposure to failure early in a crisis period. Fundamentals of bank financial conditions, reflected in the CAMELS acronym, are good predictors of failures. Their relative importance though may change over time, which creates uncertainty as to one's choice of model specification. We use the failure experience of banks during the S & L crisis to build a prediction model applied to the recent financial crisis that accounted for specification uncertainty. When Bayesian model averaging is used, the accuracy of the logit model's estimates of bank failures during the S & L crisis improve the prediction of failures in the period 2/1/2009-2/1/2010. Further, we show that BMA also improves the out-of-sample predictive performance of survival times with the Cox model for the period 2/1/2009 – 2/1/2015.

Based on the experience during the S & L crisis, policymakers at year-end 2008 with the data available to them would have been able to use a logit model estimated from the S & L crisis to correctly identify 106 of the 129 (82%) banks that actually fail in 2009 with a type – 2 error rate of 3.8%. The Cox model's estimates allow policymakers to think more generally about failure patterns over different periods of time. Policymakers in 2008, using our estimates, would identify that 2.0% of banks are in critical condition, 7.0% are unhealthy, and 91% are healthy,

which is quite comparable, despite the massive intervention, to the 1.6% of banks that failed within a year, the 4.1% of banks that would fail later during the crisis, and the 92% that remained healthy. When the next crisis strikes, policymakers should look to recent events to base their inferences.

References

Arena, M., 2008. Bank failures and bank fundamentals: A comparative analysis of Latin America and East Asia during the nineties using bank-level data, *Journal of Banking and Finance* 32, 299-310.

Betz, F., S. Oprica, T. A. Peltonen, and P. Sarlin, 2014. Predicting distress in European banks, *Journal of Banking and Finance* 45, 225-241.

Brown, C.O., and I. S. Dinc, 2005. The politics of bank failures: evidence from emerging markets, *Quarterly Journal of Economics* 120, 1413-1444.

Chen, R., N. K. Chidambaran, M. B. Imerman, and B. J. Sopranzetti, 2014. Liquidity, leverage, and Lehman: A structural analysis of financial institutions in crisis, *Journal of Banking and Finance* 45, 117-139.

Cleary, S., and G. Hebb, 2016. An efficient and functional model for predicting bank distress: In and out of sample evidence, *Journal of Banking and Finance* 64, 101-111.

Cole, R. A., B. G. Cornyn, and J. W. Gunther, 1995. FIMS: A new monitoring system for banking organizations, *Federal Reserve Bulletin*, 81: 629-667.

Cole, R. A., and Gunther, J. W., 1998. Predicting bank failures: A comparison of on- and off-site monitoring systems, *Journal of Financial Services Research* 13, 103-117.

Cole, R. A., and L. J. White, 2012. Deja vu all over again: The causes of U.S. commercial bank failures this time around, *Journal of Financial Services Research* 42, 5-29.

Cole, R. A., and Q. Wu, 2010. Is hazard or probit more accurate in predicting financial distress? Evidence from U.S. bank failures, Unpublished working paper.

Collier, C., S. Forbush, D. A. Nuxoll, and J. O'Keefe, 2003. The SCOR system of off-site monitoring: Its objectives, functioning, and performance, *FDIC Banking Review* 15, 17-32.

DeYoung, R., and G. Torna, 2013. Nontraditional banking activities and bank failures during the financial crisis, *Journal of Financial Intermediation* 22, 397-421.

Edwards, W., H. Lindman, and L. J. Savage, 1963. Bayesian statistical inference for psychological research, *Psychological Review* 70, 193-242.

Fernandez, C., E. Ley, and M. F. J. Steel, 2001. Model uncertainty in cross-country growth regressions, *Journal of Applied Econometrics* 16, 563-576.

Freedman, D. A., 1983. A note on screening regression equations, *The American Statistician* 37, 152-155.

Goeman, J., 2010. L1 penalized estimation in the Cox proportional hazards model, *Biometrical Journal* 52, 70-84.

Gogas, P., T. Papadimitriou, and A. Agrapetidou, 2018. Forecasting bank failures and stress testing: A machine learning approach, *International Journal of Forecasting* 34(3), 440-455.

Greene, W. H., 1997. Econometric Analysis, 3rd edition. Prentice Hall, Upper Saddle River, New Jersey. .

Harvey, C. R., 2017. Presidential address: The scientific outlook in financial economics, *Journal of Finance* 72(4), 1399-1440.

Heagerty, P. J., and P. Saha-Chaudhuri, 2015. Package risksetROC. Retrieved from https://cran.r-project.org/web/packages/risksetROC/risksetROC.pdf

Heagerty, P.J., and Y. Zheng, 2005. Survival model predictive accuracy and ROC curves, *Biometrics* 61, 92-105.

Hosmer, D. W., S. Lemeshow, and S. May, 2008. Applied Survival Analysis: Regression Modeling of Time to Event Data. 2nd edition. Wiley, New York, New York.

Jin, J. Y., K. Kanagaretnam, and G. J. Lobo, 2011. Ability of accounting and audit quality variables to predict bank failure during the financial crisis. *Journal of Banking and Finance* 35, 2811-2819.

Lane, W. R., S. W. Looney, and J. W. Wansley, 1986. An application of the Cox proportional hazards model to bank failure, *Journal of Banking and Finance* 10, 511-531.

Le, H. H., and J. Viviani, 2018. Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios, *Research in International Business and Finance* 44, 16-25.

Leamer, E. E., 1978. Specification Searches: Ad Hoc Inference with Non-experimental Data. Wiley, New York, New York.

Mayes, D. G., and H. Stremmel, 2014. The effectiveness of capital adequacy measures in predicting bank distress, SUERF Study 2014/1. Brussels: Larcier.

Nasiripour, S., 2016. Sheila Bair called the financial crisis. Here's her new nightmare, Bloomberg.com https://www.bloomberg.com/features/2016-sheila-bair-student-debt/ Accessed 12 December 2016

Ng, J., and S. Roychowdhury, 2014. Do loan loss reserves behave like capital? Evidence from recent bank failures, *Review of Accounting Studies* 19, 1234-1279.

Raftery, A. E., 1995. Bayesian Model Selection in Social Research, in *Sociological Methodology 1995*. (editor P. V. Marsden) , Cambridge MA.: Blackwells Publishers (111-195).

Raftery, A. E., J. Hoeting, C. Volinsky, I. Painter, and K. Y. Yeung, 2018. Package BMA. Retrieved from https://cran.r-project.org/web/packages/BMA/BMA.pdf.

Raftery, A. E., D. Madigan, and J. A. Hoeting, 1997. Bayesian model averaging for linear regression models, *Journal of the American Statistical Association* 92, 179-191.

Raftery, A. E., D. Madigan, and C. T. Volinsky, 1995. Accounting for model uncertainty in survival analysis improves predictive performance (with discussion), in Bayesian Statistics 5 (editors J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), Oxford: Oxford University Press (323-349).

Schwarz, G., 1978. Estimating the dimension of a model, *The Annals of Statistics* 6, 461-464.

Shumway, T., 2001. Forecasting bankruptcy more accurately: A simple hazard model, *Journal of Business* 74, 101-124.

Thomson, J. B., 1992. Modeling the bank regulator's closure option: A two-step logit regression approach, *Journal of Financial Services Research* 6, 5-23.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society Series B* 58, 267–288.

Volinsky, C., T., D. Madigan, A. E. Raftery, and R. A. Kronmal, 1997. Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke, *Applied Statistics* 46, 433-448.

Whalen, G., 1991. A proportional hazards model of bank failure: An examination of its usefulness as an early warning tool, *Economic Review* 27, 21-31.

Wheelock, D. C., 1992. Deposit insurance and bank failures: New evidence from the 1920s, *Economic Inquiry* 30, 530-543.

Wheelock, D. C., and P. W. Wilson, 1995. Explaining bank failures: deposit insurance, regulation, and efficiency, *Review of Economics and Statistics* 77, 689-700.

Wheelock, D. C., and P. W. Wilson, 2000. Why do banks disappear? The determinants of U.S. bank failures and acquisitions, *Review of Economics and Statistics* 82, 127-138.

Table 1

Difference in mean characteristics of failing and non-failing banks

The sample of banks for each crisis period are divided between banks that either fail or do not fail during the period - S & L crisis (2/1/1985 – 2/1/1994) and the Great Financial Crisis (2/1/2009 -2/1/2015). The means of the controls reported are measured year-end 1984 and 2008.

	S & L Crisis		Great Financial Crisis	
	Failed	Non-failed	Failed	Non-failed
Loans past due 90+ days	0.0108***	0.0064	0.004***	0.0019
Nonaccrual loans	0.0143***	0.0068	0.0542***	0.0103
Foreclosed real estate	0.0073***	0.0036	0.018***	0.0036
Equity	0.0887	0.0892	0.0798***	0.1154
Net income	-0.0059***	0.0090	-0.0261***	0.0048
Securities	0.1578***	0.2842	0.1113***	0.2029
Loan loss reserves	0.0098***	0.0062	0.0194***	0.0095
Jumbo CDs	0.2183***	0.1035	0.2142***	0.1585
Cash	0.1081***	0.0923	0.0408***	0.0580
Demand deposits	0.1681***	0.1578	0.0629***	0.1081
Federal funds purchased	0.012	0.0136	0.0134	0.0167
Volatile liability expense	0.0915	0.0956	0.0426***	0.0363
Charge-offs	0.0123***	0.0055	0.0144***	0.0038
Brokered deposits	0.1227***	0.0248	0.8152***	0.3753
Non-interest expense	0.0396***	0.0323	0.0355	0.0329
Insider loans	0.0129***	0.0054	0.0174**	0.0151
Dividends	0.0027***	0.0037	0.0018***	0.0055
Age	37.5064***	55.8927	35.872***	68.2487
Size	10.4061***	10.6093	12.4985***	11.9514
Provisions for loan losses	0.0144***	0.0057	0.022***	0.0051
C & I loans	0.2109***	0.1278	0.0999	0.0976
Consumer loans	0.1305	0.1264	0.0192***	0.0455
Commercial real estate	0.0533***	0.0440	0.235***	0.1617
Agriculture loans	0.0708	0.0681	0.0106***	0.0446
Federal funds sold	0.0597	0.0598	0.0301	0.0278
Number of observations	1100	12924	422	7019

*, **, ***, indicate significant differences in means between failed and non-failed banks at the 0.10, 0.05, and 0.01 level, respectively.

Table 2

Logit model estimates of bank failure (1984)

The logit model uses a cross-section of year-end bank data from 1984 to predict failures in the period 2/1/1985 - 2/1/1986. BMA estimates reported include the posterior mean (Coef), standard deviation (SE), and effect probabilities (PEP) of the variables averaged over. The posterior model probability (PMP) of the best model averaged over is also reported. The stepwise model is selected based on Akaike's information criterion (AIC).

	Bayesian model averaging			Stepwise AIC		
	Coef	SE	PEP	Coef	SE	P-value
Constant	4.043	1.495	100.0	5.842	1.685	0.001
Loans past due 90+ days	19.454	6.364	95.6	16.608	5.000	0.001
Nonaccrual loans	17.463	3.832	100.0	14.342	4.084	< .001
Foreclosed real estate	21.660	4.094	100.0	17.731	4.234	< .001
Equity	-47.505	4.957	100.0	-42.792	5.471	< .001
Net income	-	-	-	-	-	-
Securities	-	-	-	-4.428	1.436	0.002
Loan loss reserves	-	-	-	-	-	-
Jumbo CDs	5.571	1.089	100.0	5.380	1.028	< .001
Cash	-0.144	0.928	3.1	-6.348	2.846	0.026
Demand deposits	-	-	-	-	-	-
Federal funds purchased	-	-	-	-	-	-
Volatile liability expense	-	-	-	-	-	-
Charge-offs	-	-	-	-	-	-
Brokered deposits	-	-	-	-	-	-
Non-interest expense	-	-	-	-	-	-
Insider loans	-	-	-	-	-	-
Dividends	-	-	-	-	-	-
Age	0.004	0.006	30.3	0.012	0.004	0.003
Size	-0.859	0.139	100.0	-0.891	0.141	< .001
Provisions for loan losses	-	-	-	7.887	4.577	0.085
C & I loans	5.075	1.162	100.0	3.716	1.104	0.001
Consumer loans	-	-	-	-	-	-
Commercial real estate	-	-	-	-	-	-
Agriculture loans	6.389	0.959	100.0	4.298	0.961	< .001
Federal funds sold	-	-	-	-6.463	2.782	0.020
Observations	14024			14024		
Models averaged over	4					
PMP of the best model	.62					

Table 3

Out-of-sample type-1 error rate for year ahead failures

The out-of-sample type - 1 error rate (1 - sensitivity) associated with a type - 2 error rate (1 - specificity) of 10%. The logistic specifications are estimated using year-end data from the indicated estimates year, which are then applied to year end data in the prediction year to form out-of-sample predictions. Predictions in 1985 (2008) refer to out-of-sample predictions of failures in the period 2/1/1986 - 2/1/1987 (2/1/2008 – 2/1/2009) based on financial conditions at year-end 1985 (2008) and estimates from the model estimated using year-end data from 1984.

Estimates Year	Prediction Year	BMA	Stepwise	Prediction Year	BMA	Stepwise
1984	1985	7.6%	6.9%	2008	33.3%	39.5%
1985	1986	7.7%	7.2%	2008	24.0%	22.5%
1986	1987	23.6%	22.0%	2008	39.5%	44.2%
1987	1988	1.5%	3.1%	2008	8.5%	16.3%
1988	1989	4.8%	4.8%	2008	13.2%	14.7%
1989	1990	1.7%	3.3%	2008	11.6%	15.5%
1990	1991	13.9%	14.8%	2008	16.3%	17.8%
1991	1992	2.4%	4.9%	2008	12.4%	14.0%

Table 4

Cox proportional hazards model estimates of bank failure - S & L Crisis

Cox model estimates of the time to bank failure in days between 2/1/1985 - 2/1/1994. The control variables are measured using year-end bank data from 1984. BMA estimates reported include the posterior mean (Coef), standard deviation (SE), and effect probabilities (PEP) of the variables averaged over. The posterior model probability (PMP) of the best model averaged over is also reported. The stepwise model is selected based on Akaike's information criterion (AIC).

	Bayesian model averaging			Stepwise AIC		
	Coef	SE	PEP	Coef	SE	P-value
Loans past due 90+ days	9.886	2.072	100.0	9.099	2.065	< .001
Nonaccrual loans	7.231	1.863	100.0	6.797	1.861	< .001
Foreclosed real estate	10.331	1.680	100.0	9.762	1.712	< .001
Equity	-1.411	1.111	68.8	-1.335	0.700	0.056
Net income	-7.043	3.384	84.8	-10.910	1.321	< .001
Securities	-4.158	0.370	100.0	-4.697	0.373	< .001
Loan loss reserves	0.461	1.992	6.4	-	-	-
Jumbo CDs	5.309	0.259	100.0	5.396	0.259	< .001
Cash	-0.097	0.360	8.4	-1.150	0.472	0.015
Demand deposits	0.346	0.662	25.8	1.609	0.512	0.002
Federal funds purchased	0.017	0.177	1.6	-	-	-
Volatile liability expense	-	-	-	-	-	-
Charge-offs	-	-	-	-	-	-
Brokered deposits	-	-	-	-	-	-
Non-interest expense	-0.886	2.190	17.2	-6.304	2.185	0.004
Insider loans	4.387	0.795	100.0	4.492	0.834	< .001
Dividends	0.454	2.519	4.1	11.150	6.304	0.077
Age	0.005	0.001	100.0	0.005	0.001	< .001
Size	-0.172	0.034	100.0	-0.201	0.034	< .001
Provisions for loan losses	1.944	4.178	19.7	-	-	-
C & I loans	2.793	0.325	100.0	2.367	0.308	< .001
Consumer loans	-	-	-	-	-	-
Commercial real estate	2.864	0.623	100.0	2.246	0.614	< .001
Agriculture loans	2.636	0.356	100.0	2.254	0.348	< .001
Federal funds sold	-0.269	0.582	21.5	-1.490	0.497	0.003
Observations	14024			14024		
Models averaged over	23					
PMP of the best model	.30					

Table 5

Bayesian model averaging estimates of bank failures during the Great Financial Crisis

The logit model estimates whether banks fail in 2/1/2009 - 2/1/2010, and the Cox model estimates time to bank failure in days between 2/1/2009 - 2/1/2015. The control variables for both models are measured using year-end bank data from 2008. BMA estimates reported include the posterior mean (Coef), standard deviation (SE), and effect probabilities (PEP) of the variables averaged over. The posterior model probability (PMP) of the best model averaged over is also reported.

	Logit Model			Cox Model		
	Coef	SE	PEP	Coef	SE	PEP
Constant	0.018	0.757	100.0	-	-	-
Loans past due 90+ days	-	-	-	21.967	5.216	100.0
Nonaccrual loans	22.251	3.795	100.0	11.315	1.197	100.0
Foreclosed real estate	10.640	9.409	62.3	12.706	2.189	100.0
Equity	-69.461	6.325	100.0	-34.146	2.468	100.0
Net income	-27.315	7.654	100.0	-11.518	3.902	100.0
Securities	-0.474	1.266	14.5	-2.969	0.686	100.0
Loan loss reserves	-	-	-	5.191	6.696	43.2
Jumbo CDs	-	-	-	1.968	0.528	99.8
Cash	-	-	-	-3.777	1.645	93.5
Demand deposits	-	-	-	-5.136	1.274	100.0
Federal funds purchased	-	-	-	-	-	-
Volatile liability expense	-	-	-	2.776	2.503	60.6
Charge-offs	-9.081	12.635	39.0	-6.703	8.085	54.1
Brokered deposits	1.064	0.398	94.9	0.871	0.139	100.0
Non-interest expense	-2.816	7.019	15.4	-0.990	3.142	10.9
Insider loans	-	-	-	2.849	3.639	43.7
Dividends	-	-	-	-45.398	22.326	90.3
Age	-	-	-	-0.001	0.002	36.0
Size	-	-	-	0.000	0.001	0.1
Provisions for loan losses	-	-	-	2.558	6.303	17.9
C & I loans	-	-	-	-0.150	0.499	10.6
Consumer loans	-2.260	5.232	18.8	-7.937	2.358	100.0
Commercial real estate	-	-	-	0.581	0.689	47.8
Agriculture loans	-	-	-	-0.113	0.595	4.9
Federal funds sold	-	-	-	1.866	1.503	67.6
Observations	7441			7441		
Models averaged over	18			363		
PMP of the best model	.28			.014		

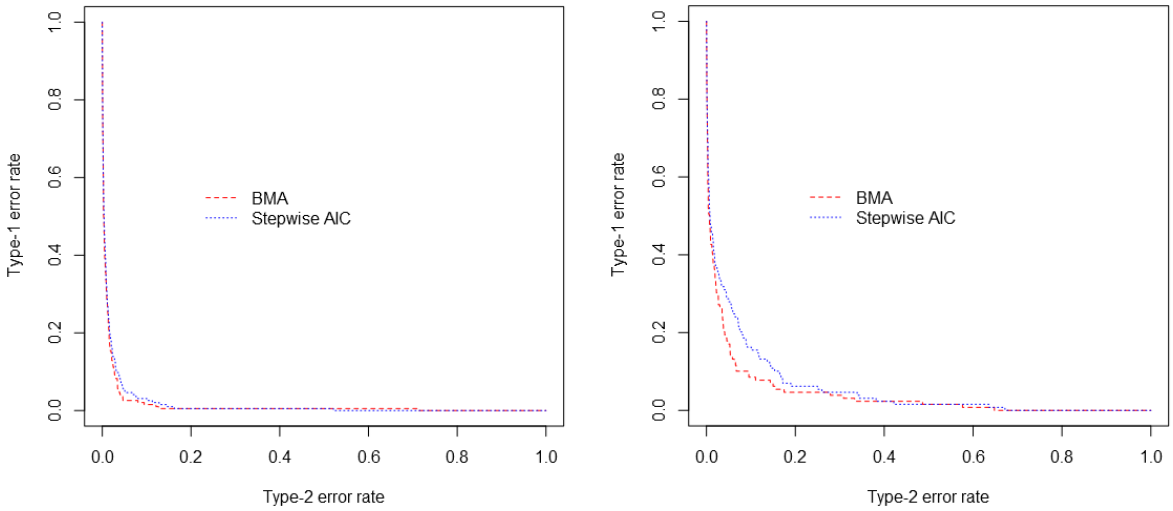

Figure 1

The out-of-sample relationship between type – 2 and type – 1 prediction errors from a logit model

Prediction errors from identifying bank failures in the year ahead using Bayesian model averaging and stepwise selection. The prediction model in the left panel uses year-end financial data from 1987 to estimate the model and uses these estimates along with year-end data from 1988 to predict bank failures in the S & L crisis period 2/1/1989 – 2/1990. The right panel also uses year-end financial data from 1987 to estimate the logit model and uses these estimates along with year-end data from 2008 to predict failures in the Great Financial Crisis period 2/1/2009– 2/1/2010.

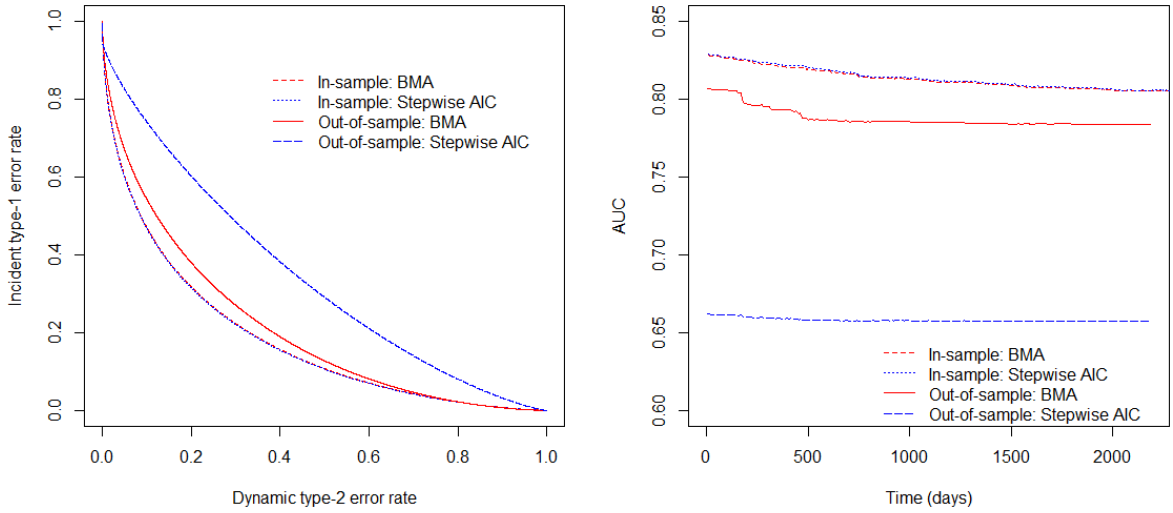

Figure 2

Prediction errors from a Cox model of the time to bank failure

The left panel plots the relation between dynamic type – 2 and incident type – 1 errors at 365 days using Bayesian model averaging and stepwise selection. In-sample accuracy is measured based on the estimates of the Cox model using year-end data from 1984 to predict time to failure during the S & L crisis period. Out-of-sample accuracy is measured using the earlier period's estimates, combined with year-end data from 2008, to determine out-of-sample risk scores during the Great Financial Crisis. The right panel plots the area under the ROC curve (AUC) at various times and both in and out-of-sample for the Cox model estimated using Bayesian model averaging and stepwise selection.