RICE UNIVERSITY

Essays on Commercial Banking: Survival, Performance, and Heterogeneous Technologies

by

Pavlos Almanidis

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Approved, Thesis Committee:

Robin C. Sickles, Chair
Reginald Henry Hargrove Chair of
Economics
Rice University

Mahmoud A. El-Gamal
Chair and Professor of Economics
Rice University

James P. Weston
Associate Professor of Finance
Jones Graduate School of Management
Rice University

Houston, Texas

April, 2011

ABSTRACT

Essays on Commercial Banking: Survival, Performance, and Heterogeneous Technologies

Pavlos Almanidis

In the first chapter, we focus on explaining the U.S. commercial banking failures during the recent financial crisis. We employ the semi-parametric mixture hazard model (MHM) with both continuous and discrete time specifications to first, distinguish between troubled and healthy banks and second, to estimate the probability and the timing of their failure. We combine the MHM with the stochastic frontier model (SFM) to explore the role of managerial inefficiency on a bank's longer term viability. We find that the discrete-time MHM which takes the managerial inefficiencies into account fits well and dominates other competing specifications by accurately predicting the timing of failures both in and out of the sample.

The second chapter explores a new class of flexible cross-sectional parametric SFMs that impose an unobservable bound on the inefficiency term. We consider

doubly truncated normal, truncated half-normal, and truncated exponential distributions to model the inefficiencies. We extend the models to the panel data setting and specify a time-varying inefficiency bound. We apply these models to analyze the performance of the U.S. commercial banking industry during 1984-2009.

In the third chapter, we address the issue of the "wrong" skewness of the least squares residuals that often arises in applied studies using the traditional SFM. Findings of "wrong" skewness imply that the SFM is misspecified and all firms are fully efficient. Based on doubly truncated normal distribution that displays both positive and negative skewness, we prove that "wrong" skewness does not necessarily imply that the SFM model is misspecified.

The fourth chapter investigates the existence of heterogeneous technologies in the U.S. commercial banking industry through the threshold effects estimation techniques, modified to allow for time-varying effects. We employ the total assets as a threshold variable and determine seven distinct technology-groups.

In the fifth chapter, we describe the commercial banking data that are extracted from the quarterly Consolidated Reports of Condition and Income (Call Reports). We detail the construction of the key variables used in this thesis, which mainly contain output quantities, input quantities and prices, bank-specific structural and geographical characteristics, as well as a number of measures of risk.

Acknowledgements

As you set out for Ithaca[1] *pray that the journey is long, full of adventure, full of knowledge. -Constantine P. Cavafy, 1911-*

First, and most importantly I would like to thank my advisor, Dr. Robin Sickles, for setting me on the path to the marvelous journey of Ithaca. His guidance, encouragement, and unique personality has made the journey the most enjoyable one. He endowed me with the knowledge and the necessary tools that undoubtedly will help me to understand and navigate challenges in my own life and the lives of others for many years to come. In short, I could not have imagined having a better advisor and mentor for my Ph.D. study and hopefully one day I would become as good an advisor to my students as Dr. Sickles has been to me.

Besides my advisor, I would like to express my deepest gratitude to Dr. Mahmoud El-Gamal for he has always been there to listen and provide me with his insightful comments and suggestions. I would also like to thank Dr. James Weston for his tough and thoughtful questions while serving as a dissertation committee member.

[1]Ithaca is the home island of Odysseus located in the Ionian Sea, in Greece. According to myth, Odysseus spent adventurous 10 years while returning home from the Troy war.

I am very grateful to all my professors from economics and statistics department for their genuine care and enthusiasm in transferring their knowledge that further evoked my sense of curiosity and a thirst for knowledge for the economic and statistic science. I would also like to thank my professors from Economic Analysis and Policy division, Aristotle University of Thessaloniki for their encouragment and inspiration. I owe much to Dr. Nikolaos Varsakelis for his advise and continuous encouragment to follow my dream of doctoral studies. I undoubtedly have been very fortunate to meet and learn a lot from these professors during my scholastic life.

I am indebted to Robert Adams at the Board of Governors of the Federal Reserve System. His very valuable advise and support have helped me to deeply understand the U.S. commercial banking industry and better fix and focus my research ideas.

To my dearest friends in Greece, who always stood by my side and truly cared about my success. I am really glad that I have you! Also, I thank all my friends and colleagues at Rice University who in one way or another provided their support and helped me to stay sane through all these years. In particular, I would like to thank Levent Kutlu for his great friendship, true interest and valuable input in some of my research and non-research related problems. I would also like to thank David Splinter and Jiaqi Hao for being a great company during my job interviews. I would like to express my sincere thanks to Martha Alexander, Enrique Patino,

and Joyce Thormodsgaard for proofreading the first and the fourth chapters and of course for being true friends and great neighbors.

I am heartily thankful to Altha Rodgers for her help in a number of ways and always with a wonderful smile on her face. I truly acknowledge her continuos support and prompt responsiveness to all of my requests during all these years.

I gratefully acknowledge the generous financial support from Social Sciences Research Institute at Rice University that funded parts of the research discussed in this thesis. Without this, I would not be able to afford the powerful computer equipment that was essential to handle large data sets and estimate complicated models. Dr. Larry Scott Baggett helped to write the source code in SAS that considerably facilitated the data extraction and filtration. I really appreciate his help.

I wish to express my love and gratitude to my family; for their endless support through the duration of my studies, understanding, and most importantly believing in me, which made me work harder toward achieving my goals.

Finally, I would like to thank my wife, Ioulia. Only she knows that I have been through and how many times I had to face cyclops and angry Posidon to arrive at Ithaca. Her encouragement, patience, and help over all these years cannot be described with human words. The full thesis is dedicated to her!

Contents

ABSTRACT ii

Acknowledgements iv

List of Tables x

List of Figures xiv

Chapter 1. Banking Crises, Early Warning Models, and Efficiency 1

1.1. Introduction 1

1.2. The Mixture Hazard Model 15

1.3. Stochastic Frontier Model combined with Mixture Hazard Model 30

1.4. Empirical Model and Data 39

1.5. Results and Predictive Accuracy 48

1.6. Conclusions 55

1.7. Appendix A: Derivation of the likelihood function 57

1.8. Appendix B: Tables and Figures 59

Chapter 2. Bounded Stochastic Frontiers with an Application to the US Banking Industry: 1984-2009[2] 70

2.1. Introduction 70

2.2. The Model 75

2.3. The Skewness Issue 80

2.4. Estimation 85

2.5. Panel Data 96

2.6. Simulations 97

2.7. Efficiency Analysis of Banking Industry 100

2.8. Conclusions 108

2.9. Appendix: First-order derivatives of the log-likelihood function 110

Chapter 3. Skewness Issue in Stochastic Frontier Models:[3] Fact or Fiction? 121

3.1. Introduction 121

3.2. Skewness issue in Stochastic Frontier Analysis 124

3.3. Skewness statistic under the bounded inefficiencies 133

3.4. Further Discussion 138

3.5. Conclusions 141

Chapter 4. Accounting for Heterogeneous Technologies in the Banking

[2]This is a version of my work with professors Junhui Qian (Shanghai Jiao Tong University) and Robin Sickles.

[3]This is a version of my work with professor Robin Sickles.

Industry: A time-varying Stochastic Frontier Model with Threshold Effects 142

4.1. Introduction 142

4.2. Heterogeneity in Stochastic Frontier Models 147

4.3. The Threshold Effects Stochastic Frontier Model 155

4.4. Empirical Model and Data 160

4.5. Empirical Results 165

4.6. Conclusions 169

Chapter 5. Commercial Banking Data 178

5.1. Balance Sheet Data 180

5.2. Income Statement Data 186

References 199

List of Tables

1.1	Per State distribution of failed banks	4
1.2	CAMELS proxy Financial Ratios	59
1.3	Structural, Geographical, and State-Specific Macroeconomic variables	60
1.4	Descriptive Statistics for CAMELS proxy financial ratios for the fourth quarter of 2007 and 2009	61
1.5	Descriptive Statistics for variables that enter the cost function for the fourth quarter of 2007 and 2009	62
1.6	Estimates from the Continuous-Time Semiparametric Proportional Mixture Hazard Model (Model I) and Discrete-Time Mixture Hazard Model (Model II)	63
1.7	Estimates from the Stochastic Frontier Continuous-Time Semiparametric Proportional Mixture Hazard Model (Model III) and Stochastic Frontier Discrete-Time Mixture Hazard Model (Model IV)	64
1.8	Cost efficiencies results for the sample of Nonfailed Banks	65

1.9	Cost efficiencies results for the sample of Failed Banks	65
1.10	In-sample classification error decomposition	65
1.11	Out-of-sample classification error decomposition	66
2.1	Key Results	79
2.2	Proportion of Positive Skewness for Simulated Residuals in the Doubly Truncated Normal Model.	83
2.3	Central Moments of ε	86
2.4	Comparisons of Various Estimators. Estimates and standard errors (in parentheses) for each model parameters from competing models (FIX, CSSW, BC, BIE)	107
2.5	Spearman Rank Correlations of Efficiencies	108
2.6	Descpriptive statistics for bank-specific variables	112
2.7	Monte Carlo results for Truncated Half Normal model. The number of repetitions $M = 1000$. Sample size $N = 200$	115
2.8	Monte Carlo results for Truncated Half Normal model. The number of repetitions $M = 1000$. Sample size $N = 1000$	116
2.9	Monte Carlo results for Doubly Truncated Normal model. The number of repetitions $M = 1000$. Sample size $N = 200$	117
2.10	Monte Carlo results for Doubly Truncated Normal model. The number of repetitions $M = 1000$. Sample size $N = 1000$	118

2.11	Monte Carlo results for Truncated Exponential model. The number of repetitions $M = 1000$. Sample size $N = 200$	119
2.12	Monte Carlo results for Truncated Exponential model. The number of repetitions $M = 1000$. Sample size $N = 1000$	120
3.1	Monte Carlo results for for Half-Normal model. The number of repetitions $M = 1000$.	140
4.1	Estimation Results: Threshold Values, Cost Efficiency, Returns to Scale, Technical Change, Return on Assets, Return on Equity, Profit Margin, and Asset Utilization.	166
4.2	Summary statistics for selected periods	171
4.3	Technology-Group Estimation Results	172
5.1	Balance Sheet Data	188
5.2	Average Input Prices	189
5.3	1984Q1 Balance Sheet-Asset Side data (in millions of U.S. dollars)	190
5.4	1984Q1 Balance Sheet-Liability and Equity Capital Side data (in millions of U.S. dollars)	191
5.5	1993Q1 Balance Sheet-Asset Side data (in millions of U.S. dollars)	192

5.6	1993Q1 Balance Sheet-Liability and Equity Capital Side data (in millions of U.S. dollars)	193
5.7	2000Q1 Balance Sheet-Asset Side data (in millions of U.S. dollars)	194
5.8	2000Q1 Balance Sheet-Liability and Equity Capital Side data (in millions of U.S. dollars)	195
5.9	2010Q2 Balance Sheet-Asset Side data (in millions of U.S. dollars)	196
5.10	2010Q2 Balance Sheet-Liability and Equity Capital Side data (in millions of U.S. dollars)	197
5.11	Income Statement data (in millions of U.S. dollars)	198

List of Figures

1.1	The number of failed and troubled banks for 2007.Q3-2010.Q2 period	3
1.2	Per State Distribution of failed banks	5
1.3	Illustration of the bank financial health condition during the period of financial turmoil	17
1.4	Financial ratios over the 2007.Q4-2010.Q2 period. Solid line is for non-failed banks and dashed line is for failed banks.	66
1.5	Distribution of estimated cost efficiencies obtained from Models III, IV and the Random Effects Model.	67
1.6	Estimated Average Returns to Scale and Technological Change from Models III and IV	67
1.7	SPMHM - Survival profile of the average failed bank (2008-2009)	68
1.8	DTMHM - Survival pofile of the average failed bank (2008-2009)	68
1.9	Model III - Survival profile of the most and the least efficient bank	69

1.10	Model IV - Survival profile of the most and the least efficient bank	69
2.1	Quantile-Quantile plot	112
2.2	Averaged Efficiencies from each estimator	113
2.3	Estimated Inefficiency Bound	114
4.1	Distribution of log of Total Assets	175
4.2	Distribution of banks in groups over time	175
4.3	Averaged cost efficiencies for seven groups and the pooled sample	176
4.4	Average returns to scale measure for seven groups and the pooled sample	176
4.5	Average technical change for seven groups and the pooled sample	177
5.1	Number of banks in asset size percentiles	180

CHAPTER 1

Banking Crises, Early Warning Models, and Efficiency

1.1. Introduction

Financial crisis that started in summer of 2007 has led the U.S. and international economies to an unprecedented meltdown, creating political instability and uncertainty worldwide. According to many, it also could be characterized as the worst economic crisis since the Great Depression of the 1930's. It appears that the crisis originally started in the secondary market for residential mortgages after the dramatic increase in delinquencies and default rates on subprime residential-mortgage-backed securities (RMBS) as a result of the collapse of housing bubble during the second half of the year 2006. It very quickly spread to the banking industry as many banks, in particular large banks, were highly involved in this market until recently, fact that caused a substantial portion of these financial institutions experience widespread distress that led to closures, mergers, takeovers, or injection of heavy doses of government funds. More than 290 banks and thrifts failed or more correctly were forced into closure by regulatory agencies in three years from late 2007 to the middle of October of 2010. At the same time, the

number of troubled or problem banks on the watch list of the Federal Deposit Insurance Corporation (FDIC) has dramatically increased. The sharp increase in the number of failed and troubled banks since 2007 is illustrated in figure 1.1. Among the states that experienced the most failures are California, Georgia, Florida, and Illinois, accounting for more than half of all banking failures. Table 1.1 displays the number of failures per state within these three years, while figure 1.2 provides a map of banking failures per state for the referenced period.

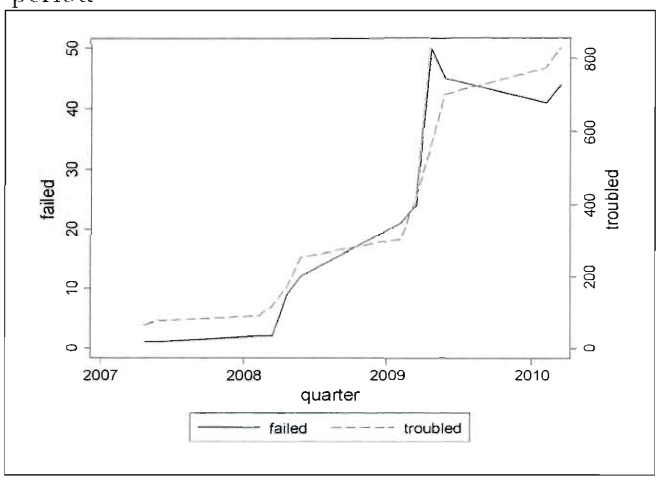
The distinguishing characteristic of the current banking failures from those of the earlier crises of the 1980's and 1990's is that the failures were not limited to the small financial institutions. The rapid credit expansion, as a result of over-optimism about the economic growth, and the bad quality loans and investments made in good times have mainly taken their toll on large multi-billion dollar financial institutions. Approximately one out of five failed banks had asset sizes of over $1 billion. As recently as 2008, 36% of failed banks were large banks, among them the largest bank failure in the history of U.S., that of Washington Mutual with $307 billion in assets.[1] That same year saw Lehman Brothers file for Chapter 11 bankruptcy protection and IndyMac bank, with $32 billion in assets, was taken over by the FDIC.[2] These large financial institution failures created large uncertainties about the exposure of other financial institutions (healthy and troubled) to

[1]Continental Illinois Bank and Trust Company of Chicago that failed in 1984 had one-seventh of Washington Mutual's assets.

[2]Chapter 11 permits reorganization under the bankruptcy laws of the United States. A financial institution filling for Chapter 11 bankruptcy protection usually proposes a plan of reorganization to keep its business alive and pay its creditors over time.

additional risks, reduced the availability of credit from investors to banks, drained the capital and money markets of confidence and liquidity, triggered the failure of smaller community banks, and raised the fears of severe instability in the financial system and the global economy.[3] A greater attention was paid to the larger institutions at danger, commonly described as too-big-to-fail, which received financial and other assistance from regulatory authorities as they thought their failure could impose a greater systemic risk that could substantially damage the economy and lead to conditions similar to, or possibly exceeding, those of the Great Depression.

Figure 1.1. The number of failed and troubled banks for 2007.Q3-2010.Q2 period



As the number of failures is still rising, one may reasonably ask when failures will begin to fall. On one hand, pessimistic scenarios predict the number of failures

[3]Community banks are banks with assets sizes of $1 billion or less. They operation is oftentimes limited to the rural communities and small cities. They usually engage in traditional banking activities and provide more personal-based services.

Table 1.1. Per State distribution of failed banks

State	number of failed banks	State	number of failed banks
Alabama	4	North Carolina	2
Arkansas	1	Nebraska	2
Arizona	7	New Jersey	3
California	32	New Mexico	2
Colorado	3	Nevada	10
Florida	41	New York	4
Georgia	44	Ohio	6
Iowa	1	Oklahoma	2
Idaho	1	Oregon	6
Illinois	37	Pennsylvania	1
Indiana	1	Puerto Rico	3
Kansas	5	South Carolina	4
Kentucky	1	South Dakota	1
Louisiana	1	Texas	8
Massachusetts	1	Utah	5
Maryland	5	Virginia	2
Michigan	9	Washington	13
Minnesota	14	Wisconsin	2
Missouri	9	West Virginia	1
Mississippi	1	Wyoming	1

in 2010 to reach 200, the most since the end of the savings and loan (S&L) crisis of early 1990's. They also predict that this rate of failure will continue at the same pace in subsequent two years.[4] On the other hand, more optimistic scenarios expect the 2010 to be the peak year of banking distress as the worst time has

[4]This prediction is due to Gerard Cassidy and his colleagues at RBC Capital Markets, who were among the first analysts that predicted the rising number of current banking failures very early. Gerard Cassidy is the developer of the Texas ratio, a tool which is able to determine insolvent banks.

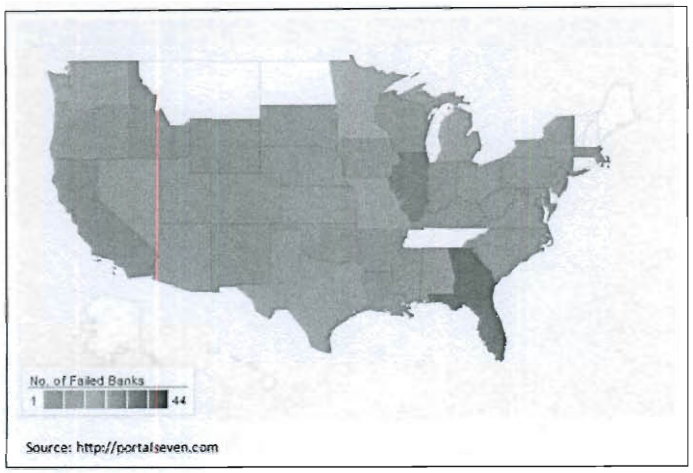


Figure 1.2. Per State Distribution of failed banks

passed and that the economy will continue to recover, with the housing market displaying apparent signs of stabilization.

Regulatory authorities have always considered banking failures as a major public policy concern due to their special role in the economic network and in implementation of an effective monetary policy. In addition, failures of certain banks could possibly lead to contagion or domino effects and thus negatively affect the safety and soundness of the banking industry and of the entire economy. Head-on confronting of crises usually involves taking a series of extraordinary costly actions and sacrificing valuable economic resources. There are typically two approaches to calculate the costs of a banking crisis: the narrow fiscal or quasi-fiscal costs, which involve large government guarantees and central bank bailouts, and the system-wide economic costs, which include output loss, increases in unemployment rate,

missed business opportunities, and etc. It is too early to give approximate figures for the fiscal or economic costs as the consequences of the banking crisis are still unfolding. Nevertheless, we can make some tentative observations regarding the costs of banking failures to the FDIC's Deposit Insurance Fund (DIF) and the banking industry output and employment.[5] With regard to the first, it is estimated that the 140 failures of 2009, with combined assets of about $170 billion, cost the DIF $36.4 billion, while the cost of failures through mid-October 2010 involving banks with combined assets of more than $85 billion, is expected to exceed $20 billion. Notably, only the failures of 2008 involving 9 out of 25 large banks cost the DIF $15.8 billion. The industry's assets have shrunk by 5.3% as a result of the banking crisis, which led to fewer loans (↓7.5%) and declines in the economic activity. Industry employment fell by 8.5% since 2007, translating into 188,000 lost jobs. Failed banks alone left 11,210 employees without jobs.[6]

In United States, FDIC and state banking regulatory authorities are responsible for the identification and resolution of insolvent institutions. A bank is considered at a risk of immediate closure if it is unable to fulfil its financial obligations the

[5]The Deposit Insurance Fund (DIF), which is the result of the merger of the Bank Insurance Fund (BIF) and the Saving Association Insurance Fund (SAIF) in 2006, requires each FDIC insured institution to pay an insurance premium. The amount of the insurance premium to the fund is determined based on institution's balance of insured deposits and the degree of risk it poses to the fund. The FDIC, as the receiver of the failed institution, liquidates its assets and compensates its depositors up to the insurance limit (currently $250,000). The amount not covered by the asset sales is provided by the DIF.

[6]Source: Wall Street Journal on-line article "Banks Keep Failing, No End in Sight" available at http://online.wsj.com.

next day or its capital reserves fall below the required regulatory minimum.[7] In the event of a bank failure, the FDIC either liquidates the assets of a failing bank and pays insurance to the depositors up to the amount of the insurance limit or arranges the sale of some or all of the bank's assets to another institution, which also may assume the part or all of the bank's liabilities.[8] The latter is oftentimes accomplished through financial assistance provided by regulators. The FDIC is required to resolve outstanding issues with problem banks in a manner that imposes the least cost on the deposit insurance fund and ultimately on the taxpayer. Thus, early detection of insolvent institutions is of vital importance, especially if the failure of those institutions would pose a serious systemic risk on the financial system and the economy as a whole. The FDIC and state authorities utilize on-site and off-site examination methods in order to determine which institutions are insolvent and thus should be either closed or be provided financial assistance in order to rescue them. The off-site examinations are typically based on statistical and other mathematical methods and constitute complementary tools to the on-site visits made by supervisors to institutions considered at risk. There are three advantages of off-site versus on-site examinations. First, the on-site examinations

[7]Under the current regulations issued by the Basel Committee on Banking Supervision (Basel II), a bank is considered as failed if its ratio of Tier 1 (core) capital to risk-weighted assets is 2% or lower. This ratio must exceed 4% to avoid supervisory intervention and prompt corrective actions as underlined in Federal Deposit Insurance Corporation Improvement Act (FDICIA) of 1992. A bank with ratio of 6% or above is considered as well-capitalized.

[8]The coverage limit is temporary set at $250,000 until the end of 2013 in order to protect the funds of depositors and prevent any potential depositor runs that could harm even the healthy institutions (Emergency Economic Stabilization Act, 2008). This limit will return to the permanent limit of $100,000 in 2014.

are more costly as they require the FDIC to bear the cost of visits and to retain extra staff during times when economic conditions are stable. Second, the on-site examinations are usually time-consuming and cannot be performed with high frequency during periods of wide-spread banking distress. Costs associated with on-site supervision of course increase with the number of troubled banks. Third, the off-site examinations can help allocate and coordinate the limited on-site examination resources in an efficient way with priority given to financial institutions facing the most severe challenges. The major drawback of the statistically based off-site tools is that they incorporate estimation errors which also affect the classification of banks as failures and nonfailures. An effective off-site examination tool must aim at identifying problem banks sufficiently prior to the time when a marked deterioration of their financial health would occur, which would force supervisors to undertake the necessary corrective actions needed to remedy the financial turmoil. Therefore, it is desirable to develop a model which would timely identify future failures with a high degree of accuracy and would not unnecessary flag healthy banks as being at risk of closure.

Accurate statistical models that serve as early warning tools and can be alternatives to or complementary to the costly on-site visits made by supervisors to institutions considered at risk have been well documented in the banking literature. Early warning models mostly refer to models that can identify and predict the realization of some event with high probability well in advance. These models have been successfully applied to study banking and other financial institutions'

failures in the U.S. and in other countries. As the literature that deals with bankruptcy prediction of financial and non-financial institutions is vast and there are a myriad of papers that specifically refer to the banking industry failures, we will discuss only few papers that are closely related to our work and are viewed as early warning models.

The more widely-used statistical models for bankruptcy prediction are the single-period static probit/logit models and methods of discriminant analysis.[9] These models usually estimate the probability that a firm with specific characteristics will fail or survive within a certain time interval. The timing of the failure is not provided by such models. Shumway (2001), in his bankruptcy prediction application, demonstrates with a simple example the inconsistency and the inefficiency (in a statistical sense) of these static models, as well as the superiority of a model such as his dynamic hazard model that utilizes multi-period observations. Others in this literature have employed the Cox proportional hazard model (PHM) to explain banking failures and develop early warning models.[10] Typically, in this model the dependent variable is time to occurrence of some specific event (failure in case of banks) which can be equivalently expressed either through the probability distribution function or the hazard function, the latter of which provides the

[9]For applications of the probit/logit models and the methods of discriminant analysis see Altman (1968), Meyer and Pifer (1970), Deakin (1972), Martin (1977), Lane et al. (1986), Cole and Gunther (1995, 1998), Cole and Wu (2010), among others.

[10]The thorough discussion of the Cox proportional hazard model can be found in Cox (1972), Lancaster (1990), Kalbfleisch and Prentice (2002), and Klein and Moeschberger (2003). The application of this model to study U.S. commercial banking failures is found in Lane et al. (1986), Whalen (1991), and Wheelock and Wilson (1995, 2000).

instantaneous risk of failure at some specific time conditional on the survival up to this time. The Cox PHM has three advantages over the static probit/logit models: (i) it provides not only the measure of probability of failure (survival) but also the probable timing of failure (ii) it accommodates censored observations, those observations that survive through the end of the sample period (iii) it does not make strong assumptions about the distribution of time to failure. The disadvantage of this model is that it requires the hazard rate to be proportional between any two cross-sectional observations and the inclusion of time-varying covariates is not as straightforward as with other models. To remedy these two shortcomings researchers recently turned their attention to the discrete time hazard model (DTHM).[11] The DTHM assumes that the failure occurs at discrete times and requires the covariates to be unchanged within a given time (month, quarter, or year). Inclusion of time-varying regressors that change over different periods allows for more efficient estimation and improved predictions as more recent prospective information is added to the retrospective information often used in less dynamic approaches.

In this chapter we develop an early warning model based on the Mixture Hazard Model (MHM) of Farewell (1977, 1982) with continuous and discrete time specifications.[12] MHM effectively combines the static model, which is used to

[11]See Shumway (2001), Halling and Hayden (2006), Cole and Wu (2009), and Torna (2010) for applications of discrete-time hazard models.

[12]Application of the discrete-time verison of the MHM are found in Gonzalez-Hermosillo et al.(1997), Yildirim (2008) and Topaloglu and Yildirim (2009).

identify insolvent banks, and the duration model, which provides estimates of the probability of failure along with the timing of closure of the troubled banks. In our study we view the financial crisis as a negative shock that affects banks in an unequal way. Well capitalized, well prepared, and prudently managed institutions may feel little relative distress during the financial turmoil. On the other hand, poorly managed banks that previously engaged in risk business practices will increase their probability of being on the FDIC watch list and subsequently forced into closure or merger with a surviving bank by regulatory authorities. Unlike the standard duration model, which assumes that all banks are at the risk of failure, we will implicitly assume that there is a proportion of banks that will survive for a sufficiently long time after the end of crisis and thus are not in this absorption state. In other words, we assume that the probability of failure for a bank that has never been on the watch list is arbitrarily close to zero. The MHM is appropriate in dealing with this issue as it is able to distinguish between healthy and at-risk of failure banks. Our model also recognizes the fact that insolvency and failure are two different events. The realization of the first event is largely attributed to the actions undertaken by the bank itself, while the second usually occurs as a result of regulators' intervention following their insolvency. Supervisors tend not to seize an insolvent bank unless it has no realistic probability of survival and its closure does not threaten the soundness and the stability of the financial system through its contribution to systemic risk.

One of our (testable) assumptions concerns the fact that banks with low performance, as calculated by the radial measure of realized outcome to the maximum potential outcome, will increase their probability of failure. Inefficiently managed banks could cumulatively save valuable funds by employing the best-practice technologies, which have shown to be important, especially during periods of banking crisis when money markets suffer from poor liquidity. Barr and Siems (1994) and Wheelock and Wilson (1995, 2000) were the first to consider the inefficiency as a potential influential factor explaining U.S. commercial banking failures during the earlier crisis. Barr and Siems (1994) estimate the efficiency scores with Data Envelopment Analysis (DEA) techniques, which are used in a static model to predict banking failures. Wheelock and Wilson (1995, 2000) estimate the Cox proportional hazard model with inefficiency scores included among other regressors, allowing inefficiency to affect the probability of failure as well as the probability being acquired by other bank. They employ three measures of radial technical inefficiency, namely the parametric cost inefficiency measure, the nonparametric input distance function measure, and the inverse of the nonparametric output distance function measure. The first two appear to have statistically significant positive effects on the probability of failure, while only the first measure significantly decreases the acquisition probability. The estimation of these models is conducted in two stages. The first stage involves the parametric or nonparametric estimation of inefficiency scores. In the second stage these scores are used as an explanatory variables in addition to other variables to investigate their effect on failure

probability. Tsionas and Papadogonas (2006) criticize the two-step approach as it may entail an error-in-variables bias as well as introduce an endogenous auxiliary regressor. They propose a single step joint estimation procedure to overcome these problems. We follow a similar approach.

Another challenge that we face in this study is the incomplete information associated with the troubled banks on the watch list of the FDIC. Each quarter the FDIC releases the number of problem banks but their names and identities are not disclosed. Based on our earlier assumption we can deduce that a bank that failed was on this list. Based on available information we make a prediction of which banks are on this list through an expectation-maximization (EM) algorithm which is designed to address this problem of missing information. Torna (2010), who also studies the recent U.S. commercial banking failures, identifies the number of troubled banks on the watch list through their tier 1 capital ranking. Banks are ranked according to their tier 1 capital and the number of banks with the lowest value are selected to match the number provided by FDIC in each quarter. Other ratios, such as Texas ratio, also can be utilized to deduce the problem banks. The Texas ratio was developed by Gerard Cassidy to predict banking failures in Texas and New England during recessionary periods of the 1980's and 1990's. It is defined as the ratio of nonperforming assets to total equity and loan-loss reserves. Banks with ratios close to one are identified as high risk. There are at least two limitations to these approaches besides their crude approximation. First, they ignore other variables that play a pivotal role in leading banks to a

distressed state. For example, the ratio of nonperforming loans is one of the major indicators of difficulties that bank will face in near future even if their capital ratio is at a normal level. Second, financial ratios that are used to classify banks as healthy or troubled cannot be subsequently employed as determinants due to possible endogeneity problem.

Another contribution of this study is that we follow a forward stepwise procedure in model building and selecting the relevant covariates that is not only based on the conventional measures of the goodness-of-fit and statistical tests but also on the contribution of these covariates to the predictive accuracy. As in Gonzalez-Hermosillo et al. (1997), we also include state-specific macroeconomic variables to control for factors that differentially impact particular states. The unequal distribution of banking failures among the states is revealed in Table 1. Industry-specific variables that could potentially capture the sector's condition as well as contagion effects cannot be identified in the Cox proportional hazard model and univariate probit/logit models. In the first case the constant in general is not identified, while in the second case these variables will be mixed with the constant term and thus will not be identified as well.

The remainder of the chapter is organized as follows. Section 2 describes the potential decision rule adopted by the regulatory authorities in determining and closing insolvent banks, which naturally will lead to the mixture hazard model (MHM). Two variants of the MHM are discussed, the continuous-time semiparametric proportional MHM and discrete-time MHM. In section 3 we discuss the

joint MHM-SFM. Section 4 deals with empirical specification issues and the data description. Estimation results for the model parameters and predictive accuracy are provided in section 5 along with a comparison of various models and specifications. Section 6 contains our main conclusions.

1.2. The Mixture Hazard Model

Before describing the mixture hazard model formulation in detail we establish a few definitions and describe the potential rules adopted by regulatory authorities to determine unsound banks that subsequently fail or survive. Other regulatory closure rules can be found in Kasa and Spiegel (2008). Let H_{it} define the financial health stock of bank i at time t and assume that there is a threshold level of it, H_{it}^{*}, such that if financial health falls bellow this level then the bank is considered at risk of closure by regulatory authorities. Formally, the difference between H_{it}^{*} and H_{it} can be represented as a function of bank-specific financial ratios, and structural and geographical macroeconomic variables

$$h_{it}^{*} = H_{it}^{*} - H_{it} = x_{it}'\beta + e_{it} \tag{1.1}$$

where e_{it} represents the error term, which is assumed to be identically and independently distributed across observations and over time.[13]

[13]The *iid* assumption of the error term can be relaxed in the panel data context by assuming $e_{it} = \mu_i + \xi_{it}$ with $\mu_i \sim N(0, \sigma_\mu^2)$ and $\xi_{it} \sim N(0, \sigma_\xi^2)$ independent of each other. This adds an additional complication to the model and it is not pursued in this paper.

We consider three simplified scenarios that can describe the path of financial health of a bank during periods of financial turmoil, which are represented in figure 1.3. Case I describes a situation in which bank financial health sharply declines and falls far below its threshold level. Banks in Case I are considered as high priorities by the FDIC and are placed at the top of the list of at-risk of failure banks. Case II is the scenario under which the bank experiences some difficulties and is considered as "troubled" by regulatory authorities. However, this bank recovers either by its own means or by receiving some financial assistance from regulators. Finally, Case III refers to a bank that is financially sound before and during crisis. With some very exceptional cases, such a bank will not be considered at a risk of closure during the current crisis. Notice that these scenarios are very simplified ones. In practice, banks can enter and exit the watch list multiple times or remain on the watch list for multiple quarters. In addition, the current health stock may depend on the history of its past realizations, which also affect the bank's survival. For the purposes of our further analysis what is required is that bank is to be considered troubled at least once during the sample period as we will assume that the probability that a healthy bank fails is close to zero or equivalently, a bank will not be seized by the FDIC unless it is considered as problem bank. This is the usual practice adopted by the FDIC.

The financial health of a particular bank is a composite and oftentimes a subjective index and its lower bound is not observable. Therefore, h_{it}^{*} is also not observable even to regulatory agencies that have only partial information about

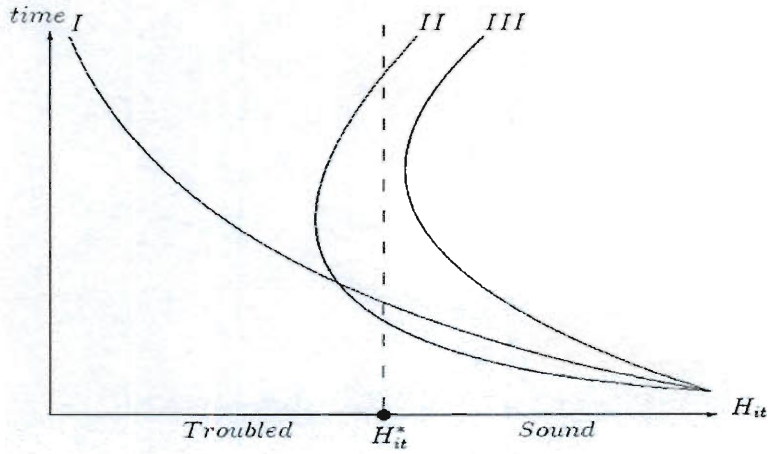


Figure 1.3. Illustration of the bank financial health condition during the period of financial turmoil

individual banks' financial health. Instead we can define a binary variable h_{it} such that

$$h_{it} = \begin{cases} 1 & \text{if } h_{it}^* > 0 \\ 0 & \text{if } h_{it}^* \leq 0 \end{cases}$$

Given this, the probability that a bank will become financially unhealthy is given by

$$\begin{aligned} p &= P(h_{it} = 1) = P(h_{it}^* > 0) \\ &= P(e_{it} > -x_{it}'\beta) = F_e(x_{it}'\beta) \end{aligned}$$

where F_e is the cumulative distribution function (cdf) of the random error e. If e is assumed to be standard normally distributed (probit model) then

$$F_e(x'_{it}\beta) = \int_{-\infty}^{x'_{it}\beta} (2\pi)^{-1/2} \exp(-t^2/2)dt$$

and

$$F_e(x'_{it}\beta) = \frac{\exp(x'_{it}\beta)}{1 + \exp(x'_{it}\beta)}$$

if e is logistically distributed (logit model) (McFadden 1974, 1981; Train, 2003).

However, as it is also discussed in the introductory section, information about a particular bank being at risk is not disclosed by regulator authorities. Therefore, h_{it} is only observed for banks that actually failed and is not observed for those that did not. Hence, we are faced with an incomplete information type problem as the only information that is available to us is the total number of problem institutions and not their names. In next section, we show how to deal with this type of missing information problem by transforming the incomplete elements into complete data. For now we treat the data as complete and, in order to derive the observed likelihood function, we further define, as in the standard hazard model, a nonnegative random variable T which represents the duration of a bank in a state of operation or the time until of occurrence of some specific event, such as failure

in our case.[14] This is characterized by the conditional probability density function (pdf), f_T and the cumulative distribution function (cdf), F_T. The survivor function of a particular bank given that it is characterized as a problem bank is then given by

$$S^p(t; w_i) = \Pr(T > t | h_i = 1; w_i) \tag{1.2}$$

S^p represents the probability that problem bank will survive for a period longer than t and w_i is the set of individual-specific, macroeconomic, structural, and geographical variables that are related to a bank's survival.

On the other hand, we let

$$F^p(t; w_i) = \Pr(T \leq t | h_i = 1; w_i) \tag{1.3}$$

represent the probability that a problem bank will fail by time t, which is the complement to the survivor function.

Likewise, we can define these probability measures for sound banks as

$$S^s(t; w_i) = \Pr(T > t | h_i = 0; w_i) \tag{1.4}$$

and

[14]Bank that ceased their operation due to reasons other than failure, such as merger and voluntary liquidation, or remained inactive or are no longer regulated by the Federal Reserve have censored duration times.

$$F^s(t; w_i) = \Pr(T \leq t | h_i = 0; w_i) \tag{1.5}$$

The survivor and failure functions of bank i are then expressed as

$$\begin{aligned} S(t; x_i, w_i) &= \Pr(T > t; x_i, w_i) = \Pr(T > t | h_i = 1; w_i) \Pr(h_i = 1; x_i) \\ &+ \Pr(T > t | h_i = 0; w_i) \Pr(h_i = 0; x_i) \end{aligned} \tag{1.6}$$

and

$$\begin{aligned} F(t; x_i, w_i) &= \Pr(T \leq t; x_i, w_i) = \Pr(T \leq t | h_i = 1; w_i) \Pr(h_i = 1; x_i) \\ &+ \Pr(T \leq t | h_i = 0; w_i) \Pr(h_i = 0; x_i). \end{aligned} \tag{1.7}$$

Let the binary variable d_i take on a value of 1 for observations that fail at time t and 0 for observations that are right censored when the bank does not fail by the end of the sample period or disappears during the period for reasons other than failure (mergers, acquisitions, incomplete data, etc). Then the likelihood function for bank i is given by

$$
\begin{aligned}
L(\theta; x, w) &= f(t; x_i, w_i)^{d_i} S(t; x_i, w_i)^{1-d_i} \\
&= [F_e(x_i'\beta) f^p(t; w_i) + (1 - F_e(x_i'\beta)) f^s(t; w_i)]^{d_i} [F_e(x_i'\beta) S^p(t; w_i) \\
&\quad + (1 - F_e(x_i'\beta)) S^s(t; w_i)]^{1-d_i}
\end{aligned}
$$

where θ is the parameter vector, and x and w are covariates associated with the probability of being troubled and of having failed, respectively.

What is generally observed for the failed U.S. commercial banks is that prior to their failure they experience an extensive period of financial distress. Regulatory authorities tend to close banks that have been characterized as troubled either by means of off-site or on-site examinations. Therefore, we assume that the probability that a healthy bank fails instantaneously is arbitrarily close to zero and thus the hazard rate for such banks is also arbitrarily close to zero. Hence, the above likelihood function reduces to

$$
L_i(\theta; x, w) = [F_e(x_i'\beta) \lambda_i^p(t; w_i) S^p(t; w_i)]^{d_i} [F_e(x_i'\beta) S^p(t; w_i) + (1 - F_e(x_i'\beta))]^{1-d_i} \tag{1.8}
$$

where

$$
\lambda^p(t; w_i) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T > t, h_i = 1; w_i)}{\Delta t} = \frac{f^p(t; w_i)}{S^p(t; w_i)}
$$

represents the hazard rate or probability that a troubled bank will fail during the period $(t, t + \Delta t)$, given that it was in operation for t periods earlier. The hazard rate, survivor and failure functions are mathematically equivalent in the sense that they convey the same information about the distribution of duration times and specifications of any can be derived from the specification of one of these distributions.

After rearranging the expression in (1.8) and taking the product over all cross-sectional units, the sample likelihood function is given by[15]

$$\begin{aligned} L(\theta; x, w, d) &= \prod_{i=1}^{n} L_i(\theta; x, w, d) \\ &= \prod_{i=1}^{n} F_e(x_i'\beta)^{h_i}(1 - F_e(x_i'\beta))^{1-h_i}\{\lambda_i(t; w_i)\}^{d_i h_i}\{S_i(t; w_i)\}^{h_i} \end{aligned} \tag{1.9}$$

where we drop the superscript from measures pertaining to problem banks in what follows. If h_i is completely observed for each individual bank, as it is to regulators, then the log-likelihood can be maximized by conventional techniques of maximum likelihood estimation (MLE) to obtain consistent and asymptotically efficient estimates of the model parameters. Given that h_i is partially observed to outside observers we need to address this incomplete information utilizing other methods and thus turn to the expectation-maximization (EM) algorithm or simulated maximum likelihood (SML), which are designed to handle problems of incomplete

[15]The derivation of the likelihood function is provided in the Appendix A.

sample information. In the next section we describe the continuous-time semiparametric mixture hazard model under the assumptions of proportional hazard (Cox, 1972) and the absence of any effect of inefficiency measures on the probability and the timing of event, as well as the full implementation of the EM algorithm. These assumptions are subsequently relaxed in following sections where we consider discrete-time mixture hazard model with time-varying covariates, as well as the performance (efficiency) of each bank as a factor explaining their difficulties and failures. We will refer to the continuous-time mixture hazard model as Model I and the discrete-time mixture hazard model as Model II. Following the standard nomenclature of the medical and biological sciences where the MHM was initially applied, we will often refer to the logistic part of the model as the incidence part and to the hazard part as the latency part. We will use these terms interchangeably throughout the chapter.

1.2.1. Semiparametric Continuous-Time Proportional Mixture Hazard Model

Following Kuk and Chen (1992) and Sy and Taylor (2000) we specify a semiparametric proportional mixture hazard model with the full log-likelihood function for i^{th} individual bank with observed data (t_i, d_i, x_i, w_i) expressed by

$$
\begin{aligned}
L_i(\theta; x, w, d) &= \log L_i(\theta, \lambda_0; x, w, d) = h_i \log(F_e(x_i'\beta)) \\
&\quad +(1 - h_i)\log(1 - F_e(x_i'\beta)) + d_i h_i \log(\lambda_i(t; w_i)) + h_i \log(S_i(t; w_i)) \\
&= L_{1i}(\beta; x_i, h_i) + L_{2i}(\alpha, \lambda_0; w_i, h_i)
\end{aligned}
$$

with

$$\lambda_i(t; w_i) = \lambda_0(t)\exp(w_i'\alpha) \tag{1.10}$$

and

$$S_i(t; w_i) = S_0(t)^{\exp(w_i'\alpha)} \tag{1.11}$$

where $\lambda_0(t)$ and $S_0(t)$ are unspecified conditional baseline hazard and baseline survivor functions, respectively. These are non-negative functions of time only and are assumed to be common for all individuals at risk. The censoring is assumed to be noninformative and statistically independent of the events of distress and failure.

Given that h_i is only partially observed, the SPMHM can be estimated by the Expectation-Maximization (EM) algorithm. The EM algorithm is an efficient iterative procedure for maximizing complex likelihood functions and handling incomplete or missing data. Each iteration of the algorithm consists of two steps: expectation (E) and maximization (M) step. The expectation step involves the

projection of an appropriate functional (likelihood or log-likelihood function) containing the augmented data on the space of the original, incomplete data. That is, the missing data are first estimated given the observed data and a current estimate of the model parameters. In the maximization step the function is maximized while treating the incomplete data as known. Iterating between these two steps yields estimates that under suitable regularity conditions converge to the maximum likelihood estimates (MLE). For more discussion on the EM algorithm and its convergence properties see Dempster et al. (1977) and McLachlan and Krishnan (1996).

To implement the EM algorithm we first need to take the expectation of the full log-likelihood function with the respect to h_i and the data, which completes the E-step of the algorithm. Linearity of $L(\cdot)$ with respect to h_i in this case facilitates the calculations and analysis considerably.

The log-likelihood function of i^{th} observation in the M-step is given by

$$\begin{aligned} E^{(M)}_{h|X,W,\theta,\lambda_0}\left[L_i(\theta; x, w, d)\right] &= \tilde{h}_i^{(M)} \log(F_e(x_i'\beta)) + (1 - \tilde{h}^{(M)}) \log(1 - F_e(x_i'\beta)) \\ &\quad + \tilde{h}_i^{(M)} d_i \log(\lambda_i(t; w_i)) + \tilde{h}_i^{(M)} \log(S_i(t; w_i)) \end{aligned}$$

where $\tilde{h}_i$ is the probability that the i^{th} bank will eventually belong to the group of problem banks conditioned on observed data and the model parameters. It represents the fractional allocation to the problem banks and is given by

$$\begin{aligned}\tilde{h}_i^{(M)} &= E\left[h_i|\theta^{(M)}, Data\right] = \Pr(h_i^{(M)} = 1|t_i > T_i) \\ &= \begin{cases} \frac{F_e(x_i'\beta^{(M)})S_i(t;w_i)}{F_e(x_i'\beta^{(M)})S_i(t;w_i)+(1-F_e(x_i'\beta^{(M)}))} & \text{if } d_i = 0 \\ 1 \quad \text{otherwise} \end{cases}\end{aligned} \tag{1.12}$$

The observed full likelihood function (1.9) is then expressed by

$$\begin{aligned}L(\theta; x, w, \tilde{h}^{(M)}) &= \prod_{i=1}^{n} F_e(x_i'\beta)^{\tilde{h}_i^{(M)}}(1 - F_e(x_i'\beta))^{1-\tilde{h}_i^{(M)}} \\ &\quad \{\lambda_0(t)\exp(w_i'\alpha)\}^{d_i\tilde{h}_i^{(M)}}\{\exp(-\tilde{h}_i^{(M)}\Lambda_0\exp(w_i'\alpha))\}\end{aligned} \tag{1.13}$$

where $\Lambda_0 = \int_0^t \lambda_0(v)dv$ is the baseline cumulative hazard function. The nuisance baseline hazard function λ_0 is not specified parametrically. It is estimated non-parametrically from the profile likelihood function as

$$\hat{\lambda}_0(t) = \frac{N(t_i)}{\sum_{j\in R(t_i)} \tilde{h}_j \exp(w_j'\alpha)} \tag{1.14}$$

and the baseline cumulative hazard function is then calculated as

$$\hat{\Lambda}_0(t) = \sum_{t_i \le t} \frac{N(t_i)}{\sum_{j\in R(t_i)} \tilde{h}_j \exp(w_j'\alpha)} \tag{1.15}$$

where $N(t_i)$ is the number of failures and $R(t_i)$ is the set of all individuals at risk at time t_i, respectively. Notice that in standard model where $h = 1$ with probability one (1.15) reduces to Breslow's (1972) estimator in the case when ties are present at time t_i.[16] Substituting (1.14) and (1.15) into (1.13) leads to the M-step log-likelihood

$$\begin{aligned}\tilde{L}(\theta; x, w, \tilde{h}) &= \sum_{i=1}^{n} \{\tilde{h}_i \log F_e(x_i'\beta) + (1 - \tilde{h}_i) \log(1 - F_e(x_i'\beta))\} \\ &\quad + \sum_{i=1}^{N} \{w_i'\alpha - N(t_i) \log \left(\sum_{j \in R(t_i)} \tilde{h}_j \exp(w_j'\alpha) \right)\} \\ &= L_1(\beta; x, \tilde{h}) + \tilde{L}_2(\alpha; w, \tilde{h})\end{aligned} \tag{1.16}$$

The second term in above expression is the Cox-type partial log-likelihood function for parameter α that handles the ties using the Peto (1972) and Breslow (1974) approximation method.

The full implementation of the EM algorithm involves the following four steps:

- Step 1: Provide an initial estimate for the parameter β and estimate the ordinary Cox partial likelihood model to obtain the starting values for α and $\hat{\lambda}_0$.

[16] See Johansen (1983), Sy and Taylor (2000), and Klein and Moeschberger (2003) on this argument. Sy and Taylor (200) propose an alternative of the product-limit estimator to estimate the nuisance baseline hazard function which also can handle the zero-tail constraint in the survivor function for the last event time. It is empirical issue whether this constraint holds for the Breslow type estimate of the survivor function.

- Step 2 (E-step): Compute $\tilde{h}_i$ from (1.12) based on the current estimates and the observed data.
- Step 3 (M-step): Update the estimate of parameter β using L_1 and update the estimate of parameter α and hence $\hat{\lambda}_0$ using the partial log-likelihood $\tilde{L}_2$ and equation (1.14), respectively.
- Step 4: Iterate between steps 2 and 3 until convergence is reached.

After the estimates of model parameters are obtained their standard errors are calculated from the inverse of the standard information matrix. Sy and Taylor (2000) provide the components of this matrix based on the complete data log-likelihood function in their appendix.

1.2.2. Discrete-Time Mixture Hazard Model with time-varying covariates

In order to incorporate time-varying regressors in the model we consider the discrete-time mixture hazard model.[17] However, this requires that these regressors remain unchanged in the time window $[t, t+1]$, an assumption which is tenable if we consider the fact that banks report their data on a quarterly basis. The hazard rate in the discrete-time hazard model is given by

[17] See Cox and Oaks (1984), Kalbfleisch and Prentice (2002), and Bover et al. (2002) for discussion on discrete-time proportional hazard models.

$$P(t \leq T < t+1 | T > t, h_{it} = 1) = 1 - \exp\left[-\exp(w'_{it}\alpha) \int_t^{t+1} \lambda(v)dv\right]$$
$$= 1 - \exp\left(-\exp(\omega_t + w'_{it}\alpha)\right) = F^*\left(\omega_t + w'_{it}\alpha\right) \quad (1.17)$$

where $\omega_t = \ln \int_t^{t+1} \lambda(v)dv$ and $F^*(\cdot)$ is the extreme value cumulative distribution function. If we assume a parametric function for the baseline hazard function, then ω_t will also be a parametric function of time. Meyer (1990) proposes methods of estimating such models, both parametrically and nonparametrically.

The hazard rate in (1.17) can also be expressed in terms of the logistic distribution if we note that

$$\ln \int_t^{t+1} \lambda(v; w)dv = \ln\left[1 + \exp q(t; w)\right] \quad (1.18)$$

where

$$q(t; w) = \ln\left(\frac{1 - \exp(-\exp(\Lambda(t; w))}{\exp(-\exp(\Lambda(t; w)}\right) \quad (1.19)$$

which implies that $q(\cdot)$ is logistically distributed and can be linearly approximated by $w'_{it}\alpha$. Hence,

$$P(t \leq T < t+1 | T > t, h_{it} = 1; w_{it}) = \frac{\exp(w'_{it}\alpha)}{1 + \exp(w'_{it}\alpha)} \quad (1.20)$$

By noting that $\lambda_{ij}(t;w) = 1 - \frac{S(t_{ij})}{S(t_{i,j-1})}$ for $j = 1, 2, ..., t_i$, and writing the survivor function as the product of conditional survival probabilities $S_{ij}(t;w) = \prod_{j=1}^{t_i} \frac{S(t_{ij})}{S(t_{i,j-1})}$ with $S(t_{i0}) = 1$, we have

$$S_{ij}(t;w,u) = \prod_{j=1}^{t_i} \left(\frac{1}{1+\exp(w_{ij}'\alpha)} \right), \tag{1.21}$$

which relates the survivor function to the hazard function and is a nonincreasing step function of time. By substituting (1.20) and (1.21) into (1.9) we obtain the likelihood function for the Discrete-Time Mixture Hazard Model (DTMHM).

1.3. Stochastic Frontier Model combined with Mixture Hazard Model

In this section we consider the performance of an individual bank as a determinant of both the probability of being troubled and the timing of the event of failure. The efficiency performance of a firm relative to the best practice (frontier) technology was formally considered by Debreu (1951) and Farrell (1957). Aigner and Chu (1968) proposed a deterministic frontier model where the performance was measured parametrically by the deviation of the observed outcome from the optimal outcome. This formulation suffered from an assumption that the entire deviation from the ideal frontier outcome was solely attributed to inefficiency, which was under the control of the firm. Aigner et al. (1977), Meeusen and van den Broeck (1977), and Battese and Cora (1977) introduced the parametric stochastic

frontier model (SFM).[18] In SFM the error term is assumed to be multiplicative and composed of two parts; a one-sided term that captures the effects of inefficiencies relative to the stochastic frontier and a two-sided term that captures random shocks, measurement errors and other statistical noise, and allows random variation of frontiers across firms. The initial SFM was formulated in a cross-sectional context and was later extended to panel data models, which allow the researcher to consistently unconditional efficiency scores. Excellent surveys on frontier models and their applications are found in Kumbhakar and Lovell (2000) and Greene (2007).

The general stochastic frontier panel model for the i^{th} firm is given by

$$y_{it} = g(z_{it};\gamma)\exp(\varepsilon_{it}) \tag{1.22}$$

where the dependent variable y_{it} could represent cost, output, profit, revenue etc, z_{it} is a vector of independent regressors, and $g(\cdot)$ is the frontier function, which can be either linear or non-linear in coefficients and covariates. Depending on the particular dual representation of technology specified, $\varepsilon = \upsilon \pm u(= \log y_{it} - \log g(z_{it};\gamma))$ represents the composed error term, with υ_{it} representing the noise and u_i the inefficiency process. The noise term is assumed to be iid normally distributed with zero mean and constant variance. Inefficiencies are also assumed to be iid random variables with distribution function defined on the domain of

[18] Nonparametric alternatives measuring inefficiency were introduced by Charnes et al. (1978) and are generically referred as data envelopment analysis (DEA). Deprins et al. (1984) generalized the DEA model to what is referred to as the class of free disposal hull (FDH) models.

positive numbers ($u \in R_+$). Both, v and u, are assumed to be independent from each other and from regressors.[19] We follow Pitt and Lee (1981) and assume that the inefficiency process is a time-invariant random effect which follows the half-normal distribution ($u_i \sim N^+(0, \sigma_u^2)$).

Under the above assumptions the joint distribution of the noise and the inefficiency term is given by

$$\begin{aligned} f_{v,u}(v_{it}, u_i) &= f_v(v_{it}) f_u(u_i) = f_v(\varepsilon_{it} \pm u_i) f_u(u_i) \\ &= \frac{2}{(2\pi)^{(T_i+1)/2} \sigma_v^{T_i} \sigma_u} \exp\left[-\frac{(\varepsilon_{it} \pm u_i)'(\varepsilon_{it} \pm u_i)}{2\sigma_v^2} - \frac{u_i^2}{2\sigma_u^2} \right] \end{aligned}$$

After integrating u_i from this expression we obtain the marginal density of the composed error term, which for the production or profit frontier model is derived as

$$f_\varepsilon(\varepsilon_{it}) = \frac{2}{(2\pi)^{T_i/2} \sigma_v^{T_i-1} \sigma} \exp\left[-\frac{\varepsilon_{it}' \varepsilon_{it}}{2\sigma_v^2} + \frac{\bar{\varepsilon}_i^2 \lambda^2}{2\sigma^2} \right] \left[1 - \Phi\left(\frac{T_i \bar{\varepsilon}_i \lambda}{\sigma} \right) \right] \tag{1.23}$$

where $\sigma = \sqrt{\sigma_v^2 + T_i \sigma_u^2}$, $\lambda = \sigma_u / \sigma_v$, and $\bar{\varepsilon}_i = (1/T_i) \sum_{t=1}^{T_i} \varepsilon_{it}$.[20] The parameter λ is the signal-to-noise ratio and measures the relative allocation of total variation

[19] The assumption of independence of the inefficiency term and the regressors is restrictive but is necessary for our current analysis. Its validity can be tested with Hausman-Wu specification test. In the panel data context this assumption can be relaxed by assuming that inefficiencies are fixed effects or random effects correlated with all or some of the regressors (Hausman and Taylor, 1981; Schmidt and Sickles, 1984; Cornwell et al., 1990).

[20] The cost frontier is obtained by reversing the sign of the composed error.

to the inefficiency term. In practice we can use an alternative parameterization called the $\gamma-$parameterization which specifies

$$\gamma = \frac{\sigma_u^2}{\sigma^2}$$

This reparameterization is more desirable as γ has compact support which facilitates the numerical procedure of maximum likelihood estimation, hypothesis testing, and establishing the asymptotic normality of this parameter.

It can be also shown (see Jondrow et al., 1982) that the conditional distribution of the inefficiency term is given by

$$f_{u|\varepsilon}(u_i|\varepsilon_{it}) = \frac{f_{\varepsilon,u}(\varepsilon_i, u_i)}{f_\varepsilon(\varepsilon_i)} = \frac{\frac{1}{\sigma}\phi(\frac{u_i-\mu_i^*}{\sigma_*})}{\left[1-\Phi\left(-\frac{\mu_i^*}{\sigma_*}\right)\right]} \tag{1.24}$$

where $f_{u|\varepsilon}(\cdot)$ represents the normal distribution truncated at 0 with mean $\mu_i^* = -T_i\bar{\varepsilon}_i\sigma_u^2/\sigma^2 = -T_i\bar{\varepsilon}_i\gamma$ and variance $\sigma_*^2 = \sigma_u^2\sigma_v^2/\sigma^2 = \gamma\sigma^2(1-\gamma T_i)$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively. The mean or the mode of this conditional distribution provides an estimate of the technical inefficiency of each firm in the sample. Horrace and Schmidt (1996) derive the prediction interval for inefficiency scores based on quantiles of $f_{u|\varepsilon}(\cdot)$.

In the absence of any effect of the inefficiencies on the probability and timing of failure, (1.23) and (1.24) can be employed to obtain the maximum likelihood estimates of model parameters and efficiency scores. However, consistent and efficient parameter estimates cannot be based solely on the frontier model when there

is feedback between this measure of economic frailty, and the likelihood of failure and the ensuing tightening of regulatory supervision. There is a clear need for joint estimation of the system when the decision of firm is affected by these factors.

Formally, we will assume as above that the censoring is noninformative and statistically independent of h_i. Following Tsionas and Papadogonas (2006) we will also assume that the censoring and h_i are independent of the composed error term, conditional on inefficiency and the data. Hence, given h_i, the observed joint density function of the entire system, after integrating out the latent inefficiency, can be written as

$$\begin{aligned} L_i(y_i, h_i, d_i|\Omega_i, \theta') &= \int_0^\infty F_e(x_i'\beta + \delta_1 u_i)^{h_i}(1 - F_e(x_i'\beta + \delta_1 u_i))^{1-h_i} \\ &\quad \times\{\lambda_i(t; w_i, u_i)\}^{d_i h_i}\{S_i(t; w_i, u_i)\}^{h_i} \underbrace{f_\upsilon(\varepsilon_{it} \pm u_i) f(u_i)}_{f_\varepsilon(\varepsilon) f_{u|\varepsilon}(u|\varepsilon)} du_i \\ &= f_\varepsilon(\varepsilon_{it}) \int_0^\infty F_e(x_i'\beta + \delta_1 u_i)^{h_i}(1 - F_e(x_i'\beta + \delta_1 u_i))^{1-h_i} \\ &\quad \times\{\lambda_i(t; w_i, u_i)\}^{d_i h_i}\{S_i(t; w_i, u_i)\}^{h_i} f_{u|\varepsilon}(u|\varepsilon) du_i \end{aligned} \tag{1.25}$$

The hazard rate and survival function for SPMHM are now given by

$$\lambda_i(t; w_i, u_i) = \lambda_0(t) \exp(w_i'\alpha + \delta_2 u_i)$$

and

$$S(t; w_i) = S_0(t)^{\exp(w_i'\alpha + \delta_2 u_i)}$$

respectively and $\Omega_i = \{x_i, w_i, z_i\}$ denotes the set of covariates, while θ' is the vector of the structural and distributional parameters.

This is a general model which combines the stochastic frontier model with logistic regression and the proportional hazard model. Either of these three models are special cases. If, for example, there is no association between inefficiency and probability of being troubled or failed ($\delta = (\delta_1, \delta_2) = (0, 0)$), then (1.25) consists of two distinct parts, the stochastic frontier and the mixture hazard. Both can be estimated separately using the previously outlined methods.

The integral in the joint likelihood (1.25) has no closed form solution and thus the maximization of this function requires numerical techniques, such as simulated maximum likelihood (SML) or Gaussian quadrature.[21] In SML the sample of draws from $f_{u|\varepsilon}(\cdot)$ are required to approximate the integral by its numerical average (expectation). As such, the simulated log-likelihood function for the i^{th} observation becomes

[21] Tsionas and Papadogonas (2006) employ the gaussian quadrature in estimation of the model wherein the technical inefficiency has potential effect on firm exit.

$$
\begin{aligned}
L_i \;=\; & \log L_i(y_i, h_i, d_i|\Omega_i, \Theta') = Constant - \frac{(T_i - 1)}{2}\log\sigma^2(1-\gamma T_i) \qquad (1.26)\\
& -\frac{1}{2}\log\sigma^2 + \log\left(1 - \Phi\left(\frac{T_i\bar{\varepsilon}_i\lambda}{\sigma}\right)\right) - \frac{\varepsilon_{it}'\varepsilon_{it}}{2\sigma^2(1-\gamma T_i)} + \frac{\bar{\varepsilon}_i^2\gamma}{2\sigma^2(1-\gamma)}\\
& + \log\frac{1}{S}\sum\nolimits_{s=1}^{S}\{F_e(x_i\beta + \delta_1 u_{is})^{h_i}(1 - F_e(x_i\beta + \delta_1 u_{is}))^{1-h_i}\\
& \qquad \times(\lambda_i(t; w_i, u_{is}))^{d_i h_i}(S_i(t; w_i, u_{is}))^{h_i}\}
\end{aligned}
$$

where u_{is} is a random draw from the truncated normal distribution $f_{u|\varepsilon}(\cdot)$ and S is the number of draws. We utilize the inverse cdf method to efficiently obtain draws from this distribution as

$$
u_{is} = \mu_i^* + \sigma_*\Phi^{-1}\left[U_{is} + (1 - U_{is})\Phi(-\frac{\mu_i^*}{\sigma_*})\right] \qquad (1.27)
$$

where U is a random draw from uniform $U[0,1]$ distribution or a Halton draw.

By substituting (1.27) into (1.26) and treating the $h_i's$ as known we can maximize the log-likelihood function $L = \sum_i L_i$ by employing standard optimization techniques and obtain the estimates of the model parameters.

Finally, after obtaining the model parameters, the efficiency scores are obtained as the expected values of the conditional distribution in the spirit of Jondrow et al. (1982)

$$\hat{u}_i = E\left[u_i|\hat{\varepsilon}_i, \tilde{h}_i, d_i\right] = \frac{\int_0^\infty u_i G(u_i;\Theta) f_{u|\varepsilon}(u|\varepsilon) du_i}{\int_0^\infty G(u_i;\Theta) f_{u|\varepsilon}(u|\varepsilon) du_i} \tag{1.28}$$

$$G(u_i;\Theta) = \tilde{F}(x_i'\beta + \delta_1 u_i)^{\tilde{h}_i}(1 - \tilde{F}(x_i'\beta + \delta_1 u_i))^{1-\tilde{h}_i}\{\lambda_i(t; w_i, u_i)\}^{d_i\tilde{h}_i}\{S(t; w_i, u_i)\}^{\tilde{h}_i}.$$

The integrals in the numerator and denominator are calculated numerically by SML method. It is straightforward to check that if δ is zero then (1.28) collapses to the Jondrow et al. formula for production frontiers

$$\hat{u}_i = E\left[u_i|\hat{\varepsilon}_i\right] = \mu_* + \sigma_* \frac{\phi(\frac{\mu_*}{\sigma_*})}{\Phi(\frac{\mu_*}{\sigma_*})} \tag{1.29}$$

The predicted efficiency of i^{th} firm is given by $TE_i = \exp(-\hat{u}_i)$ or it can be calculated as $TE_i^* = E\left[\exp(-u_i)|\hat{\varepsilon}_i, \tilde{h}_i, d_i\right]$ as suggested by Battese and Coeli (1988). The latter measure is optimal in the sense that it gives lower mean squared error of prediction than the latter one.

The EM algorithm for the stochastic frontier mixture model involves the following steps:

- Step 1: Provide initial estimates of the parameter vector θ'. Set the initial value of parameters δ_1 and δ_2 equal to zero and obtain the initial value of the baseline hazard function from (1.14). Consistent starting values of the variances of the noise and inefficiency terms are based on method of

moments estimates

$$\begin{aligned}\hat{\sigma}_u^2 &= [\sqrt{2/\pi}(\frac{\pi}{\pi-4})\hat{m}_3]^{2/3} \\ \hat{\sigma}_v^2 &= \hat{m}_2 - (\frac{\pi-2}{\pi})\hat{\sigma}_u^2\end{aligned}$$

where $\hat{m}_2$ and $\hat{m}_3$ are the estimated second and third sample moments of the OLS residuals, respectively. Estimates of σ and γ parameters are obtained through the relevant expressions provided above.

- Step 2 (E-step): Compute $\tilde{h}_i$ based on the current estimates and the observed data from

$$\tilde{h}_i^{(M)} = E\left[h_i|\theta^{(M)}, Data\right] = \Pr(h_i^{(M)} = 1|t_i > T_i)$$

$$\begin{cases} \frac{F_e(x_i'\beta^{(M)}+\delta_1^{(M)}u_i)S_i(t;w_i,u_i)}{F_e(x_i'\beta^{(M)}+\delta_1^{(M)}u_i)S_i(t;w_i,u_i)+(1-F_e(x_i'\beta^{(M)}+\delta_1^{(M)}u_i))} & \text{if } d_i = 0 \\ 1 & \text{otherwise} \end{cases}$$

- Step 3 (M-step): Update the estimate of parameters by maximizing L via simulated maximum likelihood technique.
- Step 4: Iterate between steps 2 and 3 until convergence.

1.4. Empirical Model and Data

In this section we outline the empirical specification that we follow to estimate the four models. We also describe the data used in this study and the stepwise forward selection procedure that is employed in model building and the variable selection.

1.4.1. Empirical Specification

Following Whalen (1991) we employ a model with a two-year timeline to estimate the probability of distress and failure, as well as the timing of the failure of the bank with a certain set of characteristics. In the SPMHM the time to failure is measured in months (1-24) starting from December 31, 2007. The sample consists of 125 banks that failed during 2008 and 2009 and 5,843 nonfailed banks. The covariates used in estimation are derived from 2007.Q4 Consolidated Reports of Condition and Income (Call Reports). Banks that disappear from the sample for reasons other than failure or did not experience the event through the end of the sample period have censored duration times. That is, for these banks what we observe is the maximum duration in the sample but there is no further information on their status. Nine banks voluntarily liquidated and are excluded from the sample since we are modeling the regulatory decision. In addition, we exclude banks that were chartered and started to report their data after the first quarter of 2007. These are typically referred as "de novo" banks and require a special treatment (DeYoung,

1999, 2003). The holdout sample consists of 92 banks that failed during 2010 including the third quarter, as well as 5,674 surviving banks. This sample will be used to assess the model's out of sample predictive accuracy.

For the DTMHM time to failure is measured in quarters as banks report their data on a quarterly basis. The sample consists of eight quarters of observations on banks that either failed or survived during the 2008-2009 period. The number of failed banks is the same as in the SPMHM. The holdout sample consists of two quarters of bank observations in 2010. Rather than calculating the probable time of failure, in this case we estimate the probability that a certain bank will fail during 2010. This excludes the fourth quarter of 2010 due to lack of data.

We employ the cost frontier in the Stochastic Frontier specification. The cost frontier describes the minimum level of cost given a certain output level and input prices. It is dual to the production frontier, and it is oftentimes used to describe the technology employed by the firms in regulated industries (Shephard, 1953). The gap between the actual and minimum cost is a measure of total (cost) inefficiency which is composed of two parts: technical inefficiency, which arises from excess usage of inputs, and allocative inefficiency, which results from a non-optimal mix of inputs. We do not make this decomposition but rather estimate overall cost inefficiency. We adopt the intermediation approach of Sealey and Lindley (1977) according to which banks are viewed as financial intermediaries that collect deposits and other funds and transform them into loanable funds by using capital and labor. Deposits are viewed as inputs as opposed to outputs, which is assumed

in the production and value-added approaches (Baltensperger, 1980; Berger and Humphrey, 1992).

As in Kaparakis et al. (1994) and Wheelock and Wilson (1995) we specify a multiple output-input short-run stochastic cost frontier with a quasi-fixed input. Following the standard banking literature we specify a translog functional form to describe the cost function[22]

$$
\begin{aligned}
\ln C_{it} \;=\; & \alpha_{0+} \sum_{m=1}^{5} \alpha_m \ln y_{mit} + \sum_{k=1}^{4} \beta_k \ln w_{kit} \\
& + \frac{1}{2} \sum_{m=1}^{5} \sum_{j=1}^{5} \alpha_{mj} \ln y_{mit} \ln y_{jit} + \theta_1 t + \frac{1}{2} \theta_2 t^2 \\
& + \frac{1}{2} \sum_{k=1}^{4} \sum_{n=1}^{4} \beta_{kn} \ln w_{kit} \ln w_{nit} + \eta_1 \ln X_{it} + \frac{1}{2} \eta_2 (\ln X_{it})^2 \\
& + \sum_{m=1}^{5} \sum_{k=1}^{4} \delta_{mk} \ln y_{mit} \ln w_{kit} + \sum_{m=1}^{5} \lambda_{1x} \ln y_{mit} \ln X_{it} \\
& \sum_{k=1}^{4} \lambda_{2x} \ln w_{kit} \ln X_{it} + \sum_{m=1}^{5} \lambda_{mt} \ln y_{mit} t + \sum_{k=1}^{4} \phi_{kt} \ln w_{kit} t + \upsilon_{it} + u_i
\end{aligned}
$$

[22]Translog function provides a second-order differential approximation to an arbitrary function at a single point. It does not restrict the share of a particular input to be constant over time and across individual firms. Additional flexibility can be attained by considering the Fourier-flexible functional form, which includes Fourier trigonometric terms in addition to the standard translog terms. It requires specific truncation of the data and as it is documented in Berger and Mester (1997) there is no essential difference in average efficiencies and ranking of firms between this and the translog functional specification.

with symmetry and linear homogeneity in input price restrictions imposed by considering capital as the numeraire and dividing the total cost and other input prices by its price. Thus

$$\alpha_{mj} = \alpha_{jm} \text{ and } \beta_{kn} = \beta_{nk}$$

$$\sum_{k=1}^{4} \beta_k = 1 \ , \ \sum_{k=1}^{4} \beta_{kn} = \sum_{k=1}^{4} \delta_{mk} = \sum_{k=1}^{4} \lambda_{2x} = \sum_{k=1}^{4} \phi_{kt} = 0$$

where C is the observed short-run variable cost of an individual bank at each time period, y_m is the value of m^{th} output, $m = 1, .., 5$. Outputs are real estate loans ($yreln$), commercial and industrial loans ($yciln$), installment loans ($yinln$), securities ($ysec$), and off-balance sheet items ($yobs$). The $w's$ represent input prices for total interest-bearing deposits (dep), labor (lab), purchased funds ($purf$), and capital (cap). The quasi-fixed input (X) consists of noninterest-bearing deposits. Kaparakis et al. (1994) assume that the bank takes the level of noninterest-bearing deposits as exogenously given and since there is no market price associated with this input, the quantity of it should be included in the cost function instead of its price. We also include the time and its interaction with outputs and input prices to account for non-neutral technological change.

After the technology parameters are estimated we can estimate the scale economies and technological change measures along a particular output/price ray. The scale

economies are defined as the degree to which a firm's total cost of producing financial services decreases as its output of services increase proportionally and they are derived as the sum of partial derivatives of the cost with respect the outputs. That is,

$$\begin{aligned} Scale_{it} &= \sum_{m=1}^{5} \frac{\partial \ln C_{it}}{\partial \ln y_{mit}} \\ &= \sum_{m=1}^{5} \left[\alpha_m + \sum_{j=1}^{5} \alpha_{mj} \ln y_{jit} + \sum_{j=1}^{4} \delta_{mk} \ln w_{kit} + \lambda_{1x} \ln X_{it} + \lambda_{mt} t \right] \end{aligned} \tag{1.30}$$

A value of this measure less than one indicates the presence of (short-run) economies of scale and would indicate that the bank is operating below its optimal scale level and thus can reduce its cost by expanding output. If the measure is greater than one then the bank experiences (short-run) diseconomies of scale and should reduce its output level to achieve optimal input usage. The reciprocal of scale economies of course is returns to scale (RTS). Technological change is derived from the first order derivative of the cost function with respect to time evaluated at output and input price levels, as well as the level of the quasi-fixed factor.

1.4.2. Data

The data used in this study are extracted from three main sources. The first source is the public-use quarterly Call Reports for all U.S. commercial banks that are collected and administrated by the Federal Reserve Bank of Chicago and the

FDIC. The majority of the data are from this source, which mainly consists of bank-specific variables. The second source is the FDIC website which provides information regarding failed banks and industry-level indicators. The third source is the website of the Federal Reserve Bank of St. Louis (FRED), which provides information on regional-specific macroeconomic variables.[23]

More than forty individual-specific financial ratios, state-specific macroeconomic, geographical, and structural variables are constructed from variables obtained form these sources as potential determinants of banking distress and failure. We apply the stepwise forward selection procedure (Klein and Moeschberger, 2003) to select the most relevant explanatory variables based on global and local tests, as well as the Akaike Information Criterion (AIC). Although these procedures are very useful for deciding how many and which variables to include in the model, we also select variables based on their contribution to the prediction accuracy of the model. The purpose of doing this is to identify some variables which introduce a spurious relationship to the model and hence destroy the predictive ability of the model. We prefer to exclude these variables from the model even though they are statistically significant and have the correct sign. In addition, other confounder variables, which are highly correlated with the significant variables, are excluded from the model due to high degrees of multicollinearity and attendant complications in obtaining the numerical solutions of our multivariate models.

[23]Websites for the data sources are: (i) Federal Reserve Bank of Chicago (http://www.chicagofed.org), (ii) FDIC (http://www.fdic.gov)(iii) Federal Reserve Bank of St. Louis (http://www.stlouisfed.org).

The final set of variables entering both the incidence and the latency part includes capital adequacy, asset quality, managerial quality, earnings, liquidity, and sensitivity (the so-called "CAMELS"), six structural and geographical variables, and four state-specific variables. We use the same set of explanatory variables in both the incidence and latency part of our model in order to capture the different effects that these have on the probability that a particular bank is troubled, as well as the probability and timing of the resolution of the bank's troubles by the FDIC. Tables 1.2 and 1.3 provide our mnemonics for the variable names as well as their formal definitions.

The first variable in table 1.2 is the tier 1 risk-based capital ratio. It is defined as the ratio of tier 1(core) capital to risk adjusted on and off balance sheet assets. Banks with a high level of this ratio are thought to have sufficient capital to absorb any losses occurring during the crisis and hence have a higher chance of survival. We expect a negative sign for this variable in both incidence and latency. Under current regulation, banks with a ratio above 4% are less likely to cause regulatory intervention. The next variable is the ratio of nonperforming loans to total loans, which consists of total loans and lease financing receivables that are nonacrual, past due 30-89 days and still accruing, and past due 90 days or more and still accruing. This variable is a primary indicator of the quality of loans made by banks and it is one of the influential factors explaining their distress and failure. The higher this ratio, the higher the probability that the bank will enter the watch list and subsequently fail. The next five ratios also reflect the asset quality of

banks. We expect the ratio of allowance for loan and lease loss to average total loans to have a positive effect on a bank's survival. Higher ratios may signal banks to anticipate difficulties in recovering losses and thus this variable may positively impact incidence. Similarly, charge-offs on loan and lease loss recoveries provide a signal of problematic assets that increase the probability of insolvency and failure. Provision for loan and lease losses are based upon the management's evaluation of loans and leases that the reporting bank has the intent to hold. Such a variable can expect to decrease the probability of distress and increase the probability of survival.

Two of the three management quality proxies that we include are constructed from the balance sheet items of the reporting banks. The ratio of the full-time employees to average assets has an ambiguous sign in both parts. However, we conjecture a negative sign on this variable as the FDIC may face constraints in seizing large banks with a large number of employees. The intermediation ratio shows the ability of a bank to successfully transform deposits into loans and thus we expect its impact to be negative. Earnings are also expected to have a negative effect on both parts. From the liquid assets we expect cash and core deposits to have negative signs, while the direction of the effect of Jumbo CD's is uncertain. Banks with more rate sensitive liabilities, repricing within a year more than assets, ex ante should be considered as riskier. The state-specific variables that we include in the model are expected to have a positive impact on survival with the exception

of the unemployment rate, which is expected to have a negative effect on their viability. The structural and geographical variables have ambiguous signs.

For the SF part of the model we use 14 quarters (2007.Q1-2010.Q2) of bank observations from the Call Reports. After deleting observations with obvious reporting errors and substantial outliers the final number of banks range from 6,342 in the first quarter of 2007 to 5,642 in the second quarter of 2010. The unbalanced panel has a total of 83,936 observations.

Means and standard deviations of financial ratios for the fourth quarter of 2007 and 2009 are reported in Table 1.4. The last column of this table reports the p-values of the hypothesis of no difference between the means of variables of failed and nonfailed banks based on the two-group mean comparison test for these two periods. This shows the difference in financial health of nonfailed and failed banks.[24] Table 1.5 reports the descriptive statistics of variables that enter the cost function for the sample of failed and nonfailed banks for the same selected periods. It is worth noting that, on average banks that failed had issued more real estate loans than their nonfailed peers prior to the crisis. This is consistent with the prevailing view that failed banks were highly engaged in residential mortgage loans, which experienced an unusually high default rates after the collapse of the housing bubble in 2006. Failed banks also paid higher salaries than did the nonfailed banks. For the DTMHM it is informative to look at the evolution of these variables

[24]Notice that there are troubled banks among the nonfailed banks that fail in subsequent periods. Hence, the difference must be larger from that reportered in the table.

over time. These are given in figure 1.4 where the large discrepancy between the financial health of failed and non-failed banks is evident for most of the financial ratios. For example, the key financial variables, such as capital adequacy and nonperforming loans, display significant mean differences which amplify as we move in time.[25]

1.5. Results and Predictive Accuracy

Table 1.6 reports the results for the continuous-time semiparametric and discrete-time MHM under the assumption that inefficiencies have no effect on the probability of incidence and latency. Both models produce qualitatively similar results. The influential factors that were a priori believed to have a strong effect on both probabilities turn out to have the correct sign and are statistically significant at any conventional significance level in both models. Results indicate that there is a large marginal effect of tier 1 capital ratio on the incidence probability. Other measures of earnings proxies and asset quality also have a large and significant effect on this probability. In other words, well capitalized banks with positive earnings and quality loans are less likely to appear on the FDIC watch list. In contrast, banks that are already on this list will increase their probability of failure in the industry if their capital ratio is insufficient, ratio of nonperforming loans is high and earnings are negative, and have a decreasing trend. Certificates of deposits

[25]We reject the null hypothesis of no mean difference for these two variables between failed and nonfailed banks at any conventional significance level based on the two-group mean comparison test with Satterthwaite's degrees of freedom for all periods.

and core deposits have the expected effect though not a statistically significant one. On the other hand, cash has a positive and significant effect. One explanation of this could be, after controlling for profitability, banks that remain cash idle have a higher opportunity cost. It would only stand to reason for these banks to be costly and inefficient. Banks with a large number of full-time employees have less chances to fail, as it is also seen by its negative sign in both parts. Those who successfully transform deposits into vehicles of investment are considered potentially stronger, while others with more rate sensitive liabilities appear to be less promising. The state-specific variables have the expected economic congruences which appear to be nonsignificant in the incidence part. We would expect these variables to significantly affect the probability of incidence of banks in states with higher unemployment rates, lower growth in personal income, limited construction permits, and falling housing prices, all of which would give cause for an on-site inspection. Only two of the four geographical variables have a significant effect. Banks that are FR members have a higher probability of failure than those that are not. This is associated with behavior consistent with moral hazard. Such banks have felt secure as members of Federal Reserve system and hence may have assumed higher risks than they would have had they not had FR banking. The positive result of the FR district code indicates that the probability of insolvency and failure is higher for banks in the Atlanta (6) district than for banks in the Boston (1) district and it is lower than for banks in the San Francisco (12) district. Recall that Washington, D.C. is the reference district. The size of the bank, as it

is measured by the natural logarithm of its gross total assets, has a negative and significant sign only in the incidence part of model II, which implies that larger banks are less likely to find themselves on the watch list and then subsequently fail. Older and well-established banks have lower failure probabilities than their younger counterparts.

Table 1.7 contains the results for the continuous-time semiparametric and discrete-time MHM with the stochastic frontier specification. With few exemptions, the results are qualitatively similar to those reported in Table 1.6. Inefficiency has a positive effect on incidence and failure probability. The effect is only significant on the latter probability and this is consistent with the view that bank performance is not the criterion for on-site examination but rather a factor affecting a bank's longer term viability. The distributional parameters are significant at the 1% significant level. The descriptive statistics for the efficiency scores obtained from models III and IV, as well as from the standard time-invariant random effects (RE) model for the sample of nonfailed and failed banks are summarized in Tables 1.8 and 1.9, respectively. There is a small but statistically significant difference between average efficiencies of failed and nonfailed banks in models III and IV. This difference is not statistically significant for efficiencies derived from the random effects model. figure 1.5 plots the distribution of inefficiencies (non-truncated) obtained from these three models. It is interesting to note that the RE model reports some surviving banks as extremely inefficient while the most efficient banks are banks that failed. Hence, we suspect that the two-step approach

would yield the opposite sign on inefficiency component from what we would expect. The difference in average efficiencies from the single step estimation can be mainly attributed to the fact that distressed banks that subsequently fail typically devote their efforts to overcome the difficulties and clean up their balance sheets. These impose additional costs on banks and worsen an already bad situation.

In figure 1.6 we plot the estimated average returns to scale and the technical change for failed and nonfailed banks for the full sample period. Both type of banks display increasing returns to scale. Nonfailed banks appear to be improving their scale efficiencies after the second quarter of 2009. These banks also display a negative technological change (digress) for all periods as estimated by both specifications. The failed banks had experienced a slight technological progress during the last periods. This points out to the fact that banks that failed in 2009 were relatively younger banks that employed newer technology than did older banks.

In figures 1.7 and 1.8 we plot the survival profile of the average bank that failed during 2008-2009 period for all four models. The average survival profile based on results of the SPMHM is constructed from (1.11) and is based on the average characteristics of failed banks. Similarly, the average profile based on the DTMHM results is calculated from (1.21), which is a step function with steps occurring at discrete failure times. From figure 1.7 it can be seen that average failed banks in SPMHM are predicted to have a duration time of 22 months. After controlling for inefficiencies the time to failure drops to 21 months. Based on the DTMHM results, figure 1.8 demonstrates that a bank with the same characteristics

as the representative failed bank will survive up to 7 quarters after accounting for inefficiency.

It is also interesting to look at the survival profile of the most and the least efficient banks derived from models III and IV. figure 1.9 displays the survival profiles obtained from SPMHM. The least efficient bank with the efficiency score of 0.149 is predicted to fail in 8 months. This bank was closed by FDIC in the end of August of 2008. On the other hand, the most efficient bank with the efficiency score of 0.971 has a survival probability of one throughout the sample period. This is also illustrated in figure 1.10, where the least efficient bank with the efficiency score of 0.154 is predicted to fail by the fifth quarter, using the DTMHM results. This bank failed in the third week of April of 2009.[26] The most efficient bank with the efficiency score of 0.969 has an estimated survival probability that exceeds 0.95.

We next examine our results by recasting our model estimates as early warning tools that can correctly classify failed and nonfailed banks within our sample used for estimation as well as in our hold-out sample. The tests are based on two types of errors, similar to those that arise in any statistical hypothesis testing. These are type I and type II errors (see Lane et al. 1986; Whalen, 1991; and Thompson, 1992 among others). A type I error is defined as the error due to classifying a failed bank as a nonfailed bank, while a type II error arises from classifying a non-failed bank as a failed bank. There is a trade-off between these two type of errors and

[26]The identity of the least efficient bank is not the same in these two models. However, the identity of the most efficient bank is the same.

both are important from a public policy standpoint. Models with low type I error are more desirable since timely identification of failed banks allows the regulator to undertake any prompt corrective action to ensure the stability and the soundness of the financial system. On the other hand, models with high type II error will be unnecessary flagging some banks as failures while they are not, and hence could waste regulators' time and resources. However, it is oftentimes hard to interpret the costs of a type II error since various constraints faced by FDIC could delay the resolution of an insolvent bank. Thompson (1992) attributes this to information, administrative, legal and political constraints, among others. Whalen (1991) notes that some type II error predictions actually represent failures that occur in the near future and so should be considered as a success of the model rather than its failure.

Table 1.10 reports the in-sample predictive accuracy for the four models based on type I, type II, and overall classification error. Overall classification error is a weighted sum of type I and type II errors. In what follows we set the weights at 0.5 for both errors. Clearly this weighting scheme is arbitrary and alternative weighting schemes could be based on different risk preference assumptions, implicit and explicit costs of regulation, etc. In our predictive accuracy analysis each bank is characterized as a failure if its survival probability falls bellow a probability cutoff point, which we base on the sample average ratio of failed to nonfailed banks (0.021). The DTMHM specification yields a lower type I error than does the SPMHM. This is to be expected since the DTMHM incorporates multiperiod

observations for each bank and thus is more informative on bank financial health than the single-period cross-sectional observations. There is a significant drop in type I error in both specifications when the performance of a bank is added to the model as an additional factor. On the other hand, type II error is increased in the DTMHM and it is doubled when inefficiency is included. Based on the overall classification error, Model IV seems to perform slightly better than Model III, but it largely outperforms the Models I and II.

Table 1.11 presents the errors that judge the out-of sample classification accuracy of the models. The SPMHM errors are based on the survival profile of banks using the 2009 end-year data and can predict failures that may occur through the end of 2011. We consider banks that have failure times of up to nine months in order to compare the errors from this model with those based on the discrete-time alternative. The survival probabilities in the DTMHM are calculated from (1.21). In order to account for the third quarter failures we keep the financial ratios the same from the second quarter to the third quarter since the Call Reports of 2010 have not yet been released as of the date of this analysis. We first compare these results with the in-sample classifications. There is a significant drop in type I error for all four models. This is mainly due to the fact that the data used to calculate the survival profiles of each banks are more informative than what was used to estimate the model parameters and is reasonable since the end of 2009 is considered the peek year of the banking crisis during which the financial health of some banks deteriorated significantly (see Table 3.9). The inter-model comparison is the

same as above with Model IV favored over the other models based on predictive accuracy.

1.6. Conclusions

Massive banking failures during the financial turmoil of the last three years has resulted in enormous financial losses and costs to the U.S. economy, not only in terms of the bailouts by regulatory authorities in their attempt to restore liquidity and stabilize the financial sector, but also in terms of the lost jobs in banking and other sectors of economy, failed businesses, and ultimately slow growth of the economy as a whole. The design of early warning models that accurately predict the failures and their timing is of crucial importance in order to ensure the safety and the soundness of the financial system. Early warning models that can be used as off-site examination tools are useful for at least three reasons. They can help direct and efficiently allocate the limited resources and time of on-site examination so that banks in immediate help are examined first. They are less costly than on-site visits made by supervisors to institutions considered at risk and can be performed with high frequency to examine the financial condition of the same bank. Finally, they can predict failures at a reasonable length of time prior to the marked deterioration of bank's condition and allow supervisors to undertake any prompt corrective action that will have the minimal cost to a taxpayer.

In this chapter we considered early warning models that attempt to explain the recent failures in the U.S. commercial banking sector. We employed a duration analysis model combined with a static logit model to determine troubled banks which subsequently fail or survive. Both, continuous and discrete time version of the mixed model were specified and estimated. These effectively translated the bank-specific characteristics, state-related macroeconomic variables, and geographical and structural variables into the risk measures. Capital adequacy and nonperforming loans were found to play a pivotal role in determining and closing insolvent institutions. State-specific variables appeared to significantly affect the probability of failure but not insolvency. The discrete-time model outperformed the continuous-time model as it is able to incorporate time-varying covariates, which contain more and reacher information. We also found that managerial efficiency does not significantly affect the probability of a bank being troubled but plays an important role in their longer term survival. Inclusion of the efficiency measure led to improved prediction in both models.

1.7. Appendix A: Derivation of the likelihood function

In this appendix we show the derivation of the sample likelihood function given in expression (1.9). For this purpose, we first note that at time t each bank can fall into four mutually exclusive states of nature:

$$States = \begin{cases} h_i = 1, d_i = 1 & \text{with prob. } p\lambda_i^p(t;w_i)S_i^p(t;w_i) \\ h_i = 0, d_i = 1 & \text{with prob. } (1-p)\,\lambda_i^s(t;w_i)S_i^s(t;w_i) \\ h_i = 1, d_i = 0 & \text{with prob. } pS_i^p(t;w_i) \\ h_i = 0, d_i = 0 & \text{with prob. } (1-p)\,S_i^s(t;w_i) \end{cases}$$

Then

$$\begin{aligned} L(\theta;x,w,d) &= \prod_{i=1}^n L_i(\theta;x,w,d) \\ &= \prod_{i=1}^n \left\{ \left(p\lambda_i^p(t;w_i)S_i^p(t;w_i)\right)^{h_i} \left((1-p)\,\lambda_i^s(t;w_i)S_i^s(t;w_i)\right)^{1-h_i} \right\}^{d_i} \\ &\quad \times \left\{ \left(pS_i^p(t;w_i)\right)^{h_i} \left((1-p)\,S_i^s(t;w_i)\right)^{1-h_i} \right\}^{1-d_i} \\ &= \prod_{i=1}^n p^{h_i}(1-p)^{(1-h_i)} \left[\lambda_i^p(t;w_i)\right]^{d_i h} \\ &\quad \times \left[\lambda_i^s(t;w_i)\right]^{d_i(1-h_i)} \left[S_i^p(t;w_i)\right]^{h_i} \left[S_i^s(t;w_i)\right]^{1-h_i} \end{aligned}$$

By assumption, $\lambda_i^s(t;w_i) = 0$ if and only if $h_i = 0$ and $d_i = 0$ i.e., a bank is healthy and is not observed failing. Similarly $S_i^s(t;w_i) = 1$ if and only if $h_i = 0$ i.e., a bank is healthy. The final sample likelihood function is then given by

$$L(\theta; x, w, d) = \prod_{i=1}^{n} p^{h_i} (1-p)^{(1-h_i)} \left[\lambda_i^p(t; w_i)\right]^{d_i h_i} \left[S_i^p(t; w_i)\right]^{h_i}$$

which implies that the completely healthy banks contribute to the likelihood function only through their probability being troubled.

1.8. Appendix B: Tables and Figures

Table 1.2. CAMELS proxy Financial Ratios

	Capital Adequacy (C)
tier1	Tier 1 (core) capital/risk-weighted assets
	Assset Quality (A)
rnpl	Nonperforming loans/total loans
alll	Allowance for loan and lease loss/average loans and leases
reln	Commercial real estate loans/total loans
coffs	Charge-off on loans and leases/average loans and leases
lrec	Recoveries on loan and lease losses/loans and leases
llp	Provision for loan and lease losses /loans and leases
	Managerial Quality (M)
fte	Number of fulltime equivalent employees/average assets
imr	Total loans/total deposits
u	Random Effects inefficiency score
	Earnings (E)
oi	Total operating income/average assets
roa	Net income (loss)/average assets
roe	Net income (loss)/total equity
	Liquidity (L)
cash	Noninterest-bearing balances, currency, and coin/average assets
cd	Total time deposits of USD 100,000 or more/total assets
coredep	Core deposits/total assets
	Sensitivity (S)
sens	1-Year rate sensitive assets minus liabilities/total assets

Table 1.3. Structural, Geographical, and State-Specific Macroeconomic variables

Structural and geographical variables	
chtype	Charter type (1if state chartered, 0 otherwise)
frsmb	FRS membership indicator (1 if Federal Reserve member, 0 otherwise)
ibf	International banking facility (1 if bank operates an ibf, 0 otherwise)
frsdistrcode	FRS district code (Boston(1), New York (2), Philadelphia (3), Cleveland (4), Richmond (5), Atlanta (6), Chicago (7), St. Louis (8), Minneapolis (9), Kansas City (10), Dallas (11), San Francisco (12), Washington., D.C. (0-referense district))
lgta	log of total assets
age	Age (measured in quarters)
State-Specific Macroeconomic variables	
ur	Unemployment rate
chpi	% Change in personal income
chphi	% Change in house price index
chnphu	Change in new private housing units authorized by building permits

Table 1.4. Descriptive Statistics for CAMELS proxy financial ratios for the fourth quarter of 2007 and 2009

	Non-Failed Banks				Failed Banks				p-value
Variable	2007.Q4		2009.Q4		2007.Q4		2009.Q4		2007.Q4/2009.Q4
	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	
tier1	0.1072	0.0333	0.1025	0.0301	0.1011	0.0384	0.0253	0.0219	0.049/0.000
alll	0.0129	0.0065	0.0169	0.0094	0.0161	0.0100	0.0450	0.0221	0.000/0.000
reln	0.4583	0.1703	0.4590	0.1652	0.6254	0.1440	0.6070	0.1372	0.000/0.000
rnpl	0.0260	0.0231	0.0435	0.0395	0.0539	0.0528	0.2114	0.1043	0.000/0.000
roa	0.0097	0.0076	0.0023	0.0158	0.0030	0.0137	-0.0636	0.0296	0.000/0.000
roe	0.0957	0.0771	0.0064	0.3224	0.0178	0.1980	-1.5493	55.774	0.000/0.801
cd	0.1563	00749	0.1644	0.0757	0.2082	0.1010	0.2302	0.1092	0.000/0.000
coredep	0.8228	0.0741	0.8334	0.0659	0.8068	0.0798	0.9054	0.0665	0.013/0.000
coffs	0.2493	0.3995	0.5687	0.6542	0.2902	0.3376	1.5778	1.1923	0.135/0.000
lrec	0.0008	0.0033	0.0008	0.0021	0.0004	0.0012	0.0014	0.0017	0.001/0.011
llp	0.0027	0.0067	0.0120	0.0169	0.0092	0.0119	0.0724	0.0370	0.000/0.000
fte	0.3169	0.1290	0.2838	0.1174	0.2737	0.1999	0.2134	0.0796	0.007/0.000
imr	0.8126	0.2053	0.7792	0.1873	0.9647	0.1526	0.8149	0.1271	0.000/0.014
sens	-0.1433	0.1490	-0.1348	0.1270	-0.0505	0.1748	-0.2095	0.1440	0.000/0.022
cash	0.0323	0.0182	0.0310	0.0269	0.0209	0.0135	0.0230	0.0307	0.000/0.000
oi	0.0718	0.0121	0.0581	0.0115	0.0786	0.0178	0.0511	0.0111	0.000/0.000
N	5843		5674		125		92		

The calculation of the mean and standard deviation for 2007.Q4 sample of failed banks is based on banks that failed between 2007.Q4-2009.Q4. Figures for 2009.Q4 are based on failures of 2010.Q1-2010.Q3

Pvalues under the null hypothesis of no difference between the means of variables of failed and nonfailed banks based on the twogroup mean comparison test with Satterthwaite's degrees of freedom.

Table 1.5. Descriptive Statistics for variables that enter the cost function for the fourth quarter of 2007 and 2009

	Non-Failed Banks				Failed Banks				p-value
Variable	2007.Q4		2009.Q4		2007.Q4		2009.Q4		2007.Q4/2009.Q4
	Mean	S.D	Mean	S.D	Mean	S.D	Mean	S.D	
Cost	53395	955229	46259	780129	42967	105935	24717	31302	0.488/0.049
yreln	485820	7426706	625015	9671132	532470	1432030	329015	436220	0.745/0.003
yciln	165465	3061182	179833	3128254	67053	181888	44812	65765	0.021/0.001
yinln	78985	1587319	110304	2343919	15900	68652	6952	17501	0.003/0.001
ysec	195045	3420833	335307	6913427	118358	379112	56332	108157	0.155/0.003
yobs	136245	2736033	212630	3947440	74323	290364	104491	215157	0.146/0.060
X	163124	3369938	250018	5307998	58179	124646	58179	65623	0.694/0.718
wdep	0.0297	0.0066	0.0165	0.0052	0.0367	0.0065	0.0248	0.0054	0.001/0.000
wlab	55.078	14.016	58.598	14.645	66.127	16.279	69.801	17.747	0.000/0.000
wcap	0.2875	0.2119	0.2900	0.2131	03041	0.2507	0.3551	0.2949	0.409/0.050
wpurf	0.0447	0.0083	0.0267	0.0076	0.0458	0.0091	0.0311	0.0096	0.132/0.000
N	5843		5674		125		92		

The calculation of the mean and standard deviation for 2007.Q4 sample of failed banks is based on banks that failed between 2007.Q4-2009.Q4.

Figures for 2009.Q4 are based on failures of 2010.Q1-2010.Q3. The price of labor is measured in thousands of US dollars. Pvalues under the null hypothesis of no difference between the means of variables of failed and nonfailed banks based on the twogroup mean comparison test with Satterthwaite's degrees of freedom.

Table 1.6. Estimates from the Continuous-Time Semiparametric Proportional Mixture Hazard Model (Model I) and Discrete-Time Mixture Hazard Model (Model II)

	Model I		Model II	
Variable	Latency	Incidence	Latency	Incidence
Intercept		-2.5989		4.9130
lgta	0.0797	0.0607	0.0531	-0.3320***
age	-0.0004*	0.0004	-0.0003	0.0001
tier 1	-48.417***	-86.791***	-47.060***	-88.728***
alll	-9.5829**	16.473**	-8.8615*	8.8671
reln	4.4321***	2.0116	3.7811***	3.9762***
rnpl	7.2555***	6.3838***	6.1802***	9.6447***
roa	-6.1672	-11.248**	-7.2727	-8.8145
roe	0.0003	0.0002	0.0003	0.0003
cd	1.0098	1.6651	1.2499	0.8245
coredep	-2.7654	-1.2140	-2.5466	-3.1272
coffs	0.2351***	0.3168***	0.2319***	0.2703**
lrec	38.162**	14.463	35.681*	37.945
llp	-10.427**	-15.501**	-11.688**	-13.155**
fte	-0.8228	-3.0004**	-0.8329	-3.1287**
imr	-4.2141***	-1.7238	-3.7792***	-4.4020***
sens	2.3255***	2.5869**	2.0025**	5.6444***
cash	6.7983***	6.7497**	6.9211***	4.5465
oi	-3.9670	-6.1756	-3.1670	-4.9651
ur	0.1198***	0.0196	0.0655*	0.0548
chpi	-15.091*	-10.555	-20.313**	-10.645
chhpi	-8.1375*	-3.1678	-9.8817**	-5.4824
chnphu	-0.6570***	0.0006	-0.5246**	0.0047
chtype	-0.2151	0.4441	0.0223	-0.7143
frsmb	0.4707***	0.4018*	0.4617***	0.3466
ibf	1.1171	1.4405	1.2816*	-2.5959***
frsdistrcode	0.2465***	0.2615***	0.2295***	0.2457***
LogL	–1763.87		-1714.92	
N	5968		38571	

p*<0.1, p**<0.05, p***<0.01

Table 1.7. Estimates from the Stochastic Frontier Continuous-Time Semiparametric Proportional Mixture Hazard Model (Model III) and Stochastic Frontier Discrete-Time Mixture Hazard Model (Model IV)

	Model III		Model IV	
Variable	Latency	Incidence	Latency	Incidence
Intercept		-2.6934		4.7694*
lgta	-0.0408	-0.0087	-0.0742	-0.3466***
age	-0.0004*	0.0004	-0.0004*	0.0001
tier 1	-48.647***	-86.280***	-48.452***	-88.684***
alll	-8.5073*	17.003**	-8.8587*	8.8881
reln	4.6588***	2.1044*	4.4871***	3.9288***
rnpl	6.9014***	6.0653***	6.7347***	9.5835***
roa	-6.1672	-11.451***	-6.4175	-8.8129
roe	0.0002	0.0001	0.0002	0.0002
cd	0.8641	1.5840	0.7565	0.7329
coredep	-2.3913	-1.0244	-1.5432	-2.9196
coffs	0.2447***	0.3232***	0.2516***	0.2720**
lrec	37.309**	14.661	37.219**	38.569
llp	-11.175**	-15.784**	-11.654**	-13.211**
fte	-2.1781**	-3.8780**	-2.8670***	-3.3559**
imr	-3.7660***	-1.4553	-3.2640***	-4.2466***
sens	2.2143***	2.5264**	2.0894**	5.6036***
cash	7.4166***	7.1461**	7.6605***	4.6012
oi	-3.9483	-6.4722	-4.1980	-5.0126
ur	0.1210***	0.0234	0.1208***	0.0555
chpi	-15.567**	-9.7061	-15.551*	-10.639
chhpi	-8.1886*	-3.2387	-8.1802**	-5.4864
chnphu	-0.6300***	-0.0001	-0.6171***	0.0046
chtype	-0.1496	0.4875	-0.1293	-0.7224
frsmb	0.4960**	0.3994*	0.4977***	0.3487
ibf	1.1325	1.4718*	1.1295*	-2.5923***
frsdistrcode	0.2612***	0.2725***	0.2663***	0.2469***

p*<0.1, p**<0.05, p***<0.01

table 1.7 Cont'd

	Model III		Model IV	
Variable	Latency	Incidence	Latency	Incidence
SFM				
δ_1		0.2062		0.0343
δ_2	0.3058***		0.4137***	
σ	0.0552***		0.0548***	
γ	0.5173***		0.5278***	
LogL	-67701		-66360	
N	5968		38571	

p*<0.1, p**<0.05, p***<0.01

Table 1.8. Cost efficiencies results for the sample of Nonfailed Banks

	Mean	Standard Deviation	Minimum	Maximum
Model III	0.6817	0.0691	0.3167	0.9705
Model IV	0.7295	0.1630	0.1992	0.9688
Random Effects	0.6466	0.0662	0.4636	0.9650

The top and bottom 5% of inefficiencies scores are trimmed

Table 1.9. Cost efficiencies results for the sample of Failed Banks

	Mean	Standard Deviation	Minimum	Maximum
Model III	0.6721	0.1022	0.1499	0.8722
Model IV	0.6804	0.0824	0.1539	0.8488
Random Effects	0.6408	0.0798	0.3845	0.8626

The top and bottom 5% of inefficiencies scores are trimmed

Table 1.10. In-sample classification error decomposition

	I	II	III	IV
Type I error	0.3840	0.2882	0.1123	0.0644
Type II error	0.0047	0.0051	0.0231	0.0476
Overall classification error	0.1937	0.1465	0.0581	0.0573

Overall classification error is a simple average of type I and type II errors

Table 1.11. Out-of-sample classification error decomposition

	I	II	III	IV
Type I error	0.2283	0.1630	0.0543	0.0217
Type II error	0.0049	0.0062	0.0244	0.0503
Overall classification error	0.1157	0.0840	0.0394	0.0360

Overall classification error is a simple average of type I and type II errors

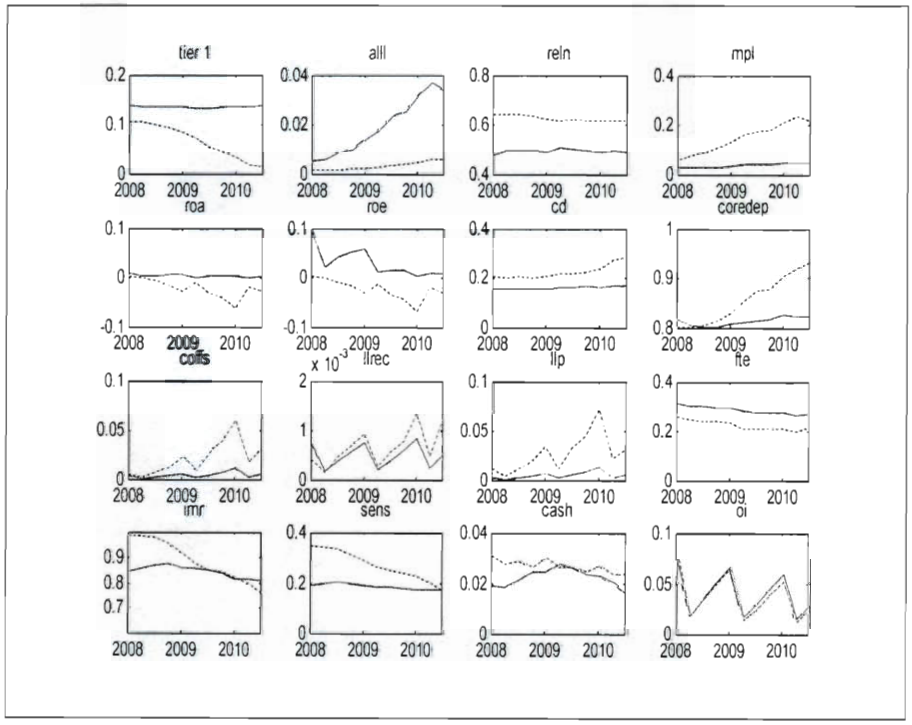


Figure 1.4. Financial ratios over the 2007.Q4-2010.Q2 period. Solid line is for non-failed banks and dashed line is for failed banks.

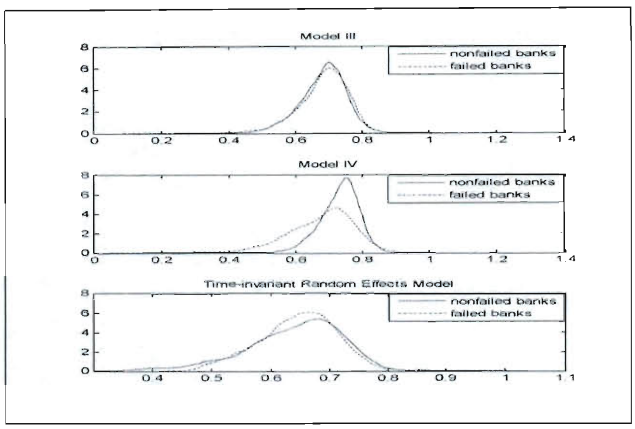


Figure 1.5. Distribution of estimated cost efficiencies obtained from Models III, IV and the Random Effects Model.

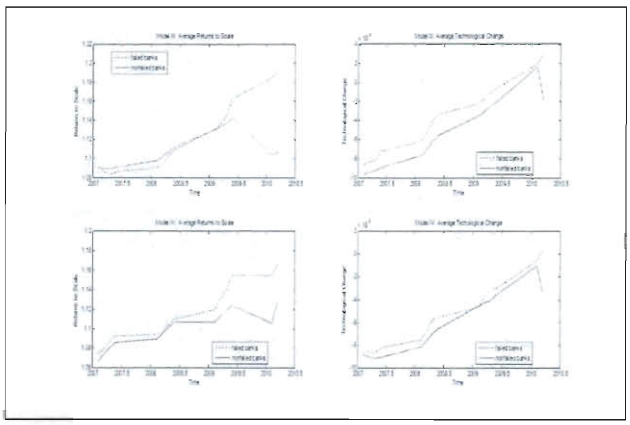


Figure 1.6. Estimated Average Returns to Scale and Technological Change from Models III and IV

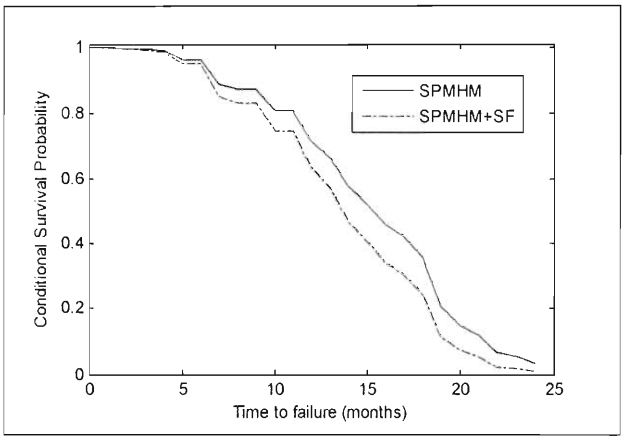


Figure 1.7. SPMHM - Survival profile of the average failed bank (2008-2009)

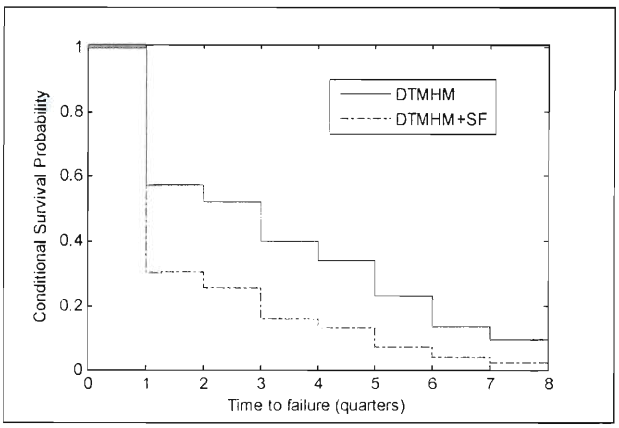


Figure 1.8. DTMHM - Survival pofile of the average failed bank (2008-2009)

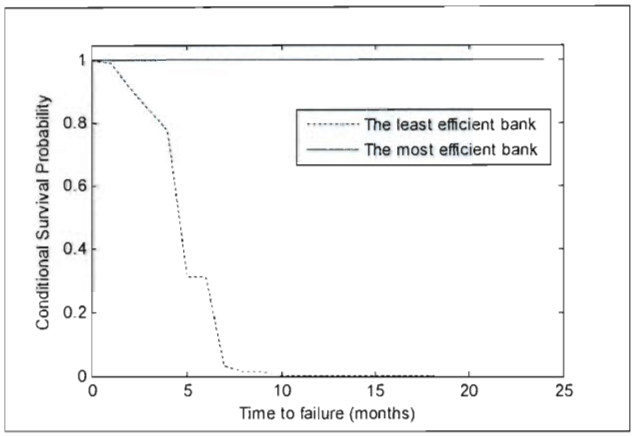


Figure 1.9. Model III - Survival profile of the most and the least efficient bank

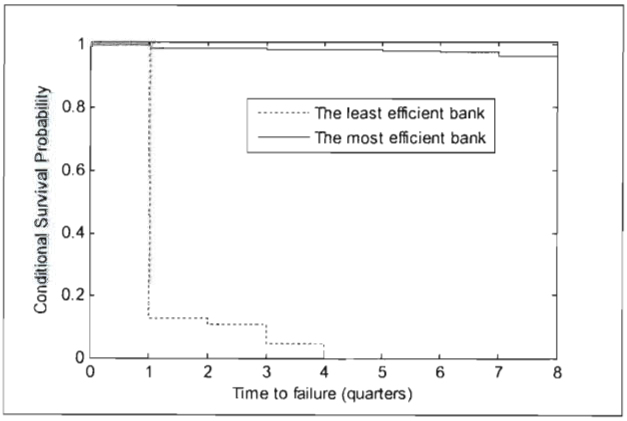


Figure 1.10. Model IV - Survival profile of the most and the least efficient bank

CHAPTER 2

Bounded Stochastic Frontiers with an Application to the US Banking Industry: 1984-2009[1]

2.1. Introduction

The parametric approach to estimate stochastic production frontiers was introduced by Aigner, Lovell, and Schmidt (1977), Meeusen and van den Broeck (1977), and Battese and Corra (1977). These approaches specified a parametric production function and a two-component error term. One component, reflecting the influence of many unaccountable factors on production as well as measurement error, is considered "noise"and is usually assumed to be normal. The other component describes inefficiency and is assumed to have a one-sided distribution, of which the conventional candidates include the half normal (Aigner, et al., 1977), truncated normal (Stevenson, 1980), exponential (Meeusen and van den Broeck, 1977) and gamma (Greene 1980a,b, Stevenson, 1980). This stochastic frontier production function has become an iconic modeling paradigm in econometric research,

[1]This is a version of my work with professors Junhui Qian (Shanghai Jiao Tong University) and Robin Sickles.

rate making decisions in regulated industries across the world, in evaluating outcomes of market reforms in transition economies, and in establishing performance benchmarks for local, state, and federal governmental activities.

In this chapter we propose a new class of parametric stochastic frontier models with a more flexible specification of the inefficiency term, which we view as improvement on the basic iconic stochastic frontier production model. Instead of allowing unbounded support for the distribution of productive (cost) inefficiency term in the right (left) tail, we introduce an unobservable upper bound to inefficiencies or a lower bound to the efficiencies, which we call the inefficiency bound. The introduction of the inefficiency bound makes the parametric stochastic frontier model more appealing for empirical studies in at least two aspects. First, it is plausible to allow only bounded support in many applications of stochastic frontier models wherein the extremely inefficient firms in a competitive industry or market are eliminated by competition. Bounded inefficiency makes sense in this setting since the extremely inefficient stores will be forced to close and thus individual production units constitute a truncated sample.[2] This is consistent with the arguments of Alchian (1950) and Stigler (1958) wherein firms are at any point in time not in a static long run equilibrium, but rather are tending to that situation as they are buffeted by demand and cost shocks. As a consequence, even if we correctly specify a family of distributions for the inefficiency term, the stochastic

[2]In addition, the frequent use of balanced panels in empirical studies would in effect eliminate those failing firms from the sample and thus would provide more merit to the bounded inefficiency model.

frontier model may still be misspecified. This particular setting is one in which the inefficiency bound is informative as an indicator of competitive pressures and/or the extent of supervisory oversight by direct management or by corporate boards. In settings in which firms can successfully differentiate their product, which is the typical market structure and not the exception, or where there are market concentrations that may reflect collusive behavior or conditions for a natural monopoly and regulatory oversight, incentives to fully exploit market power or to instead make satisficing decision are both possible outcomes. Much more likely is that it is not one or the other but some middle ground between the two extremes that would be found empirically.[3]

A second justification for our introduction of the inefficiency bound into the classical stochastic production frontier model is that our model points to an explanation for the finding of "wrong "skewness in many applied studies using the traditional stochastic frontier, and thus to the potential of our bounded inefficiency model to solve the "incorrect" skewness problem. Researchers have often found positive instead of negative skewness in many samples examined in applied work, which may point to the stochastic frontier being incorrectly specified. However,

[3] "The quiet life hypothesis" (QLH) by Hicks (1935) argues that, due to management's subjective cost of reaching the optimal profits, firms use their market power to allow inefficient allocation of resources. Increasing competitive pressure is likely to force management to work harder to reach optimal profits. Another hypothesis that relates market power and efficiency is "the efficient structure hypothesis" (ESH) by Demsetz (1973). ESH argues that firms with superior efficiencies or technologies have lower costs and therefore higher profits. These firms are assumed to gain larger market shares which lead to higher concentration. Recently Kutlu and Sickles (2010) have constructed a model in which the dynamic game is played out and have tested for the alternative outcomes, finding support for the QLH in certain airlines city-pair markets and the ESH in others.

we conjecture that the distribution of the inefficiency term may itself be negatively skewed, which may happen if there is an additional truncation on the right tail of the distribution. One such specification in which this is a natural consequence is when the distribution of the inefficiency term is doubly truncated normal, that is, a normal distribution truncated at a point on the right tail as well as at zero. As normal distributions are symmetric, the doubly truncated normal distribution may exhibit negative skewness if the truncation on the right is closer to the mode than that on the left. We also consider the truncated half normal distribution, which is a special case of the former, and the truncated exponential distribution. Although these two distributions are always positively skewed, the fact that there is a truncation on the right tail makes the skewness very hard to identify empirically. That is to say, when the true distribution of the one-sided inefficiency error is bounded (truncated), the extent to which skewness is present in any finite sample may be substantially reduced, often to the extent that negative sample skewness for the composite error is not statistically significant. Thus the finding of positive skewness may speak to the weak identifiability of skewness properties in a bounded frontier model.

In addition to proposing new parametric forms for the classical stochastic production frontier model, we also show that our models are identifiable, and in which cases the identification is local or global. Initial consistent estimates are based on method of moments estimates, based on explicit analytic expressions which we derive, and which either can be used in a two-step method of scoring or as starting

values in solving the normal equations for the relevant sample likelihood, based on the parametric density functions whose expressions we also provide. As the regularity conditions for the maximum likelihood estimation are satisfied, we employ it in order to obtain consistent and asymptotically efficient estimates of the model parameters, including this of the inefficiency bound. We conduct Monte Carlo experiments to study the finite sample behavior of our estimators. We also extend the model to the panel data setting and allow for a time-varying inefficiency bound. By allowing the inefficiency bound to be time-varying, we contribute another time-varying technical efficiency model to the efficiency literature. Our model differs from those most commonly used in the literature, e.g., Cornwell, Schmidt, and Sickles (1990), Kumbhakar (1990), Battese and Coelli (1992), and Lee and Schmidt (1993) in that, while previous time-varying efficiency models are time-varying in the mean or intercept of individual effects, our model is time-varying in the lower support of the distribution of individual effects. This explicitly allows for the level of competitive pressures on firms and other factors that may force the firms to exit the industry to change over time as demand and cost conditions change.

The outline of this chapter is as follows. In Section 2 we present the new models and derive analytic formula for density functions and expressions that allow us to evaluate inefficiencies. Section 3 deals with the "wrong" skewness issue inherent in the traditional stochastic frontier model. Section 4 discusses the identification of the new models and the methods of estimation. Section 5 presents Monte Carlo

results on the finite sample performance of the bounded inefficiency model vis-a-vis classical stochastic frontier estimators. The extension of the new models to panel data settings and specification of the time-varying bound is presented in section 6. In Section 7 we give an illustrative study of the efficiency of U.S. banking industry in 1984-2009. Section 8 concludes.

2.2. The Model

We consider the following Cobb-Douglas production model,

$$y_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_{i,k} + \varepsilon_i \tag{2.1}$$

where

$$\varepsilon_i = v_i - u_i. \tag{2.2}$$

For every production unit i, y_i is the log output, x_{ik} the k-th log input, v_i the noise component, and u_i the (nonnegative) inefficiency component. We maintain the usual assumption that v_i is $iid\ N(0, \sigma_v^2)$, u_i is iid, and v_i and u_i are independent from each other and from regressors. Clearly we can consider other more flexible functional forms for production (or cost) that are linear or linear in logarithms, such as the generalized Leontief or the transcendental logarithmic, or ones that are nonlinear. The only necessary assumption is that the error process ε_i is additively separable from the functional forms we employ in the stochastic production (cost) frontier.

As described in the introduction, our model differs from the traditional stochastic frontier model in that u_i is of bounded support. Additional to the lower bound, which is zero and which is the frontier, we specify an upper bound to the distribution of u_i (lower bound in the case of the cost frontier $\varepsilon_i = v_i + u_i$). In particular, we assume that u_i is distributed as doubly truncated normal, the density of which is given by

$$f(u) = \frac{\frac{1}{\sigma_u}\phi(\frac{u-\mu}{\sigma_u})}{\Phi(\frac{B-\mu}{\sigma_u}) - \Phi(\frac{-\mu}{\sigma_u})}\mathbf{1}_{[0,B]}(u),\ \sigma_u > 0, B > 0,$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf of the standard normal distribution, respectively, and $1_{[0,B]}$ is an indicator function. It is a distribution obtained by truncating $N(\mu, \sigma_u^2)$ at zero and $B > 0$. The parameter B is the upper bound of the distribution of u_i and we may call it the inefficiency bound. The inefficiency bound may be a useful index of competitiveness of a market or an industry.[4] In the banking industry, which we examine in section 7, the inefficiency bound may also represent factors that influence the financial health of the industry. It may be natural to extend this specification and treat the bound as a function of individual specific covariates z_i, such as $\exp(\delta' z_i)$, which would allow identification of bank-specific measures of financial health.

[4]The inefficiency bound has a natural role in gauging the tolerance for or ruthlessness against inefficient firms. It is also worth mentioning that, using this bound as the "inefficient frontier," we may define "inverted" efficiency scores in the same spirit of "Inverted DEA" described in Entani, Maeda, and Tanaka (2002).

Using the usual nomenclature of stochastic frontier models, we may call the model described above the normal/doubly truncated normal model, or simply, the doubly truncated normal model. The doubly truncated normal model is rather flexible. It nests the truncated normal ($B = \infty$), half normal ($\mu = 0$ and $B = \infty$), and truncated half normal models ($\mu = 0$). One desirable feature of our model is that the doubly truncated normal distribution may be positively or negatively skewed, depending on the truncation parameter B. This feature provides us with an alternative explanation for the "wrong skewness" problem prevalent in empirical stochastic frontier studies. This will be made more clear later in this section and in the following chapter. Another desirable feature of our model is that, like the truncated normal model, it can describe the scenario that only a few firms in the sector are efficient, a phenomenon that is described in the business press as "few stars, most dogs", while in the truncated half normal model and the truncated exponential model (in which the distribution of u_i is truncated exponential), most firms are implicitly assumed to be relatively efficient.

In Table 2.1 we provide detailed properties of our model. In particular, we present the density functions for the error term ε_i, which is necessary for maximum likelihood estimation, and the analytic form for $E[u_i|\varepsilon_i]$, which is the best predictor of the inefficiency term u_i under our assumptions, and the conditional distribution of u_i given ε_i, which is useful for making inferences on u_i. The results for the truncated half normal model, a special case of the doubly truncated normal model ($\mu = 0$), are also presented. Finally, we also provide results for the truncated

exponential model, in which the inefficiency term u_i is distributed according to the following density function,

$$f(u) = \frac{1}{\sigma_u(1 - e^{-B/\sigma_u})} e^{-\frac{u}{\sigma_u}} \mathbf{1}_{[0,B]}(u) \tag{2.3}$$

The truncated exponential distribution can be further generalized to the truncated gamma distribution, which shares the nice property with the doubly truncated normal distribution that it may be positively or negatively skewed.

For the doubly truncated normal model and the truncated half normal model, the analytic forms of our results use the so-called $\gamma-$parameterization, which specifies

$$\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}, \quad \gamma = \sigma_u^2/\sigma^2 \tag{2.4}$$

By definition $\gamma \in [0, 1]$, a compact support, which is desirable for the numerical procedure of the maximum likelihood estimation. Another parameterization, initially employed by Aigner et al. (1977), is the $\lambda-$parameterization

$$\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}, \quad \lambda = \sigma_u/\sigma_v \tag{2.5}$$

We may check that when $B \to \infty$, the density function for ε_i in the doubly truncated normal model reduces to that of the truncated normal model introduced by Stevenson (1980). Furthermore, if $\mu = 0$, it reduces to the likelihood function for the half normal model introduced by Aigner et al. (1977). Similarly, the truncated

Table 2.1. Key Results

$f(\varepsilon)$ is the density of $\varepsilon = v - u$, $\mathbb{E}(u|\varepsilon)$ is the conditional mean of u given ε, and $f(u|\varepsilon)$ is the conditional density of u given ε. $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively. And $\mathbf{1}_{[0,B]}(\cdot)$ is an indicator function.

Model	$f(\varepsilon)$	$\mathbb{E}(u\|\varepsilon)$	$f(u\|\varepsilon)$
Doubly truncated normal	$\left[\Phi(\frac{B-\mu}{\sigma_u}) - \Phi(\frac{-\mu}{\sigma_u})\right]^{-1} \cdot \left[\frac{1}{\sigma}\phi(\frac{\varepsilon+\mu}{\sigma})\right] \cdot \left[\Phi(\frac{(B+\varepsilon)\lambda+(B-\mu)\lambda^{-1}}{\sigma}) - \Phi(\frac{\varepsilon\lambda-\mu\lambda^{-1}}{\sigma})\right]$	$\mu_* + \sigma_* \frac{\phi(-\frac{\mu_*}{\sigma_*})-\phi(\frac{B-\mu_*}{\sigma_*})}{\Phi(\frac{B-\mu_*}{\sigma_*})-\Phi(-\frac{\mu_*}{\sigma_*})}$	$\frac{\frac{1}{\sigma_*}\phi(\frac{u-\mu_*}{\sigma_*})}{\Phi(\frac{B-\mu_*}{\sigma_*})-\Phi(-\frac{\mu_*}{\sigma_*})}\mathbf{1}_{[0,B]}(u)$
	$\sigma = \sqrt{\sigma_u^2+\sigma_v^2}$, $\lambda = \sigma_u/\sigma_v$	$\mu_* = \frac{\mu\sigma_v^2-\varepsilon\sigma_u^2}{\sigma^2}$, $\sigma_* = \frac{\sigma_u\sigma_v}{\sigma}$	
Truncated half normal	$[\Phi(\frac{B}{\sigma_u}) - 1/2]^{-1} \cdot \frac{1}{\sigma}\phi(\frac{\varepsilon}{\sigma}) \cdot [\Phi(\frac{(B+\varepsilon)\lambda+B\lambda^{-1}}{\sigma}) - \Phi(\frac{\varepsilon\lambda}{\sigma})]$	$\mu_* + \sigma_* \frac{\phi(-\frac{\mu_*}{\sigma_*})-\phi(\frac{B-\mu_*}{\sigma_*})}{\Phi(\frac{B-\mu_*}{\sigma_*})-\Phi(-\frac{\mu_*}{\sigma_*})}$	$\frac{\frac{1}{\sigma_*}\phi(\frac{u-\mu_*}{\sigma_*})}{\Phi(\frac{B-\mu_*}{\sigma_*})-\Phi(-\frac{\mu_*}{\sigma_*})}\mathbf{1}_{[0,B]}(u)$
		$\mu_* = -\frac{\varepsilon\sigma_u^2}{\sigma^2}$, $\sigma_* = \frac{\sigma_u\sigma_v}{\sigma}$	
Truncated exponential	$\frac{e^{\frac{\varepsilon}{\sigma_u}+\frac{\sigma_v^2}{2\sigma_u^2}}[\Phi(\frac{B+\varepsilon}{\sigma_v}+\frac{\sigma_v}{\sigma_u})-\Phi(\frac{\varepsilon}{\sigma_v}+\frac{\sigma_v}{\sigma_u})]}{\sigma_u(1-e^{-B/\sigma_u})}$	$\mu_* + \sigma_v \frac{\phi(-\frac{\mu_*}{\sigma_v})-\phi(\frac{B-\mu_*}{\sigma_v})}{\Phi(\frac{B-\mu_*}{\sigma_v})-\Phi(-\frac{\mu_*}{\sigma_v})}$	$\frac{\frac{1}{\sigma_v}\phi(\frac{u-\mu_*}{\sigma_v})}{\Phi(\frac{B-\mu_*}{\sigma_v})-\Phi(-\frac{\mu_*}{\sigma_v})}\mathbf{1}_{[0,B]}(u)$
		$\mu_* = -\varepsilon - \frac{\sigma_v^2}{\sigma_u}$	

exponential model reduces to the exponential model introduced by Meeusen and van den Broeck (1977).

2.3. The Skewness Issue

A common and important methodological problem encountered when dealing with empirical implementation of the stochastic frontier model is that the residuals may be skewed in the wrong direction. In particular, the ordinary least squares (OLS) residuals may show positive skewness even though the composed error term $v-u$ should display negative skewness, in keeping with $u's$ positive skewness. This problem has important consequences for the interpretation of the skewness of the error term as a measure of technical inefficiency. It may imply that a nonrepresentative random sample had been drawn from an inefficiency distribution possessing the correct population skewness (see Carree, 2002; Greene, 2007; Almanidis and Sickles, 2009; Simar and Wilson, 2010). This is considered a finite sample "artifact" and the usual suggestion in the literature and by programs implementing stochastic frontier models is to treat all firms in the sample as fully efficient and proceed with straightforward OLS based on the results of Olson et al. (1980) and Waldman (1982). As this would suggest setting the variance of the inefficiency term to zero, it would have problematic impacts on estimation and on inference. Simar and Wilson (2010) suggest a bagging method to overcome the inferential problems when a half-normal distribution for inefficiencies is specified. However, a finding of positive skewness in a sample may also indicate that inefficiencies are

in fact drawn from a distribution which has positive skewness. Carree (2002) considers one-sided distribution for inefficiencies that can have negative or positive skewness. He employes the binomial distribution, wherein continuous inefficiencies fall into discrete "inefficiency categories". Besides being discrete, the binomial distribution implicitly assumes that only a very small fraction of the firms attain a level of productivity close to the frontier, especially when u_i is negatively skewed.

Our model addresses the "wrong skewness" problem in the spirit of Carree (2002), but with a more appealing distributional specification on the efficiency term. For the doubly truncated normal model, let $\xi_1 = \frac{-\mu}{\sigma_u}$, $\xi_2 = \frac{B-\mu}{\sigma_u}$, and $\eta_k \equiv \frac{\xi_1^k \phi(\xi_1) - \xi_2^k \phi(\xi_2)}{\Phi(\xi_2) - \Phi(\xi_1)}$, $k = 0, 1, ..., 4$. Note that η_0 is the inverse Mill's ratio and it is equal to $\sqrt{2/\pi}$ in the half normal model, and that ξ_1 and ξ_2 are the lower and upper truncation points of the standard normal density, respectively. The skewness of the doubly truncated normal distribution is then given by

$$S_u = \frac{2\eta_0^3 - \eta_0(3\eta_1 + 1) + \eta_2}{(1 - \eta_0^2 + \eta_1)^{3/2}}. \tag{2.6}$$

It can be checked that when $B > 2\mu$, S_u is positive. And when $B < 2\mu$, S_u is negative. Since $B > 0$ by definition, it is obvious that only when $\mu > 0$ is it possible for u_i to be negatively skewed. And the larger μ is, the larger range of values B may take such that u_i is negatively skewed. Consider the limiting case where a normal distribution with $\mu \to \infty$ is truncated at zero and $B > 0$. An infinitely large μ means that there is effectively no truncation on the left at all and

that any finite truncation on the right gives rise to a negative skewness. Finally, for both the truncated half normal model ($\mu = 0$) and the truncated exponential model, the skewness of u_i is always positive.

Consequently, the doubly truncated normal model has a residual that has an ambiguous sign of the skewness, which depends on an unobservable relationship between the truncation parameter B and μ. We argue that this ambiguity theoretically could explain the prevalence of the "wrong" skewness problem in applied stochastic frontier research. Here, the term wrong is set in quotes to point out that the conventional wisdom that positive skewness is inconsistent with the standard stochastic frontier production model errors skewness is not necessarily the correct wisdom. When the underlying data generating process for u_i is based on the doubly truncated normal distribution, increasing sample size does not solve the "wrong" skewness problem. The skewness of the OLS residual ε may be positively skewed even when sample size goes to infinity. Hence the "wrong" skewness problem also may be a large sample problem.

In finite samples, we may use simulations to show that our model is capable of generating residuals with "wrong" skewness with higher frequency than do traditional stochastic frontier models (Simar and Wilson, 2010). We generate samples of the residuals $\varepsilon = v - u$ with u being doubly truncated normal. We then calculate the proportion of samples with positively skewed residuals in 1000 repeated experiments. We set the parameter μ to 1 and examine the proportions of positive

Table 2.2. Proportion of Positive Skewness for Simulated Residuals in the Doubly Truncated Normal Model.

	n	$B = 1$	$B = 2$	$B = 5$
	50	0.519	0.505	0.480
	100	0.481	0.501	0.516
	200	0.495	0.473	0.514
$\lambda = 0.1$	500	0.487	0.503	0.539
	10^3	0.520	0.516	0.510
	10^4	0.504	0.483	0.512
	10^5	0.532	0.492	0.437
	50	0.517	0.485	0.503
	100	0.545	0.491	0.459
	200	0.551	0.490	0.486
$\lambda = 0.5$	500	0.520	0.488	0.431
	10^3	0.564	0.514	0.453
	10^4	0.684	0.491	0.397
	10^5	0.759	0.496	0.107
	50	0.565	0.536	0.367
	100	0.524	0.513	0.317
	200	0.529	0.512	0.224
$\lambda = 1$	500	0.567	0.514	0.155
	10^3	0.576	0.524	0.063
	10^4	0.709	0.501	0
	10^5	0.943	0.503	0

skewness when B is $1, 2, 5$, and 10. We also experiment with different values of λ and sample sizes from 50 to 10^5. The results are reported in Table 2.2.

The first column ($B = 1$) shows that the proportion of the samples with the positive ("wrong") skewness increases as the sample size gets larger. It appears to

converge to one as the sample size increases, especially when the signal-noise-ratio λ is large. The second column corresponds to the case where $B = 2\mu$. In this case there is about a $50 - 50$ chance that we generate a sample with positive skewness. In other words, the positive skewness appears to be statistically insignificant in most of the cases. The third column ($B = 5$) corresponds to the case where the distribution of inefficiencies is positively skewed. The results in this column are similar with those reported in Simar and Wilson (2010) for traditional stochastic frontier models.

Our simulation results confirm that the skewness issue is also a large sample issue, since for $B < 2\mu$ the proportion of the samples with positive skewness converges to one. This would mean that if the true data generating process is based on inefficiencies that are drawn from a doubly truncated normal distribution, and if a researcher fails to recognize this and finds a skewness statistic with the wrong sign, then she may erroneously reject her model. Moreover, if there is the potential for increasing the sample size and the researcher keeps increasing it and finds continuously positive signs of skewness, then she may erroneously conclude that all firms in her sample are super efficient. The bounded inefficiency model, the doubly truncated normal model in particular, avoids this problem.

As a conclusion of this section, the doubly truncated normal model generalizes the stochastic frontier model in a way that allows for negative as well as positive skewness for the residual. In addition, although the truncated half normal and the truncated exponential models have correct (negative) skewness in the limit, the

existence of the inefficiency bound reduces the identifiability of negative skewness in finite samples, often to the extent that "wrong" skewness appears. This implies that finding a "wrong" skewness does not necessarily mean that the stochastic frontier model is inapplicable. It may only be that we are studying a market or an industry in which firms do not fall below some minimal level of efficiency in order to remain in the market or industry. Hence the traditional unbounded support for the inefficiency term would be misspecified and should be substituted with the model of bounded inefficiency.

2.4. Estimation

2.4.1. Identification

Identification of our model may be done in two parts. The first part is concerned with the parameters describing the technology, and the second part identifies the distributional parameters using the information contained in the distribution of the residual. For models without an intercept term the identification conditions for the first part are well known and are satisfied in most of the cases. The structural parameters can be consistently obtained by applying straightforward OLS. However, for models containing an intercept term there is a need to bias correction it using the distributional parameters since $E[\varepsilon] = -E[u] \neq 0$ (see Afriat, 1972 and Richmond, 1974). Therefore, the identification of the second part, which is based on method of moments requires a closer examination. Table 2.3 lists

Table 2.3. Central Moments of ε

Moment	Doubly-truncated-normal
ψ_1	$-\mu - \sigma_u \eta_0$
ψ_2	$\sigma_u^2 (1 - \eta_0^2 + \eta_1) + \sigma_v^2$
ψ_3	$-\sigma_u^3 (2\eta_0^3 - 3\eta_1\eta_0 - \eta_0 + \eta_2)$
ψ_4	$\sigma_u^4 (3 + 3\eta_1 + \eta_3 - 2\eta_0^2 - 4\eta_0\eta_2 + 6\eta_0^2\eta_1 - 3\eta_0^4)$
	$+6\sigma_u^2\sigma_v^2 (1 - \eta_0^2 + \eta_1) + 3\sigma_v^4$
ψ_5	$-10\sigma_v^2\sigma_u^3 (2\eta_0^3 - 3\eta_1\eta_0 - \eta_0 + \eta_2)$
	$-\sigma_u^5 (\eta_4 + 4\eta_2 - 5\eta_0\eta_3 + 10\eta_0^2\eta_2 - 10\eta_0^3\eta_1 + 10\eta_0^3 - 15\eta_0\eta_1 + 4\eta_0^5 - 7\eta_0)$
	See the text for the definitions of η_k, $k = 0, ..., 4$.
	Truncated-exp.
ψ_1	$-\sigma_u \left(1 - \frac{\kappa}{e^\kappa - 1}\right)$
ψ_2	$\sigma_v^2 + \sigma_u^2 \frac{e^{2\kappa} - \left(\kappa^2+2\right)e^\kappa + 1}{e^{2\kappa} - 2e^\kappa + 1}$
ψ_3	$-\sigma_u^3 \frac{2e^{3\kappa} - \left(\kappa^3+6\right)e^{2\kappa} + \left(6-\kappa^3\right)e^\kappa - 2}{e^{3\kappa} - 3e^{2\kappa} + 3e^\kappa - 1}$
ψ_4	$\sigma_u^4 \frac{-9e^{4\kappa} + 36e^{3\kappa} - 54e^{2\kappa} + 36e^\kappa - 9 + 6\kappa^2 e^\kappa (e^{2\kappa} - 2e^\kappa + 1) + \kappa^4 e^\kappa (e^{2\kappa} + e^\kappa + 1)}{-e^{4\kappa} + 4e^{3\kappa} - 6e^{2\kappa} + 4e^\kappa - 1}$
	$+6\sigma_v^2\sigma_u^2 \frac{e^{2\kappa} - \left(\kappa^2+2\right)e^\kappa + 1}{e^{2\kappa} - 2e^\kappa + 1} + 3\sigma_v^4, \quad \kappa = B/\sigma_u.$

the population (central) moments of (ε_i) for the doubly truncated normal model and the truncated exponential model. The moments of the truncated half normal model can be obtained by setting $\mu = 0$ in the doubly truncated normal model. These results are essential for the discussion of identification and the method of moments estimation.

To examine the identification of the second part we note that under the assumption of independence of the noise and inefficiency term the following equality holds

$$\psi_4 - 3\psi_2^2 = E\left[(u - E(u))^4\right] - 3\left(E\left[(u - E(u))^2\right]\right)^2$$

This is a measure of excess kurtosis and for the truncated half-normal model is derived as

$$\psi_4 - 3\psi_2^2 = \sigma_u^4(-\xi^3\tilde{\eta}_0 + 3\xi\tilde{\eta}_0 - 4\xi^2\tilde{\eta}_0^2 - 4\tilde{\eta}_0^2 - 3\xi^2\tilde{\eta}_0^2 - 12\xi\tilde{\eta}_0^3) \tag{2.7}$$

where $\tilde{\eta}_0 = \frac{(2\pi)^{-1/2} - \xi\phi(\xi)}{\Phi(\xi) - \frac{1}{2}}$. Notice that for the normal distribution $\tilde{\eta}_0 = 0$ and thus the excess kurtosis is also zero.

After multiplying (2.7) by $\psi_3^{-4/3}$ we eliminate σ_u and the resulting function, which we denote by g has only one argument ξ

$$g(\xi) = \frac{-\xi^3\tilde{\eta}_0 + 3\xi\tilde{\eta}_0 - 4\xi^2\tilde{\eta}_0^2 - 4\tilde{\eta}_0^2 - 3\xi^2\tilde{\eta}_0^2 - 12\xi\tilde{\eta}_0^3}{\left(2\tilde{\eta}_0^3 - 3\xi\tilde{\eta}_0^2 - \tilde{\eta}_0 + \xi^2\tilde{\eta}_0\right)^{-4/3}} \tag{2.8}$$

The weak law of large numbers implies that

$$m_k = \text{plim}\frac{1}{n}\sum\nolimits_i \hat{\varepsilon}_i^k = \psi_k$$

where m_k denotes the k^{th} central sample moment of the least squares residuals $\hat{\varepsilon}$. By employing the Slutsky theorem we can specify the following function G:

$$\begin{aligned} g(\xi) &= \frac{m_4 - 3m_2^2}{m^{4/3}} \\ &\Longrightarrow \\ G(\xi) &= g(\xi) - \frac{m_4 - 3m_2^2}{m^{4/3}} \end{aligned}$$

Similarly, we can derive the function G for the normal/truncated exponential model with function g expressed by

$$g(\xi) = \frac{36e^{2\xi} - 24e^{\xi} - 24e^{3\xi} + 6e^{4\xi} - \xi^4 e^{\xi} - 4\xi^4 e^{2\xi} - \xi^4 e^{3\xi} + 6}{(6e^{2\xi} - 4e^{\xi} - 4e^{3\xi} + e^{4\xi} + 1)(-\frac{2e^{3\xi} - (\xi^3+6)e^{2\xi} + (6-\xi^3)e^{\xi} - 2}{e^{3\xi} - 3e^{2\xi} + 3e^{\xi} - 1})^{4/3}} \tag{2.9}$$

Both the truncated half normal model and the truncated exponential model are globally identified. To see this, we can examine the monotonicity of the function G with respect to the parameter ξ which will allow us to express this parameter (implicitly) as a function of sample moments and data. This condition provides the necessary and sufficient condition for global identification ala Rothenberg (1971). For the truncated half normal model, G is monotonically decreasing and for the truncated exponential model, G is monotonically increasing. Hence, in both cases, G is invertible and ξ can be identified. The identification of other parameters then follows from the third order moment of least squares residuals. Note, however, that for large values of ξ (e.g., $\xi > 5$ for the normal/truncated half-normal model

and $\xi > 20$ for the normal/truncated exponential model), the curve $g(\xi)$ is nearly flat and gives poor identification. ξ can be large for two reasons: either σ_u goes to zero or the bound parameter is relatively large. In the first case the distribution of the inefficiency process approaches the Dirac-delta distribution which makes it very hard for the distributional parameters to be identified. This limiting case is discussed in Wang and Schmidt (2008). In the second case the distribution of the inefficiency term becomes unbounded as in the standard stochastic frontier models, where it is straightforward to show the global identification (Aigner et al., 1977 and Olson et al., 1980).

It is not clear, however, that the doubly truncated normal model is globally identifiable. However, local identification can be verified. We may examine $\psi_3^{-4/3}(\psi_4 - 3\psi_2^2)$ and $\psi_3^{-5/3}(\psi_5 - 10\psi_2\psi_3)$, both of which are functions of ξ_1 and ξ_2 only and we denote them as $g_1(\xi_1, \xi_2)$ and $g_2(\xi_1, \xi_2)$, respectively. Let $\hat{g}_1$ and $\hat{g}_2$ be the sample versions of g_1 and g_2, respectively, we have the following system of identification equations,

$$\begin{aligned} G_1(\xi_1, \xi_2) &\equiv g_1(\xi_1, \xi_2) - \hat{g}_1 = 0 \\ G_2(\xi_1, \xi_2) &\equiv g_2(\xi_1, \xi_2) - \hat{g}_2 = 0 \end{aligned}$$

By the implicit function theorem, the identification of ξ_1 and ξ_2 depends on the matrix

$$H = \begin{pmatrix} \frac{\partial g_1}{\partial \xi_1} & \frac{\partial g_1}{\partial \xi_2} \\ \frac{\partial g_2}{\partial \xi_1} & \frac{\partial g_2}{\partial \xi_2} \end{pmatrix}$$

If H is of full rank, then ξ_1 and ξ_2 can be written as functions of $\hat{g}_1$ and $\hat{g}_2$; the identification of the model then follows. The analytic form of H is very complicated, but we may examine the invertibility of H by numerically evaluating g_1 and g_2 and inferring the sign of each element in H. It can be verified that the determinant of H is nonzero in neighborhoods within I_1, I_2, and I_4, the definitions of which are given as follows,

(i) $I_1 \equiv \{(\mu, B) | \mu \leq 0, B > 0\}$

(ii) $I_2 \equiv \{(\mu, B) | \mu > 0, B \in (0, 2\mu)\}$

(iii) $I_3 \equiv \{(\mu, B) | B = 2\mu > 0\}$

(iv) $I_4 \equiv \{(\mu, B) | \mu > 0, B > 2\mu\}$.

The line $I_3 \equiv \{(\mu, B) | B = 2\mu > 0\}$ corresponds to the case where $\psi_3 = 0$. Hence, the functions g_1 and g_2 are not continuous and the implicit function theorem is not applicable. Nonetheless, simulation results in the next section show that when the true values of B and μ satisfy $B = 2\mu$, both B and μ are consistently estimated. This may indicate that the restricted ($B = 2\mu$) model may be nested in the unrestricted model and the model is locally identifiable on $I_2 \bigcup I_3 \bigcup I_4$.

We may treat the doubly truncated normal model as a collection of different sub-models corresponding to the different domains of parameters. Treated

separately, each of the sub-models is globally identified. In maximum likelihood estimation, the separate treatment is easily achieved by constrained optimization on each parameter subset. For example, on the line of $\{(\mu, B)|\mu = 0, B > 0\} \subset I_1$, the doubly truncated normal model reduces to the truncated half normal model. As another useful example, the line I_2 corresponds to a sub-model that has positive skewness even asymptotically.

2.4.2. Method of Moment Estimation

The method of moments (Olson et al., 1980) may be employed to estimate our model or to obtain initial values for maximum likelihood estimation. In the first step of this approach, OLS is used to obtain consistent estimates of the parameters describing the technology, apart from the intercept. In the second step, using the distributional assumptions on the residual, equations of moment conditions are solved to obtain estimates of the parameters describing the distribution of the residual.

More specifically, we may rewrite the production frontier model in (2.1) and (2.2) as

$$y_i = (\alpha_0 - \mathbb{E}u_i) + \sum_{k=1}^{K} \alpha_k x_{i,k} + \varepsilon_i^*$$

where $\varepsilon_i^* = \varepsilon_i + (\mathbb{E}u_i)$ has zero mean and constant variance σ_ε^2. Hence OLS yields consistent estimates for ε_i^* and α_k, $k = 1, ..., K$. Equating the sample moments of

estimated residuals ($\hat{\varepsilon}_i^*$) to the population moments, one can solve for the parameters associated with the distribution of (ε_i^*).

2.4.3. Maximum Likelihood Estimation

For more efficient estimation, we may use maximum likelihood estimation (MLE). Note that with the presence of a noise term v_i, the range of residual is unbounded and does not depend on the parameter. No other standard regularity conditions might be questioned. In the remainder of this section we provide the log-likelihood functions for the bounded inefficiency model for the three parametric distributions we have considered.

The log-likelihood function for the doubly truncated normal model with $\gamma-$ parameterization is given by

$$\begin{aligned}\ln L &= -n\ln\left[\Phi(\frac{-\ln\tilde{B}-\mu}{\sigma_u(\sigma,\gamma)})-\Phi(\frac{-\mu}{\sigma_u(\sigma,\gamma)})\right]-n\ln(2\pi\sigma^2)^{1/2}\ -\sum_{i=1}^{n}\frac{(\varepsilon_i+\mu)^2}{2\sigma^2}\\ &+\sum_{i=1}^{n}\ln\{\Phi(\frac{(-\ln\tilde{B}+\varepsilon_i)\sqrt{\gamma/(1-\gamma)}-(\ln\tilde{B}+\mu)\sqrt{(1-\gamma)/\gamma}}{\sigma}\\ &-\Phi(\frac{\varepsilon_i\sqrt{\gamma/(1-\gamma)}-\mu\sqrt{(1-\gamma)/\gamma}}{\sigma})\}\end{aligned} \tag{2.10}$$

where $\varepsilon_i = y_i - x_i\alpha$, $x_i = (1, x_{ik})$, and $\alpha = (\alpha_0, \alpha_k)'$.

$$\sigma_u(\sigma,\gamma) = \sigma\sqrt{\gamma} \tag{2.11}$$

This can be expressed in terms of the $\lambda-$parameterization as in Aigner et al. (1977) by substituting γ in (2.10) with

$$\gamma(\lambda) = \frac{\lambda^2}{1+\lambda^2} \tag{2.12}$$

In addition to the $\gamma-$parameterization discussed earlier, we re-parametrize the bound parameter with another parameter $\tilde{B} = \exp(-B)$. Unlike the bound, $\tilde{B}$ takes values in compact unit interval which facilitates the numerical procedure of maximum likelihood estimation as well as establishing the asymptotic normality of this parameter. When $\tilde{B}$ lies in the interior of the parameter space, the MLE estimator is asymptotically normal (see Rao, 1973 and Davidson and MacKinnon, 1993 among others).

The log-likelihood function for the truncated half normal model is

$$\begin{aligned} \ln L \;=\; & -n\ln(\Phi(\frac{-\ln\tilde{B}}{\sigma_u(\sigma,\gamma)}) - \frac{1}{2}) - n\ln(2\pi\sigma^2)^{1/2} - \sum_{i=1}^{n}\frac{\varepsilon_i^2}{2\sigma^2} \\ & + \sum_{i=1}^{n}\ln\{\Phi(\frac{(-\ln\tilde{B}+\varepsilon_i)\sqrt{\gamma/(1-\gamma)} - \ln\tilde{B}\sqrt{(1-\gamma)/\gamma}}{\sigma}) \\ & -\Phi(\frac{\varepsilon_i\sqrt{\gamma/(1-\gamma)}}{\sigma})\} \end{aligned} \tag{2.13}$$

Finally, the log-likelihood function for the truncated exponential model is given by

$$\begin{aligned}\ln L &= -\frac{n}{2}\ln\gamma - n\ln\sigma - n\ln(1-e^{\frac{\ln\tilde{B}\gamma^{-1/2}}{\sigma}}) + \frac{n}{2}\frac{1-\gamma}{\gamma} \\ &+\frac{\gamma^{-1/2}}{\sigma}\sum_{i=1}^{n}\varepsilon_i + \sum_{i=1}^{n}\ln[\Phi(\frac{(-\ln\tilde{B}+\varepsilon_i)(1-\gamma)^{-1/2}}{\sigma} + \sqrt{\frac{1-\gamma}{\gamma}}) \\ &-\Phi(\frac{\varepsilon_i(1-\gamma)^{-1/2}}{\sigma} + \sqrt{\frac{1-\gamma}{\gamma}})]\end{aligned} \tag{2.14}$$

where $\varepsilon_i = y_i - x_i\alpha$.

Note that in practice we may also need the gradients of the log likelihood function. The gradients are complicated in form but straightforward to derive. These are provided in the appendix.

After estimating the model, we can estimate the composed error term ε_i:

$$\hat{\varepsilon}_i = y_i - \hat{\alpha}_0 - \sum x_{i,k}\hat{\alpha}_k, i = 1, \cdots, n \tag{2.15}$$

From this we can estimate the inefficiency term u_i using the formula for $E(u_i|\varepsilon_i)$ provided in Table 2.1.

One reasonable question is whether or not one can test for the absence or the presence of the bound ($H_0 : \tilde{B} = 0$ vs. $H_1 : \tilde{B} > 0$), which one may wish to test since this would suggest that the proper specification would be the standard SF model which assumes no bound as a special case of our more general bounded SF model. The test procedure is slightly complicated but still feasible. The first

complication arises from the fact that $\tilde{B}$ lies on the boundary of the parameter space under the null. Second, it is obvious from the log-likelihood functions provided above that the bound is not identified in this case and it can be shown that any finite order derivative of the log-likelihood function with respect to $\tilde{B}$ is zero. Thus the conventional Wald and Lagrange Multiplier (LM) statistics are not defined and the Likelihood Ratio (LR) statistic has a nonstandard asymptotic distribution that strictly would dominate the $\chi^2_{(1)}$ distribution. Lee (1993) derives the asymptotic distribution of such an estimate as a mixture of χ^2 distributions under the null that its value is zero, focusing in particular on the SF model under the assumption of half-normally distributed inefficiencies. Here λ is globally identified, which can also be seen using the method of moments estimator provided in Aigner et al. (1977). Lee (1993) provides useful one-to-one reparameterization which transform the singular information matrix into a nonsingular one. However, since the bound in our model case is not identified in this situation, there is no such re-parameterization and hence this procedure cannot be used. An alternative is to apply the bootstrap procedure proposed by Hansen (1996, 1999) to construct asymptotically equivalent $p - values$ to make an inference. To implement the test we treat the $\hat{\varepsilon}_i$ $(i = 1, ..., n)$ as a sample from which the bootstrap samples $\hat{\varepsilon}_i^{(m)}$ $(i = 1, ..., n; m = 1, ..., M)$ are drawn with replacement. Using the bootstrap sample we estimate the model under the null and the alternative of bounded inefficiency and construct the corresponding LR statistic. We repeat this procedure M times and calculate the percentage of times the bootstrap LR exceeds the actual

one. This provides us with the bootstrap estimate of the asymptotic $p - value$ of LR under the null.

2.5. Panel Data

In the same spirit as in Schmidt and Sickles (1984) and Cornwell et al. (1990), we may specify a panel data model of bounded inefficiencies:

$$y_{it} = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_{it,k} + \varepsilon_{it} \tag{2.16}$$

where

$$\varepsilon_{it} = v_{it} - u_{it}. \tag{2.17}$$

We assume that the inefficiency components (u_{it}) are positive, independent from the regressors, and are independently drawn from a time-varying distribution with upper bound B_t. We may set B_t to be time-invariant. However, it is certainly more plausible to assume otherwise, as the market or industry may well become more or less forgiving as time goes by, especially in settings in which market reforms are being introduced or firms are adjusting to a phased transition from regulation to deregulation.

Note that since u_{it} is time-varying, the above panel data model is in effect a time-varying technical efficiency model. Our model differs from the existing literature in that, while previous time-varying efficiency models, notably Cornwell et al. (1990), Kumbhakar (1990), Battese and Coelli (1992), and Lee and Schmidt

(1993), are time-varying in the mean or intercept of individual effects, our model is time-varying in the upper support of the distribution of inefficiency term u_i.

The assumption that u_{it} is independent over time simplifies estimation and analysis considerably. In particular, the covariance matrix of $\varepsilon_i \equiv (\varepsilon_{i1}, ..., \varepsilon_{iT})'$ is diagonal. This enables us to treat the panel model as a collection of cross-section models in the chronological order. We may certainly impose more structure on the sample path of the upper bound of u_{it}, without incurring heavy costs in terms of analytic difficulty. For example, we may impose smoothness conditions on B_t. This is empirically plausible, indeed, since changes in the market competitive conditions may come gradually. And it is also technically desirable, since imposing smoothness conditions gives us more degree of freedom in estimation, hence better estimators of model parameters. A natural way of doing this is to let B_t be a sum of weighted polynomials,

$$B_t = \sum_{i=0}^{K} b_i (t/T)^i, \quad t = 1, ..., T, \tag{2.18}$$

where (b_i) are constants. We may also use trigonometric series, splines, among others, in the modeling of B_t.

2.6. Simulations

To examine the finite sample performance of the three MLE estimators we run a series of Monte Carlo experiments for the standard cross-sectional stochastic frontier model. The data generating process is (2.1) and (2.2) with $\alpha_0 = 0$ and

$K = 2$ (two regressors and no constant term).[5] Throughout we set $\alpha_1 = 0.6$, $\alpha_2 = 0.5$. We set $\sigma_u = 0.3$ in all three submodels. To examine how the noise level (σ_v) affects the quality of estimation, we vary σ_v from $0.1, 0.2$, to 0.5. In the other dimension, we change the inefficiency bound from $0.8, 1.0$, to 1.2, to examine its impact on estimation. For both normal/truncated half normal and normal/doubly truncated normal models we use the γ-parameterization, and thus the parameters to be estimated are σ and γ as well as the production parameters. For the normal/truncated exponential model we report the estimates of parameters σ_u and σ_v themselves.

Tables 2.7 and 2.8 report results from the normal/truncated half normal model with a sample size of 200 and 1000, respectively. The results from these two tables differ only quantitatively. The first important conclusion that can be drawn is that the MLE estimators for technology parameters, α_1 and α_2, are accurately estimated. As the noise level increases, the mean squared error (MSE) of these estimates increases only marginally. The second important observation is that the estimates of the inefficiency bound have relatively smaller MSE's when the noise level is mild. When noise level is high, as when $\sigma_v = 0.5$, $\hat{B}$ becomes inaccurate. In table 2.7 distribution parameters, $\hat{\sigma}$ and $\hat{\gamma}$ display a significantly upward bias and large MSE as the signal-to-noise ratio decreases.[6] Wang and Schmidt (2008) show

[5] The results does not change very much if we include the constant term. We ommit it to save space. The results with constant term are available upon the request.

[6] It can be shown that in this case the Hessian is close to singular, which makes the estiamates of the model paramers less accurate. To our best knoweldge this pathology is shared by all likelihood based stochastic frontier models in this setting.

that the distribution of u degenerates to a point mass at $E[u]$ as λ tends to zero. This is also the case for our model. When the noise level is high relative to the variance of inefficiencies the distributional parameters are very hard identified. On the other hand, as $\lambda \rightarrow \infty$ the variance of the noise is not identified (Deterministic Frontier). Table 2.8 shows that the problem is alleviated somewhat when the sample size increases.

We now look at the doubly truncated normal model. Table 2.9 and 2.10 reports Monte Carlo results with a sample size of 200 and 1000, respectively. For both sample sizes, the technology parameter estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are quite accurate. In order to identify the distribution parameters we employ the restrictions that arise form the identification discussion of section 4. Now, the estimates of distribution parameters, σ and γ, are upward biased, especially when λ and N are relatively small. Their MSE is low for low levels of noise. In addition, the parameter μ is accurately estimated, especially then the sample size is large. The inefficiency bound is significantly distorted when the signal-to-noise ratio decreases. Finally, the case of $B = 0.8$ corresponds to the case of the "wrong" skewness. Clearly there is no problem of estimation and identification of the model parameters for this particular case.

Tables 2.11 and 2.12 provide results for the truncated exponential model with a sample size of 200 and 1000, respectively. As with the previous models, the technology parameter estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are accurately estimated. Parameters of the one-sided error term have relatively low MSE's when the noise level is mild.

It can be seen from these tables how sensitive the parameter estimates $\hat{\sigma}_u$ and B are to the level of stochastic noise. The estimated values of these parameters are highly contaminated by the noise when this dominates the inefficiency term. As expected, the finite sample problems with $\hat{\sigma}_u$ and B are lessened when we consider the larger sample size of 1000.

2.7. Efficiency Analysis of Banking Industry

2.7.1. Empirical Model and Data

We now apply the bounded inefficiency (BIE) model to an analysis of the U.S. banking industry, which underwent a series of deregulatory reforms in the early 1980's and 1990's, and experienced an adverse economic environment in the last few turbulent years of the last decade.[7] Our analysis covers a lengthy period between 1984 and 2009. What is generally observed during this period is that the number of commercial banks substantially decreased through either mergers or failures. The current number of banks is less than half of the number in 1984. It is characteristic of the fact that the number of failed banks in 2009 was about 2.75 times more than that of the period 2001-2008 due to banking crisis that was fired up in summer of 2007. As of the present time, the number of mergers and new charters has decreased, while the proportion of problematic banks has dramatically increased. However, the biggest failures occurred during the financial crisis of early 1990's

[7] These deregulations gradually allowed banks in certain states to merge with other banks across the state borders. Reigle-Neal Interstate Banking and Branching Efficiency Act that was passed by the Congress in 1994 also allowed the branching across the state lines.

where almost 400 banks failed within three years. All these facts have triggered the interest of researchers to analyze the U.S. commercial banking industry more closely and especially the performance of its institutions and their market behavior. The primary aim of our model is to capture the efficiency trends of the U.S. banking sector during all of these years until the present time, as well as to identify the toughness of the market against very inefficient firms.

Here we extend our model to the panel setting and following Adams et al. (1999) and Kneip et al. (2011), we specify a multiple output/input Cobb-Douglas stochastic output distance frontier model as follows[8]

$$Y_{it} = Y_{it}^{*\prime}\gamma + X_{it}'\beta + v_{it} - u_{it}, \tag{2.19}$$

where Y_{it} is the log of real estate loans; X_{it} is the negative of log of inputs, which include demand deposit (dd), time and savings deposit (dep), labor (lab), capital (cap), and purchased funds (purf).[9] Y_{it}^{*} includes the log of commercial and industrial loans/real estate loans (ciln) and installment loans/real estate loans (inln). In order to account for the riskiness and heterogeneity of the banks we include the log of the ratio of equity to total assets (*eqrt*) which usually measures the risk of insolvency of the banks in the banking literature.[10] The lower the ratio the more

[8] For more discussion on stochastic distance frontiers see Lovell et al. (1994).

[9] Purchased funds consist of wholesale CDs, federal funds purchased and all securities sold under agreements to resell, other borrowed money and notes issued to the U.S. Treasury, brokered deposits, and subordinated notes and debentures.

[10] We exclude from the sample banks with *eqrt* less that 0.02. Typically, these banks are close to failure and estimation of their efficiency scores require special treatments. (see Wheelock and Wilson, 2000 and Almanidis, 2010 for more discussion).

riskier a bank is considered. We assume the v_{it} are iid across i and t, and for each t, u_{it} has a upper bound B_t. Then we can treat this model as a generic panel data bounded inefficiency model as discussed in Section 2.5. Once the individual effects u_{it} are estimated, technical efficiency for a particular firm at time t is calculated as $TE = \exp(u_{it} - \max_{1 \leq j \leq N} u_{jt})$.

We use U.S. commercial banking data from 1984.Q1 through 2009.Q3. The data is a balanced panel of 4,193 commercial banks extracted from the Call Reports and the FDIC Summary of Deposits. The data set includes 431,879 observations for 103 quarterly periods. This is a fairly long panel and thus the assumption of time-invariant inefficiencies does not seem to be tenable. For this reason we compare the estimates from BIE model to the estimates from other time-varying models such as CSSW (Cornwell et al., 1990) and BC (Battese and Coelli, 1992) models, along with the baseline fixed effect estimator (FIX) of Schmidt and Sickles (1984). Descriptive statistics for the bank-level variables are given in table 2.6, where all nominal values are converted to reflect 2000 year values.

2.7.2. Results

Table 2.4 compares the parameter estimates of the bounded inefficiency (BIE) model with that of FIX, CSSW, and BC.[11] The structural parameters are statistically significant at 1% significance level and have the expected sign for all four

[11] We estimate the normal/doubly truncated normal model in order to be able to compare it with the BC model which specifies the inefficiencies to follow the truncated normal distribution.

models. The technology parameters from BIE model are somewhat different from those obtained from other models. The negative value of the coefficient of the *eqrt* implies that riskier firms tend to produce more loans, and especially real estate loans that are considered of high risk. The positive sign of the estimate of the time trend shows technological progress on average. There is a slight difference between the distributional parameters of BIE and BC model which are also statistically significant at any conventional significance level. We also tested (not reported here) other distributional specifications for BIE discussed above. The distributional parameters obtained from the normal/truncated half-normal model did not differ very much from that reported in the table, but those obtained from the normal/truncated exponential model did. However, this is not a specific to bounded inefficiency models. Similar differences have been documented in unbounded SF models as well.

We test for the extent by which the distribution of inefficiencies displays positive skewness by testing the asymmetry of the distribution of observable least squares residuals. We do so by utilizing the adjusted for skewness test statistic proposed by Bera and Premaratne (2001), which is suitable for testing for a symmetry of distributions with non-zero excess kurtosis. The test statistic is given by

$$S = \frac{n \cdot \psi_3^2}{\psi_2^3[9 + \psi_6\psi_2^{-3} - 6\psi_4\psi_2^{-2}]} \tag{2.20}$$

where ψ_i is the i^{th} population central moment of the least squares residuals and n is the number of observations in the sample. This statistic is asymptotically normally distributed and its value in our case is calculated to be 990.26, leading to rejection of the null hypothesis of symmetry at any conventional significance level. The asymmetry of the least squares residuals is also verified by quantile-quantile plot representation in figure 2.1.

We estimate the time-varying inefficiency bound using two approaches. First we estimate the bound for the panel data model without imposing any restriction on its sample path. In the second approach we specify the bound as a sum of weighted time polynomials. We choose to fit a fifth degree polynomial the coefficients of which are estimated by MLE along with the rest parameters of the model.[12] Both approaches are illustrated in figure 2.3 with their respective 95% confidence intervals. It can be seen that the inefficiency bound has had a decreasing trend up to year 2005, when the financial crisis (informally) began, and then it is increasing for the remaining periods through 2009.Q3. One interpretation of this trend can be that the deregulations of 1980's and 1990's increased competitive pressures and forced many inefficient banks to exit the industry, reducing thus the

[12]The choice of degrees of the time polynomial was based on the simple likelihood-ratio (LR) test for degrees of the polynomial ranging from 1 to 10. The maximum likelihood estimates of coefficients for this polynomial are given by
$b_0 = -3.9477e-007$, $b_1 = 0.0039509^{**}$, $b_2 = -15.816^{***}$, $b_3 = 31656^{**}$, $b_4 = -3.168e+007^{*}$, $b_5 = 1.2682e+010$

upper limit of inefficiency that banks could sustain and still remain in their particular niche market in the larger banking industry. The new upward trend can be attributed to the adverse economic environment and an increase in the proportion of banks that are characterized as "too big to fail."

Of course, for time-varying efficiency models such as CSSW, BC, and BIE, average efficiencies change over time.[13] These are illustrated in figure 2.2 along with their 95% confidence bounds. The BIE averaged efficiencies (panel 4) are significantly higher than those obtained from the fixed effect time-invariant model. However, the differences are small compared to BC and CSSW models. These small differences are not unexpected, however, since the existence of the inefficiency bound implies that the mean conditional distribution of inefficiencies is also bounded from above, resulting in higher average efficiencies. Failing to take the bound into account could possibly yield underestimated mean and individual efficiency scores (see table 2.1). We smooth the BIE averaged efficiencies by fitting ninth degree polynomial of time in order to capture their trend and also to be able to compare them with other two time-varying averaged efficiency estimates. These are represented by a curve labeled BIEsmooth. It can be seen that the efficiency trend for the BIE model is in close agreement with the CSSW model and better reflects the deregulatory reforms and consolidation of the U.S. commercial banking industry. It is increasing initially and then falls soon after the saving and loans (S&L) crisis of early 90's began. It has the decreasing pace and reaches its minium

[13]We trimmed the top and bottom 5% of inefficiencies to remove the effects of outliers.

in 1993 a year before Congress passed the Reigle-Neal Act which allowed commercial banks to merge with and acquire banks across the state lines. This spurred a new era of interstate banking and branching, which along with the Gramm-Leach-Billey Act that granted broad-based securities and insurance power to commercial banks, substantially decreased the number of banks operated in the U.S. from 10,453 in 1994 to 8,315 by the end of the millennium. After 1994 the banking industry witnessed a rapid increase in averaged efficiencies of its institutions due in part to the disappearance of inefficient banks previously sheltered from competitive pressure and due to the expansion of large banks that both financially and geographically diversified their products. The increasing trend continues until the new recessionary period of 2001 and then steadily falls thereafter until the rapid decline illustrating the effects of the 2007-2009 crisis. The CSSW model is able to show the weakness of the banking industry as early as 2005. This weakness is illustrated by the estimated inefficiency bound from the BIE model.

On the other hand, the BC model shows a slight, statistically non-significant, upward efficiency trend for all these periods ($\eta = 0.0066$). We also can look at the efficiency ranking of firms from these four estimators. Table 2.5 tabulates the Spearman rank correlations among different models, which shows that BIE efficiency ranking is in agreement with other estimators, especially with CSSW.

Table 2.4. Comparisons of Various Estimators. Estimates and standard errors (in parentheses) for each model parameters from competing models (FIX, CSSW, BC, BIE)

	FIX	*CSSW*	*BC*	*BIE*
ciln	0.2407	0.2971	0.2284	0.2838
	(0.0015)	(0.0014)	(0.0013)	(0.0012)
inln	0.2206	0.1715	0.2043	0.2609
	(0.0013)	(0.0012)	(0.0013)	(0.0013)
dd	−0.0940	−0.0935	−0.1197	−0.0996
	(0.0024)	(0.0020)	(0.0024)	(0.0020)
dep	−0.3999	−0.4037	−0.4368	−0.4053
	(0.0048)	(0.0051)	(0.0048)	(0.0034)
lab	−0.3104	−0.2219	−0.1610	−0.1892
	(0.0046)	(0.0042)	(0.0044)	(0.0020)
cap	−0.0460	−0.0464	−0.0510	−0.0965
	(0.0016)	(0.0014)	(0.0015)	(0.0015)
purf	−0.1507	−0.1658	−0.1627	−0.1665
	(0.0034)	(0.0029)	(0.0034)	(0.0031)
time	0.0057	–	0.0020	0.0021
	(0.0001)		(0.0001)	(0.0001)
eqrt	−0.1369	−0.1189	−0.0975	−0.1088
	(0.0045)	(0.0041)	(0.0044)	(0.0039)
γ	–	–	0.7980	0.7690
			(0.0115)	(0.0058)
σ	0.2210	0.2070	0.2733	0.2712
	(0.0034)	(0.0020)	(0.0045)	(0.0022)
μ	–	–	0.3240	0.3518
			(0.0139)	(0.0630)
B	–	–	–	1.5186
ATE	0.5853	0.6470	0.6410	0.6998

Table 2.5. Spearman Rank Correlations of Efficiencies

	FIX	CSSW	BC	BIE
FIX	1	.	.	.
CSSW	0.8556	1	.	.
BC	0.9662	0.8231	1	.
BIE	0.6919	0.7942	0.7168	1

In sum, figures 2.2 and 2.3 display an interesting findings: on one hand, an upward trend is observed for the average efficiency of the industry, presumably benefiting from the deregulations in the 1980's and 1990's; on the other hand, the industry appears to be more "tolerant"of less efficient banks in the last decade. Possibly, these banks have a characteristic that we have not properly controlled for and we are currently examining this issue. Given the recent experiences in the credit markets due in part to the poor oversight lending authorities gave in their mortgage and other lending activities, our results also may be indicative of a backsliding in the toleration of inefficiency that could have contributed to the problems the financial services industry faces today.

2.8. Conclusions

In this chapter we have introduced a series of parametric stochastic frontier models that have upper (lower) bounds on the inefficiency (efficiency). The model parameters can be estimated by maximum likelihood, including the inefficiency bound. In the panel data setting, we set the inefficiency bound to be varying over time, hence contributing another time-varying efficiency model to the literature.

We have examined the finite sample performance of the maximum likelihood estimator in the cross-sectional setting. We also have showed how the "wrong "skewness problem inherent in traditional stochastic frontier model can be avoided when the bound is taken into account. An empirical analysis of U.S. commercial banking industry using the new model revealed interesting trends in efficiency scores.

2.9. Appendix: First-order derivatives of the log-likelihood function

The first-order derivatives of the normal/doubly truncated normal model are calculated based on log-likelihood function (2.10) and are given by

$$\frac{\partial \ln L}{\partial a}=\sum_{i=1}^{n}\frac{(\varepsilon_i+\mu)x_i}{\sigma^2}+\frac{\sqrt{\gamma/(1-\gamma)}}{\sigma}\sum_{i=1}^{n}x_i\frac{\phi(z_{4i})-\phi(z_{3i})}{\Phi(z_{3i})-\Phi(z_{4i})}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma^2} &= \frac{n}{2\sigma^2}\frac{[(z_1\phi(z_1)-z_2\phi(z_2)]}{\Phi(z_1)-\Phi(z_2)}-\frac{n}{2\sigma^2}\\ &\quad+\sum_{i=1}^{n}\frac{(\varepsilon_i+\mu)^2}{2\sigma^4}+\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{[z_{4i}\phi(z_{4i})-z_{3i}\varphi(z_{3i})]}{\Phi(z_{3i})-\Phi(z_{4i})}\end{aligned}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \lambda} &= \frac{n}{\gamma}\frac{[(z_1\phi(z_1)-z_2\phi(z_2)]}{\Phi(z_1)-\Phi(z_2)}+(\ln(\tilde{B})+\mu)\frac{1}{\gamma^2}\sqrt{\gamma/(1-\gamma)})\phi(z_{3i})\\ &\quad+\frac{1}{\sigma}\sum_{i=1}^{n}\frac{1}{\Phi(z_{3i})-\Phi(z_{4i})}\{((-\ln(\tilde{B})+\varepsilon_i)\frac{1}{(1-\gamma)^2}\sqrt{(1-\gamma)/\gamma}\\ &\quad-(\varepsilon_i\frac{1}{(1-\gamma)^2}\sqrt{(1-\gamma)/\gamma}-\mu\lambda\frac{1}{\gamma^2}\sqrt{\gamma/(1-\gamma)})\phi(z_{4i})\}\end{aligned}$$

$$\frac{\partial \ln(L)}{\partial \mu}=\frac{n}{\sigma\sqrt{\gamma}}\frac{\phi(z_1)-\phi(z_2)}{\Phi(z_1)-\Phi(z_2)}-\sum_{i=1}^{n}\frac{(\varepsilon_i+\mu)}{\sigma^2}+\frac{\sqrt{(1-\gamma)/\gamma}}{\sigma}\sum_{i=1}^{n}\frac{\phi(z_{4i})-\varphi(z_{3i})}{\Phi(z_{3i})-\Phi(z_{4i})}$$

$$\frac{\partial \ln(L)}{\partial \tilde{B}}=\frac{n}{\tilde{B}\sigma\sqrt{\gamma}}\frac{\phi(z_1)}{\Phi(z_1)-\Phi(z_2)}-\frac{1}{\tilde{B}\sigma\sqrt{(1-\gamma)\gamma}}\sum_{i=1}^{n}\frac{\phi(z_{3i})}{\Phi(z_{3i})-\Phi(z_{3i})}$$

where $z_1=-\frac{(\ln(\tilde{B})+\mu)}{\sigma\sqrt{\gamma}}$, $z_2=\frac{-\mu}{\sigma\sqrt{\gamma}}$, $z_{3i}=-\frac{(\ln(\tilde{B})-\varepsilon_i)\sqrt{\gamma/(1-\gamma)}+(\ln(\tilde{B})+\mu)\sqrt{(1-\gamma)/\gamma}}{\sigma}$, $z_{4i}=\frac{\varepsilon_i\sqrt{\gamma/(1-\gamma)}-\mu\sqrt{(1-\gamma)/\gamma}}{\sigma}$, and $\varepsilon_i=y_i-x_i\alpha$. The first-order derivatives of log-likelihood

function of normal/truncated half-normal model are obtained after substituting $\mu = 0$ in the above expressions.

The scores for normal/truncated exponential model are derived from (2.14) as

$$\frac{\partial \ln L}{\partial a} = -\frac{\gamma^{-1/2}}{\sigma}\sum_{i=1}^{n} x_i + \frac{(1-\gamma)^{-1/2}}{\sigma}\sum_{i=1}^{n}\frac{\phi(\tilde{z}_{2i}) - \phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i}) - \Phi(\tilde{z}_{2i})}x_i$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma} &= -\frac{n}{\sigma} - \frac{n \ln \tilde{B}\gamma^{-1/2}}{\sigma^2}\frac{e^{\frac{\ln \tilde{B}\gamma^{-1/2}}{\sigma}}}{1 - e^{\frac{\ln \tilde{B}\gamma^{-1/2}}{\sigma}}} \\ &\quad + \frac{(1-\gamma)^{-1/2}}{\sigma^2}\sum_{i=1}^{n}\{\frac{\phi(\tilde{z}_{2i}) - \phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i}) - \Phi(\tilde{z}_{2i})}\varepsilon_i + \frac{\phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i}) - \Phi(\tilde{z}_{2i})}\ln \tilde{B}\}\end{aligned}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma} &= -\frac{n}{2\gamma} - \frac{n \ln \tilde{B}}{2\gamma^{3/2}}\frac{e^{\frac{\ln \tilde{B}\gamma^{-1/2}}{\sigma}}}{1 - e^{\frac{\ln \tilde{B}\gamma^{-1/2}}{\sigma}}} - \frac{n}{2\gamma^2} - \frac{1}{2\gamma^{3/2}}\sum_{i=1}^{n}\varepsilon_i \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\{\frac{\phi(\tilde{z}_{2i}) - \phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i}) - \Phi(\tilde{z}_{2i})}\left(\frac{\varepsilon_i}{\sigma(1-\gamma)^{3/2}} - \frac{1}{\gamma^2}\sqrt{\frac{\gamma}{1-\gamma}}\right) \\ &\quad -\frac{\ln \tilde{B}}{\sigma(1-\gamma)^{3/2}}\frac{\phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i}) - \Phi(\tilde{z}_{2i})}\end{aligned}$$

$$\frac{\partial \ln L}{\partial \tilde{B}} = \frac{n\gamma^{-1/2}}{\sigma\tilde{B}}\frac{e^{\frac{\ln \tilde{B}\gamma^{-1/2}}{\sigma}}}{1 - e^{\frac{\ln \tilde{B}\gamma^{-1/2}}{\sigma}}} - \frac{(1-\gamma)^{-1/2}}{\tilde{B}\sigma}\sum_{i=1}^{n}\frac{\phi(\tilde{z}_{1i})}{\Phi(\tilde{z}_{1i}) - \Phi(\tilde{z}_{2i})}$$

where $\tilde{z}_{1i} = \frac{(-\ln \tilde{B} + \varepsilon_i)(1-\gamma)^{-1/2}}{\sigma} + \sqrt{\frac{1-\gamma}{\gamma}}$ and $\tilde{z}_{1i} = \frac{\varepsilon_i(1-\gamma)^{-1/2}}{\sigma} + \sqrt{\frac{1-\gamma}{\gamma}}$.

Table 2.6. Descpriptive statistics for bank-specific variables

Variable Name	Mean	Median	SD
Real Estate loans	212968	17549	4341501
Commercial and Industrial loans	103272	4908	2143974
Installment loans	58869	4360	1417908
Demand Deposits	54913	7282	912761
Time and Savings Deposits	449003	46954	1.00E+07
Labor	186	29	2960
Capital	8196	913	129778
Purchased Funds	163785	13698	3322838
Ratio of Equity to Total Assets	0.1007	0.0936	0.0312

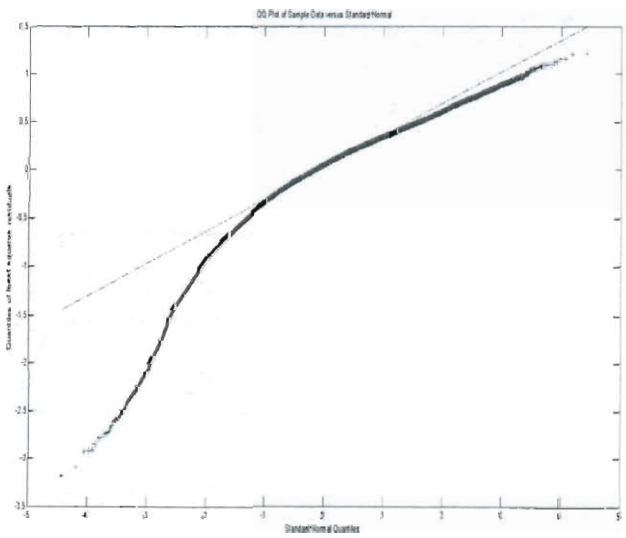


Figure 2.1. Quantile-Quantile plot

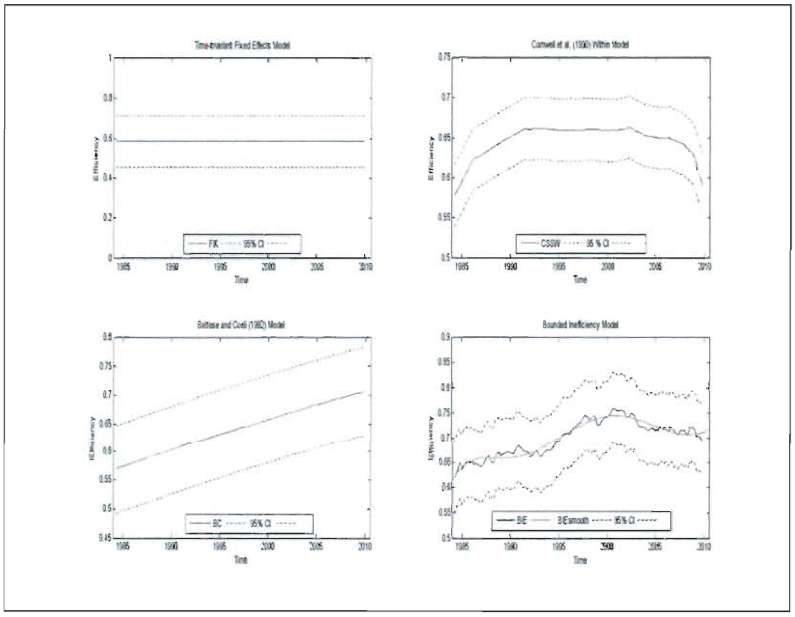


Figure 2.2. Averaged Efficiencies from each estimator

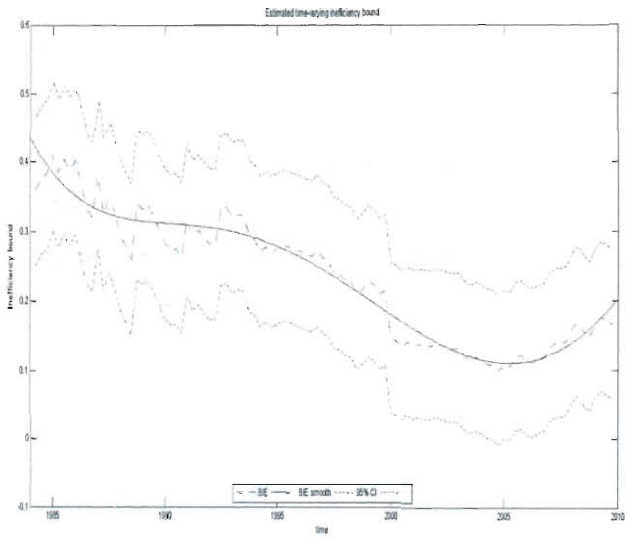


Figure 2.3. Estimated Inefficiency Bound

Table 2.7. Monte Carlo results for Truncated Half Normal model. The number of repetitions $M = 1000$. Sample size $N = 200$

			$B = 0.8$		$B = 1.0$		$B = 1.2$	
		True	AVE	MSE	AVE	MSE	AVE	MSE
$\sigma_v = 0.1$	$\hat{\sigma}$	0.3	0.3308	0.0022	0.3288	0.0013	0.3282	0.0011
	$\hat{\gamma}$	0.9	0.9079	0.0026	0.9101	0.0025	0.9107	0.0021
	$\hat{B}$		0.7987	0.0148	0.9223	0.0309	0.9600	0.0923
	$\hat{\alpha}_1$	0.6	0.6004	0.0009	0.6008	0.0009	0.6003	0.0009
	$\hat{\alpha}_2$	0.5	0.5016	0.0008	0.5022	0.0007	0.5021	0.0007
$\sigma_v = 0.2$	$\hat{\sigma}$	0.4	0.4967	0.1325	0.4464	0.0656	0.4466	0.0723
	$\hat{\gamma}$	0.7	0.7440	0.0451	0.7432	0.0354	0.7344	0.0412
	$\hat{B}$		0.8429	0.0860	0.9585	0.0990	0.9790	0.1604
	$\hat{\alpha}_1$	0.6	0.6045	0.0023	0.6024	0.0021	0.6030	0.002
	$\hat{\alpha}_2$	0.5	0.5029	0.0021	0.5061	0.0020	0.5039	0.0021
$\sigma_v = 0.5$	$\hat{\sigma}$	0.6	0.8399	0.3356	0.8538	0.3604	0.8662	0.3905
	$\hat{\gamma}$	0.3	0.4570	0.1525	0.4621	0.1636	0.4538	0.1616
	$\hat{B}$		1.0780	0.6185	1.1966	0.6121	1.2083	0.5521
	$\hat{\alpha}_1$	0.6	0.6114	0.0100	0.6169	0.0108	0.6117	0.0112
	$\hat{\alpha}_2$	0.5	0.5202	0.0116	0.5176	0.0125	0.5210	0.0127

Table 2.8. Monte Carlo results for Truncated Half Normal model. The number of repetitions $M = 1000$. Sample size $N = 1000$

			$B = 0.8$		$B = 1.0$		$B = 1.2$	
		True	AVE	MSE	AVE	MSE	AVE	MSE
$\sigma_v = 0.1$	$\hat{\sigma}$	0.3	0.3191	0.0002	0.3188	0.0002	0.3191	0.0001
	$\hat{\gamma}$	0.9	0.9020	0.0005	0.9019	0.0004	0.9027	0.0004
	$\hat{B}$		0.8045	0.0049	0.9889	0.0170	1.0918	0.0502
	$\hat{\alpha}_1$	0.6	0.6005	0.0002	0.5993	0.0002	0.6001	0.0002
	$\hat{\alpha}_2$	0.5	0.5001	0.0002	0.5013	0.0002	0.5004	0.0002
$\sigma_v = 0.2$	$\hat{\sigma}$	0.4	0.3806	0.0042	0.3735	0.0010	0.3724	0.0009
	$\hat{\gamma}$	0.7	0.7111	0.0094	0.7156	0.0056	0.7125	0.0050
	$\hat{B}$		0.8692	0.0589	1.0351	0.0808	1.1169	0.1044
	$\hat{\alpha}_1$	0.6	0.6010	0.0005	0.6027	0.0005	0.6020	0.000
	$\hat{\alpha}_2$	0.5	0.5023	0.0004	0.5021	0.0004	0.5021	0.0004
$\sigma_v = 0.5$	$\hat{\sigma}$	0.6	0.6597	0.0407	0.6568	0.0355	0.6573	0.0373
	$\hat{\gamma}$	0.3	0.3580	0.0609	0.3568	0.0597	0.3565	0.0589
	$\hat{B}$		0.9995	0.4273	1.1713	0.5255	1.2536	0.5442
	$\hat{\alpha}_1$	0.6	0.6020	0.0031	0.6028	0.0031	0.6042	0.0028
	$\hat{\alpha}_2$	0.5	0.5056	0.0028	0.5059	0.0029	0.5052	0.0026

Table 2.9. Monte Carlo results for Doubly Truncated Normal model. The number of repetitions $M = 1000$. Sample size $N = 200$

			$B = 0.8$		$B = 1.0$		$B = 1.2$	
		True	AVE	MSE	AVE	MSE	AVE	MSE
$\sigma_v = 0.1$	$\hat{\sigma}$	0.3	0.4141	0.0628	0.4227	0.0753	0.3899	0.0424
	$\hat{\gamma}$	0.9	0.9463	0.0072	0.9511	0.0085	0.9442	0.0091
	$\hat{\mu}$	0.5	0.5894	0.0329	0.5367	0.0427	0.4758	0.0377
	$\hat{B}$		0.8452	0.0239	1.0207	0.0278	1.1797	0.0411
	$\hat{\alpha}_1$	0.6	0.6046	0.0020	0.6035	0.0024	0.5990	0.0028
	$\hat{\alpha}_2$	0.5	0.5084	0.0020	0.5032	0.0023	0.5006	0.0027
$\sigma_v = 0.2$	$\hat{\sigma}$	0.4	0.4627	0.0658	0.5182	0.1026	0.4853	0.0759
	$\hat{\gamma}$	0.7	0.7937	0.0524	0.8300	0.0564	0.8173	0.0603
	$\hat{\mu}$	0.5	0.6057	0.0627	0.5875	0.0923	0.5306	0.0958
	$\hat{B}$		0.9296	0.1035	1.0963	0.1205	1.2538	0.1421
	$\hat{\alpha}_1$	0.6	0.6122	0.0040	0.6123	0.0046	0.6093	0.005
	$\hat{\alpha}_2$	0.5	0.5189	0.0045	0.5079	0.0050	0.5076	0.0058
$\sigma_v = 0.5$	$\hat{\sigma}$	0.6	0.7444	0.1064	0.7397	0.0973	0.7756	0.1238
	$\hat{\gamma}$	0.3	0.5174	0.1381	0.5338	0.1486	0.5542	0.1734
	$\hat{\mu}$	0.5	0.4491	0.1187	0.5125	0.1635	0.5647	0.2265
	$\hat{B}$		1.1524	0.6325	1.3944	0.8184	1.5888	0.9906
	$\hat{\alpha}_1$	0.6	0.6155	0.0133	0.6179	0.0150	0.6205	0.0173
	$\hat{\alpha}_2$	0.5	0.5193	0.0156	0.5172	0.0157	0.5287	0.0189

Table 2.10. Monte Carlo results for Doubly Truncated Normal model. The number of repetitions $M = 1000$. Sample size $N = 1000$

			$B = 0.8$		$B = 1.0$		$B = 1.2$	
		True	AVE	MSE	AVE	MSE	AVE	MSE
$\sigma_v = 0.1$	$\hat{\sigma}$	0.3	0.3487	0.0108	0.3336	0.0058	0.3229	0.0010
	$\hat{\gamma}$	0.9	0.9155	0.0013	0.9142	0.0015	0.9125	0.0019
	$\hat{\mu}$	0.5	0.5419	0.0088	0.5053	0.0035	0.4997	0.0044
	$\hat{B}$		0.8100	0.0056	1.0067	0.0094	1.2116	0.0157
	$\hat{\alpha}_1$	0.6	0.6025	0.0004	0.5995	0.0005	0.6014	0.0006
	$\hat{\alpha}_2$	0.5	0.5008	0.0004	0.5024	0.0005	0.5006	0.0006
$\sigma_v = 0.2$	$\hat{\sigma}$	0.4	0.4305	0.0285	0.4128	0.0236	0.3874	0.0079
	$\hat{\gamma}$	0.7	0.7348	0.0208	0.7346	0.0181	0.7260	0.0139
	$\hat{\mu}$	0.5	0.5735	0.0257	0.5298	0.0163	0.4977	0.0137
	$\hat{B}$		0.8268	0.0240	1.0441	0.0455	1.2619	0.1089
	$\hat{\alpha}_1$	0.6	0.6020	0.0008	0.6034	0.0010	0.6011	0.001
	$\hat{\alpha}_2$	0.5	0.5029	0.0008	0.5020	0.0010	0.5015	0.0012
$\sigma_v = 0.5$	$\hat{\sigma}$	0.6	0.6780	0.0400	0.7115	0.0555	0.7100	0.0527
	$\hat{\gamma}$	0.3	0.4227	0.0721	0.4601	0.0872	0.4527	0.0884
	$\hat{\mu}$	0.5	0.4771	0.0630	0.5361	0.0827	0.5253	0.0973
	$\hat{B}$		1.1325	0.7649	1.2912	0.6360	1.2649	0.6243
	$\hat{\alpha}_1$	0.6	0.6076	0.0032	0.6075	0.0032	0.6038	0.0033
	$\hat{\alpha}_2$	0.5	0.5076	0.0032	0.5086	0.0033	0.5069	0.0036

Table 2.11. Monte Carlo results for Truncated Exponential model. The number of repetitions $M = 1000$. Sample size $N = 200$

			$B = 0.8$		$B = 1.0$		$B = 1.2$	
		True	AVE	MSE	AVE	MSE	AVE	MSE
$\sigma_v = 0.1$	$\hat{\sigma}_u$	0.3	0.3062	0.0017	0.3084	0.0011	0.3006	0.0006
	$\hat{\sigma}_v$	0.1	0.0986	0.0001	0.0992	0.0001	0.0985	0.0001
	$\hat{B}$		0.7991	0.0018	0.9876	0.0027	1.1934	0.0053
	$\hat{\alpha}_1$	0.6	0.6005	0.0004	0.6038	0.0005	0.5968	0.0005
	$\hat{\alpha}_2$	0.5	0.4999	0.0003	0.4966	0.0003	0.5022	0.0004
$\sigma_v = 0.2$	$\hat{\sigma}_u$	0.3	0.3334	0.0117	0.3147	0.0048	0.3133	0.0020
	$\hat{\sigma}_v$	0.2	0.1940	0.0004	0.1962	0.0003	0.1955	0.0003
	$\hat{B}$		0.8191	0.0066	1.0150	0.0091	1.2008	0.0113
	$\hat{\alpha}_1$	0.6	0.6020	0.0014	0.6001	0.0008	0.6032	0.001
	$\hat{\alpha}_2$	0.5	0.5026	0.0009	0.5016	0.0007	0.5004	0.0008
$\sigma_v = 0.5$	$\hat{\sigma}_u$	0.3	1.0081	7.0210	0.9838	4.8403	0.7934	1.5996
	$\hat{\sigma}_v$	0.5	0.5009	0.0210	0.4869	0.0037	0.4824	0.0033
	$\hat{B}$		1.0335	0.3644	1.1115	0.3053	1.2878	0.3050
	$\hat{\alpha}_1$	0.6	0.5942	0.0108	0.6034	0.0058	0.6088	0.0060
	$\hat{\alpha}_2$	0.5	0.5166	0.0082	0.5025	0.0045	0.5266	0.0053

Table 2.12. Monte Carlo results for Truncated Exponential model. The number of repetitions $M = 1000$. Sample size $N = 1000$

			$B = 0.8$		$B = 1.0$		$B = 1.2$	
		True	AVE	MSE	AVE	MSE	AVE	MSE
$\sigma_v = 0.1$	$\hat{\sigma}_u$	0.3	0.3019	0.0008	0.3014	0.0004	0.3021	0.0003
	$\hat{\sigma}_v$	0.1	0.0992	0.0001	0.0990	0.0001	0.0988	0.0001
	$\hat{B}$		0.7992	0.0009	0.9972	0.0015	1.1975	0.0024
	$\hat{\alpha}_1$	0.6	0.5995	0.0002	0.6001	0.0002	0.6001	0.0003
	$\hat{\alpha}_2$	0.5	0.5005	0.0002	0.5003	0.0002	0.5004	0.0002
$\sigma_v = 0.2$	$\hat{\sigma}_u$	0.3	0.3186	0.0069	0.3112	0.0022	0.3074	0.0011
	$\hat{\sigma}_v$	0.2	0.1983	0.0002	0.1984	0.0002	0.1981	0.0001
	$\hat{B}$		0.8091	0.0047	1.0038	0.0060	1.1988	0.0088
	$\hat{\alpha}_1$	0.6	0.6002	0.0005	0.6014	0.0005	0.6001	0.001
	$\hat{\alpha}_2$	0.5	0.5010	0.0004	0.5004	0.0004	0.5014	0.0004
$\sigma_v = 0.5$	$\hat{\sigma}_u$	0.3	0.6274	1.2594	0.5654	0.7008	0.4593	0.3063
	$\hat{\sigma}_v$	0.5	0.5044	0.0099	0.4963	0.0034	0.5005	0.0144
	$\hat{B}$		0.9446	0.4032	1.1498	0.3430	1.2166	0.3058
	$\hat{\alpha}_1$	0.6	0.5934	0.0055	0.6028	0.0043	0.6000	0.0041
	$\hat{\alpha}_2$	0.5	0.4996	0.0048	0.5032	0.0035	0.5047	0.0038

CHAPTER 3

Skewness Issue in Stochastic Frontier Models:[1] Fact or Fiction?

3.1. Introduction

As we mentioned in the previous chapter, obtaining residuals that are skewed in the "wrong" direction constitutes one of the common and major drawbacks of the traditional stochastic frontier models (SFM). While the theory would predict a negative (positive) skewness in production (cost) frontiers in the population, researchers often discover that the sample residuals are positively (negatively) skewed. Of course, in finite samples nothing prevents the skewness statistic to have the opposite sign from that the theory would predict. Indeed this is more frequent in cases of low dominance of the inefficiency process over the two-sided noise (Carree, 2002). Simar and Wilson (2010) refer to this phenomenon as a finite sample artifact. While it may be recognized that this could arise from the models based on errors with correct skewness, researchers still consider the "wrong" skewness statistic as the indication of misspecification of the stochastic frontier model. Therefore, whenever they find the residuals skewed in the "wrong" direction they tend to believe that the model is misspecified or the data are inconsistent with

[1]This is a version of my work with professor Robin Sickles.

the SFM paradigm. Two course of actions are oftentimes undertaken: respecify the model and/or obtain a new sample which hopefully results in the desired sign of skewness. However, instead of respecifying the model, applied researchers also often respecify their interpretation of the results by assuming away inefficiencies and utilizing straightforward least squares regression approaches.[2] This weak point of the stochastic frontier models is emphasized in a series of papers, some of which try to justify that this phenomenon might arise in finite samples even for models that are correctly specified (Greene, 2007; Simar and Wilson, 2010).

The above discussion refers directly to the question raised in the title: is the "wrong" skewness just a finite sample fiction or it could be also a fact? If it is a fiction, then the bagging method, a solution proposed by Simar and Wilson (2010), should be employed to make an inference in SFMs. This method could also be generalized to the cases where efficiencies are bounded which we discussed in the previous chapter, since the normal/half-normal model is a special case of the normal/doubly truncated normal model. As we show via simulations later in this chapter, the former can be recovered from the latter without imposing any a priori restrictions on model parameters. Attributing the appearance of the "wrong" skewness exclusively to the finite sample fiction is just a one side of the same coin. We also should be concerned for the cases where this phenomenon is not a finite sample artifact but a fact.

[2]This is in particular due to the results that Olson et al. (1980) and Waldman (1982) obtain for stochastic frontier models when half-normal distribution for inefficiencies is specified.

More specifically, this chapter intends to illustrate how the bounded inefficiency formulation, discussed in the previous chapter, might overcome the issue of the "wrong" skewness in the stochastic frontier model. We first show that the imposition of an upper bound to inefficiency (lower bound to efficiency) enables the distribution of the one-sided inefficiency process to display positive and negative signs of skewness. This is in particular true for the truncated normal distribution with strictly positive mean. Imposing a bound on the truncated normal density function apart from the zero yields both positive or negative skewness depending on the position of the bound in the support of inefficiency distribution, thus justifying the occurrence of the so-called "wrong" skewness. We show that the normal/doubly truncated normal model is capable of handling and estimating the SFM with "wrong" skewness and we show that it is also quite reasonable to obtain such a pattern of residuals in large samples. Our analysis can be extended to include the gamma and the Weibull distributions as well. We perform a limited set of Monte Carlo experiments on a stochastic frontier production function with bounded inefficiency and show that when we have a positively skewed distribution of errors we can still get very reasonable maximum likelihood estimates of the disturbance and inefficiency variances, as well as other parameters of the model. An interpretation of our results is that, although a potential misspecification may occur if the stochastic frontier model is used and skewness is found to be "wrong", this can be avoided if the stochastic frontier model with bounded inefficiency is specified instead.

This chapter is structured in the following way. In section 2 the general problem of the "wrong" skewness in stochastic frontier models and its implications are discussed, as well as solutions proposed in the literature to solve it. In section 3 we show the potential of the bounded inefficiency model to address the "wrong" skewness problem and generalize Waldman's proof to formally support the use of stochastic frontier models under these circumstances. Monte Carlo simulation results and further discussion are provided in section 4. Section 5 concludes.

3.2. Skewness issue in Stochastic Frontier Analysis

3.2.1. "Wrong" skewness and its implications in frontier models

We consider here the cross-sectional classical stochastic frontier model, where the functional specification of production technology or cost is assumed to be linear in parameters. In this classical setting, the stochastic specification is $\varepsilon_i = v_i - u_i$ for production frontiers, or $\varepsilon_i = v_i + u_i$ for the case of cost frontiers. The stochastic term v_i represents the statistical noise and is usually assumed to be $iid\ \ N(0, \sigma_v^2)$ and $u_i \geq 0$ represents the inefficiency process, which also is assumed to be an iid random variable that follows some one-sided distribution. The error terms v_i and u_i are usually assumed to be statistically independent of each other and from the regressors. Under these assumptions, the distribution of the composed error term is asymmetric and non-normal implying that simple least squares applied to a linear stochastic frontier model will be inefficient and will not provide us

with an estimate of the degree of technical or cost inefficiency. However, least squares does provide consistent estimates of all parameters except the intercept since $E(\varepsilon_i) = -E(u_i) \leq 0$. Moreover,

$$E[(\varepsilon_i - E[\varepsilon_i])^3] = E[(v_i - u_i + E[u_i])^3] = -E[(u_i - E[u_i])^3] \tag{3.1}$$

which implies that the negative of the third moment of OLS residuals is a consistent estimator of the skewness of the one-sided error.

The common distributions for inefficiencies that appear in the literature are positively skewed, reflecting the fact that a large portion of the firms are expected to operate relatively close to the frontier. For production frontiers whenever we subtract the positively skewed inefficiency component from the symmetric error the composite error should display negative skewness.[3] Thus researchers find stochastic frontier models inappropriate to model inefficiencies if they obtain residuals skewed in the "wrong" direction. The typical conclusion is that, either the model is misspecified or the data is not compatible with the model. However, there can be a third interpretation as well based on the fact that inefficiencies might nave been drawn from a distribution which displays negative skewness. This simply says that if the "wrong" skewness is not a finite sample artifact but a fact, then any stochastic frontier model based on inefficiencies that are drawn form positively skewed distributions will be misspecified.

[3]We will focus on the production function but clearly all that we say about it can be said about the cost function with a sign change on the one-sided error in the composed error term.

The first formal discussion on skewness problem is found in Olson et al. (1980) in their derivation of modified ordinary least squares (MOLS) estimates as a convenient alternative to maximum likelihood estimates. They explicitly assume half-normal distribution for technical inefficiencies in their formulation. MOLS method estimates the slope parameters by OLS. These are unbiased and consistent under standard assumptions about the regressors and the error terms. OLS estimate of the constant term, however, is biased and inconsistent. The bias-corrected estimator of the constant term is obtained by adding $\sqrt{2/\pi}\sigma_u$ term, which is the expected value of the composed error term. Of course we do not know σ_u. Estimates of σ_u^2 and σ_v^2 are derived from the method of moments using the second and third moments of OLS residuals. These are consistent, but not asymptotically efficient, and are given by

$$\hat{\sigma}_u^2 = [\sqrt{\pi/2}(\frac{\pi}{\pi - 4})\hat{\mu}_3]^{2/3} \tag{3.2}$$

and

$$\hat{\sigma}_v^2 = \hat{\mu}_2 - (\frac{\pi - 2}{\pi})\hat{\sigma}_u^2 \tag{3.3}$$

where $\hat{\mu}_2$ and $\hat{\mu}_3$ are the estimated second and third moments of the OLS residuals, respectively.

It is obvious from (3.2) that a serious flaw in this method occurs whenever $\hat{\mu}_3$ is positive, since the estimated variance of inefficiencies becomes negative. This is referred as a "Type I" failure of MOLS estimators by Olson et al. (1980). Waldman (1982) proved that MLE estimate of σ_u^2 in this case is zero and that the

model parameters can be efficiently estimated by OLS. We will outline the main steps and results of Waldman's proof which are necessary benchmarks and links to our further analysis. A "Type II" failure, on the other hand, arises whenever $\hat{\mu}_2 < (\frac{\pi-2}{\pi})\hat{\sigma}_u^2$.

The log-likelihood function of normal/half-normal model is given by

$$logL = nlog(\sqrt{2/\pi}) - nlog(\sigma) + \sum_{i=1}^{n} log[1 - \Phi(\frac{\varepsilon_i \lambda}{\sigma})] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2 \qquad (3.4)$$

where $\varepsilon_i = y_i - x_i\beta$, $\lambda = \frac{\sigma_u}{\sigma_v}$, $\sigma^2 = \sigma_v^2 + \sigma_u^2$, and $\Phi(\bullet)$ denotes the cdf of the standard normal distribution. Waldman notes that there are two stationary points that potentially can characterize this log-likelihood function. Defining the parameter vector by $\theta = (\beta', \sigma^2, \lambda)$, the first stationary point would be the one for which the first derivatives of the log-liklihood function are zero while the second is the OLS solution for θ wherein the parameter λ is set to zero. The superiority of these two stationary points is then compared in cases of the wrong skewness. One way to do this is to examine the second-order derivative matrix of the log-likelihood function evaluated at these two points. The Hessian matrix evaluated at OLS solution, $\theta^* = (b', s^2, 0)$, is

$$H(\theta^*) = \begin{bmatrix} -s^{-2}\sum_{i=1}^{n} x_i x_i' & \sqrt{2/\pi}s^{-1}\sum_{i=1}^{n} x_i & 0 \\ \sqrt{2/\pi}s^{-1}\sum_{i=1}^{n} x_i & -2n/\pi & 0 \\ 0 & 0 & -n/2s^4 \end{bmatrix} \qquad (3.5)$$

where $b = (\sum_{i=1}^{n} x_i x_i')^{-1} \sum_{i=1}^{n} x_i y_i$, $s^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$, and $\hat{\varepsilon}_i$ is the least squares residual.

This matrix is singular with $k+1$ negative characteristic roots and one zero root. This essentially would require the log-likelihood function to be examined in the direction determined by the characteristic vector associated with this zero root which is given by the vector $z = (s\sqrt{2/\pi}, 1, 0)$. Departing from the point of OLS solution, the term of interest is then the sign of

$$\begin{aligned} \Delta logL &= logL(\theta^* + \delta z) - logL(\theta^*) \\ &= -\delta^2 \frac{n}{\pi} + \sum_{i=1}^{n} log[2 - 2\Phi(\hat{\varepsilon}_i \delta s^{-1} - \delta^2\sqrt{2/\pi})] \end{aligned} \tag{3.6}$$

where $\delta > 0$ is an arbitrary small number. If we expand $\Delta logL$ using a Taylor series expansion, we would obtain (see Waldman, 1982)

$$\Delta logL = (\delta^3/6s^3)\sqrt{2/\pi}[(\pi-4)/\pi]\sum_{i=1}^{n} \hat{\varepsilon}_i^3 + O(\delta^4). \tag{3.7}$$

Thus, if the term $\sum_{i=1}^{n} \hat{\varepsilon}_i^3 > 0$ then the maximum of the log-likelihood function is located at the OLS solution, which is superior to MLE. This result suggests two strategies for practitioners: apply OLS whenever the least squares residuals display positive skewness or increase the sample size, since

$$plim(\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^3) = \sigma_u^3\sqrt{2/\pi}[(\pi-4)/\pi] < 0 \tag{3.8}$$

which implies that asymptotically the sample third moment of least squares residuals converges to its population counterpart by the law of large numbers and thus the problem of the "wrong" skewness goes away.

Undoubtedly, this is true if the inefficiencies are indeed drawn from the half-normal distribution which is positively skewed. What if they are not? What if they are drawn from the distribution which displays negative skewness as well? We will attempt to give answers to these questions in the following sections.

The problem of the "wrong" skewness is also made apparent and emphasized by the two widely-used computer packages used to estimate stochastic frontiers. The first package LIMDEP 9.0, which is developed by Greene (2007), calculates and checks the skewness of the OLS residuals just before maximum likelihood estimation begins. In case the sign of the skewness statistic is positive, significantly or not, the message appears that warns the user about the misspecification of the model and suggests using OLS instead of MLE. The second software FRONTIER 4.1, produced by Coelli (1996), also first obtains the OLS estimates as a starting values for the grid search of a starting value of the γ $(= \frac{\sigma_u^2}{\sigma_u^2+\sigma_v^2})$ parameter. If the skewness is positive, the final maximum likelihood value of this parameter is very close to zero, indicating no inefficiencies. More detailed description and comparison of FRONTIER 4.1 and the earlier version 7.0 of LIMDEP can be found in Sena (1999).

Related to these results, several parametric and non-parametric test statistics have been developed to check the skewness of least squares residuals in stochastic

frontier models. Schmidt and Lin (1984) proposed the test statistic

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}} \tag{3.9}$$

where m_2 and m_3 represent the second and the third moment of the empirical distribution of the least squares residuals. The distribution of $\sqrt{b_1}$ is not standard and the application of this test requires special tables provided by D'Agostino and Pearson (1973). Coelli (1995) proposed an alternative statistic for testing whether the third moment of residuals is greater than or equal to zero

$$\sqrt{b_1^*} = \frac{m_3}{(6m_2^3/N)^{1/2}} \tag{3.10}$$

where N denotes the number of observations in the sample. Under the null hypothesis of zero skewness, the third moment of OLS residuals is asymptotically distributed as a normal random variable with zero mean and variance $6m_2^3/N$. This implies that $\sqrt{b_1^*}$ is asymptotically distributed as a standard normal variable and one can consult the corresponding statistical tables for making an inference. These two tests, although easily computed and implemented, have unknown finite sample properties. Coelli (1995) conducts Monte Carlo experiments and shows that $\sqrt{b_1^*}$ has correct size and good power in small samples, which makes it more attractive for testing for the skewness of the least squares residuals in SFM.

3.2.2. Solutions to the "wrong" skewness

Nonetheless, the standard solutions considered in the case of "wrong" skewness essentially constitute no solutions with regard to the stochastic frontier model. Setting the variance of the inefficiency process to zero based on the skewness of OLS residuals is not a very comforting solution. This solution to the problem simply would imply that all firms in the industry are fully efficient. Moreover, the estimated standard errors will not be correct if straightforward OLS is applied to the data, while data-mining techniques will introduce inferential problems and possibly biases in parameters and their standard errors (Leamer, 1978). Carree (2002), Greene (2007), and Simar and Wilson (2010) note that in finite samples, even the correctly specified stochastic frontier model is capable of producing least squares residuals with the "wrong" skewness sign with relatively high frequency. Thus another suggested solution is to get more data. Of course the availability of the data in economics is often rather limited and this alternative may not possible in many empirical settings. Another solution is to argue that the inefficiencies are drawn from an efficiency distribution with negative skewness. A major problem with this assumption is that it implies that there is only a very small fraction of the firms that attain a level of productivity close to the frontier. For example, Carree (2002) considers a distribution for inefficiencies that allows for both, negative and positive skewness.[4] He proposes a binomial distribution $b(n, p)$ which for

[4]Carree (2002) also argues that distributions with bounded range can be negatively skewed but further development of these is not pursued by the author.

a range of values of the parameter p is negatively skewed.[5] This is a discrete distribution wherein continuous inefficiencies fall into discrete "inefficiency categories". He employs the method of moments estimators as in Olson et al. (1980) and Greene (1990) and provides an explanation for how theoretically and empirically the "wrong" skewness issue may arise in stochastic frontier model.[6] Empirically, the use of the binomial distribution can be justified by a model in which the cycle of innovations and imitations occurs. This would suggest that the occurrence of positively skewed residuals would correspond to the cases where very few firms in the industry innovate while the large proportion of firms experience large inefficiencies. In contrast, as it was shown in the previous chapter, the stochastic frontier model with doubly truncated normal inefficiencies does not imply such a pattern in firms' inefficiencies, but instead it precludes the probability of occurrence of extreme inefficiencies.

[5]Other authors also considered distributions with negative skew (see Johnson et al. 1992, 1994).
[6]The shortcoming of this approach is that method-of-moments estimators may not be defined for some empirical values of the higher sample moments of the least squares residuals

3.3. Skewness statistic under the bounded inefficiencies

3.3.1. Derivation of skewness and MOLS estimates with doubly truncated normal inefficiencies

The skewness statistic is derived in chapter 2 as

$$S_u = \frac{2\eta_0^3 - \eta_0(3\eta_1 + 1) + \eta_2}{(1 - \eta_0^2 + \eta_1)^{3/2}}. \tag{3.11}$$

with

$$\eta_k \equiv \frac{\xi_1^k \phi(\xi_1) - \xi_2^k \phi(\xi_2)}{\Phi(\xi_2) - \Phi(\xi_1)} \quad k = 0, 1, 2 \tag{3.12}$$

$\xi_1 = \frac{-\mu}{\sigma_u}$ and $\xi_2 = \frac{B-\mu}{\sigma_u}$ are the lower and upper truncation points of the standard normal density function $\phi(\cdot)$, respectively. $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. η_0 represents the inverse Mill's ratio and it is equal to $\sqrt{2/\pi}$ in the normal/half-normal model.

The skewness parameter typically describes the shape of the distribution independent of location and scale. Although many non-symmetric distributions have either positive or negative sign of skewness, for the doubly truncated normal distribution the sign of skewness is ambiguous. It is either positive, whenever $B > 2\mu$, or negative when $B < 2\mu$ (for $\mu > 0$). This follows from the fact that the bound B is strictly positive by assumption.[7] The consequences of both positive and negative

[7]It should be noted that parameter μ is not restricted to be strictly positive in estimation procedure. It can take non-positive values as well.

skewness of the doubly truncated normal distribution in the SFM are not clear. The residuals can be skewed in both directions while the variance of inefficiency term is nonzero. Moreover, in finite samples the sampling variability of the skewness statistic itself could give rise to a positive or negative skewness statistic even if the population skewness parameter was negative or positive.

The second and third population central moments of the SFM residuals based on OLS are given by

$$\mu_2 = \sigma_v^2 + \sigma_u^2(1 - \eta_0^2 + \eta_1)$$

$$\mu_3 = -\sigma_u^3(2\eta_0^3 - 3\eta_1\eta_0 - \eta_0 + \eta_2)$$

from which we can obtain the method of moments estimators of σ_v^2 and σ_u^2

$$\hat{\sigma}_u^2 = [\frac{-\hat{\mu}_3}{2\eta_0^3 - 3\eta_1\eta_0 - \eta_0 + \eta_2}]^{2/3} \tag{3.13}$$

and

$$\hat{\sigma}_v = \hat{\mu}_2 - \hat{\sigma}_u^2(1 - \eta_0^2 + \eta_1) \tag{3.14}$$

Compared to (3.2), in (3.13) the positive value of $\hat{\mu}_3$ does not necessarily yield a negative variance. Here, the denominator also plays a role. Since the negative of the third moment of the OLS residuals is an unbiased and consistent estimator of the skewness of inefficiencies, one can see that the estimate of the σ_u^2 can have positive sign even in the case of positively skewed residuals. Most importantly, the "type I" failure goes away asymptotically since a positive $\hat{\mu}_3$ would imply that

the denominator of (3.13) is negative, which occurs whenever $B < 2\mu$. Thus $\hat{\sigma}_u^2$ cannot take on negative values. In cases where we have $B = 2\mu$ the ratio in (3.13) is unidentified. By applying L'Hospital rule and evaluating the limits it is straightforward to show that the variance of the inefficiency term is a strictly positive number. Only in the case when $B = 0$ is the variance of the inefficiency term zero.

We can test the extent to which the distribution of unobservable inefficiencies can display negative or positive skewness using the observable residuals based on the expression in (3.1). For this purpose we can utilize the adjusted for skewness test statistic proposed by Bera and Premaratne (2001), since the excess kurtosis is not zero. By using the standard test for skewness we will have either over-rejection or under-rejection of the null hypothesis of non-negative skewness and this will depend primarily on the sign of the excess kurtosis. In addition, since there are two points at which the doubly truncated normal distribution has zero skewness, the standard tests are not appropriate. Since the standard tests do not distinguish these two cases, application of these tests may lead researchers to accept the null hypothesis of zero variance when it is false at levels larger than nominal test size would suggest.

3.3.2. Generalization of Waldman's proof

We now examine the consistency and identifiability of the parameters of the normal/doubly truncated normal bounded inefficiency stochastic frontier model by

utilizing the same approach as in Waldman (1982). To compare and contrast the problem of the "wrong" skewness with the benchmark case of the normal/half-normal model of Aigner et al. (1977), we fix the values of the deep parameters B and μ and consider the scores of parameter vector $\theta = (\beta', \sigma^2, \lambda)$ as a function of these fixed parameters. Note that, the normal/half-normal model fixes these values at ∞ and 0, respectively. We begin by examining the second-order derivative matrix evaluated at the OLS solution point, $\theta^* = (b', s^2, 0)$:

$$H(\theta^*) = \begin{bmatrix} -s^{-2}\sum_{i=1}^{n} x_i x_i' & -\frac{1}{s^4}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \mu)x_i & 0 \\ -\frac{1}{s^4}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \mu)x_i & \frac{n}{2s^4} - \frac{1}{s^6}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \mu)^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{3.15}$$

where $b = (\sum_{i=1}^{n} x_i x_i')^{-1}\sum_{i=1}^{n} x_i y_i$, $s^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$, and $\hat{\varepsilon}_i$ is the least squares residual.

Obviously, $H(\theta^*)$ is singular with $k+1$ negative characteristic roots and one zero root. The eigenvector associated with this zero root is given by $z = (\mathbf{0}', 0, 1)$. We then need to search the sign of $\Delta logL = logL(\theta^* + \delta z) - logL(\theta^*)$ in the positive direction ($\delta > 0$), since λ is constrained to be non-negative. By expanding the $\Delta logL$, the first term in the series drops since OLS is a stationary point. The second term also vanishes since $|H(\theta^*)| = 0$. Thus, the only relevant point that remains to be considered is the third derivative of the log-likelihood function with

respect to parameter λ evaluated at the OLS solution

$$\frac{1}{6}\delta^3\frac{\partial^3 log(\theta^*)}{\partial\lambda^3} \tag{3.16}$$

Substituting for the third derivative and ignoring higher order terms, we obtain

$$\Delta logL \cong \frac{\delta^3}{6s^3}\{-2\varpi_0^3 + 3\varpi_1^3\varpi_0 + \varpi_0 - \varpi_2\}\sum_{i=1}^{n}\hat{\varepsilon}_i^3 \tag{3.17}$$

where $\varpi_k = \frac{\omega_1^k\phi(\omega_1)-\omega_2^k\phi(\omega_2)}{\Phi(\omega_2)-\Phi(\omega_1)}$, $k = 0, 1, 2$ with $\omega_1 = -\frac{\mu}{s}$ and $\omega_2 = \frac{B-\mu}{s}$.

The simple inspection of (3.17) reveals that the third order moment of the least squares residuals need not always have the opposite sign of $\Delta logL$. This will mainly depend on the relationship between the imposed bound B and the mean of the normal distribution μ. For $B < 2\mu$, ϖ_0 is negative and the term in the curly brackets becomes positive. Thus, positive skewness would imply the existence of inefficient firms in the sample. The implication of this is that whenever a researcher finds positively skewed residuals it may be the case that the inefficiencies have been drawn from a distribution that has negative skew. For $B = 2\mu$, $\Delta logL = 0$ and in this case MLE should be employed since it will be more efficient than OLS and will provide us with technical inefficiency estimates. Asymptotically the third order term of OLS residuals and the expression in curly brackets have the same sign since,

$$plim(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^3) = -\sigma_u^3(2\varpi_0^3 - 3\varpi_1^3\varpi_0 - \varpi_0 + \varpi_2)$$

which implies that we can observe the "wrong" skewness even in large samples. Thus we can argue that the problem of the "wrong" skewness is not a just finite sample issue. Positive or negative skewness of least squares residuals will always imply a positive variance of the inefficiency process in large samples. In finite samples, anything can happen. We can obtain negatively skewed residuals even if we sample from a negatively skewed distribution of inefficiencies.

3.4. Further Discussion

An important question for applied researchers is, what happens in the case if the true model is the normal/half-normal but we estimate the normal/doubly truncated normal model instead and vise-versa? To answer the first part of the question, we conduct a Monte Carlo experiment to assess the validity of the normal/doubly truncated normal model whenever the underlying true data generating process is the one proposed in Aigner et al. (1997). For this purpose, we specify a simple cross-sectional Cobb-Douglas production frontier with two inputs as:

$$y_i = \alpha_0 + \alpha_1 \ln x_{1i} + \alpha_2 \ln x_{2i} + \upsilon_i - u_i$$

where $\upsilon_i \sim^{iid} N(0, \sigma_\upsilon^2)$ and $u_i \sim^{iid} N^+(0, \sigma_u^2)$. υ_i and u_i, as previously, are assumed to be independent of each over and from regressors.

Throughout, we set $\alpha_0 = 0.9$, $\alpha_1 = 0.6$, and $\alpha_2 = 0.5$. $\ln x_{ji}|_{j=1,2}$ are drawn from $N(\mu_{xj}, \sigma_{xj}^2)$ with $\mu_{x1} = 1.5$, $\mu_{x2} = 1.8$, and $\sigma_{x1}^2 = \sigma_{x2}^2 = 0.3$. These draws are fixed across Monte Carlo replications. We keep $\sigma_u = 0.3$ and vary the σ_υ in a way

that λ^2 takes on values of 1, 10, and 100, while at the same time we vary the sample size by 100, 200, and 1000, respectively. To facilitate the numerical optimization procedure we consider the γ-parameterization instead of the λ-parameterization in the maximum likelihood estimation. We set the number Monte Carlo replications to 1000 and examine the performance of the normal/doubly truncated normal model without imposing any restrictions on model parameters. Table 3.1 reports the averaged values (AVE) of the estimates over the replications and their mean squared errors (MSE). The first case reported in the first column is where $n = 100$ and $\lambda^2 = 1$. In this case about 1/3 of the samples will have least squares residuals positively skewed according to Simar and Wilson (2010). The distributional parameters obtained from the normal/doubly truncated normal model have relatively large mean squared errors. We presume that this is due to the fictitious "wrong" skewness that yields large variances of the estimates because the determinant of Fisher's information matrix is close to zero. The normal/doubly truncated normal model cannot provide a remedy in this case and the bagging technique proposed by Simar and Wilson (2010) could be employed to make an inference. As either the sample size or signal-to-noise ratio (λ) increases, the fictitious "wrong" skewness goes away, MSE decreases and the normal/half-normal model is recovered from the normal/doubly truncated normal model. It would appear from our intuition and from our simulations that in finite samples the large estimated standard errors of the distributional parameters can serve as an indicator of the presence of "fictitious" wrong skewness in the model.

Table 3.1. Monte Carlo results for for Half-Normal model. The number of repetitions $M = 1000$.

			$n = 100$		$n = 200$		$n = 1000$	
		True	AVE	MSE	AVE	MSE	AVE	MSE
	$\hat{\sigma}$	0.42	0.4307	0.0099	0.4359	0.0079	0.4250	0.0041
	$\hat{\gamma}$	0.5	0.5634	0.1332	0.5605	0.0025	0.5565	0.0382
	$\hat{\mu}$	0.0	-0.0071	0.3069	-0.0351	0.2801	0.0082	0.2003
$\lambda^2 = 1$	$\hat{\alpha}_0$	0.9	0.9799	0.1197	0.9728	0.0728	0.9757	0.0450
	$\hat{\alpha}_1$	0.6	0.5814	0.0160	0.5997	0.0034	0.6021	0.0012
	$\hat{\alpha}_2$	0.5	0.5105	0.0139	0.5043	0.0021	0.4958	0.0013
	$\hat{\sigma}$	0.31	0.3264	0.0076	0.3256	0.0021	0.3148	0.0011
	$\hat{\gamma}$	0.91	0.9404	0.0068	0.9161	0.0006	0.9107	0.0004
	$\hat{\mu}$	0.0	-0.0398	0.1277	-0.0566	0.0439	-0.0115	0.0175
$\lambda^2 = 10$	$\hat{\alpha}_0$	0.9	0.9144	0.0247	0.8982	0.0043	0.9001	0.0026
	$\hat{\alpha}_1$	0.6	0.6079	0.0043	0.6033	0.0006	0.6012	0.0003
	$\hat{\alpha}_2$	0.5	0.5015	0.0036	0.4973	0.0007	0.5005	0.0004
	$\hat{\sigma}$	0.30	0.3064	0.0049	0.3050	0.0013	0.3030	0.0006
	$\hat{\gamma}$	0.99	0.9958	0.0001	0.9912	0.0001	0.9905	0.0001
	$\hat{\mu}$	0.0	-0.0252	0.0559	-0.0092	0.0104	-0.005	0.0050
$\lambda^2 = 100$	$\hat{\alpha}_0$	0.9	0.9102	0.0098	0.8986	0.0014	0.8951	0.0006
	$\hat{\alpha}_1$	0.6	0.5995	0.0016	0.5993	0.0108	0.6003	0.0001
	$\hat{\alpha}_2$	0.5	0.4978	0.0012	0.5016	0.0125	0.5002	0.0001

To answer the second part of the question we consider the conditional mean inefficiencies conditional on the composed error in the same spirit as in Jondrow et al. (1982). For the normal/doubly truncated normal model these are given by

$$E(u_i|\hat{\varepsilon}_i) = \mu_* + \sigma_* \frac{\phi(-\frac{\mu_*}{\sigma_*}) - \phi(\frac{B-\mu_*}{\sigma_*})}{\Phi(\frac{B-\mu_*}{\sigma_*}) - \Phi(-\frac{\mu_*}{\sigma_*})},$$

where $\mu_* = \frac{\mu\sigma_v^2 - \varepsilon\sigma_u^2}{\sigma^2}$ and $\sigma_* = \frac{\sigma_u\sigma_v}{\sigma}$. Ignoring the bound will yield incorrect estimates of the inefficiencies scores.

3.5. Conclusions

Most of the distributions for inefficiencies considered in the stochastic frontier models literature are positively skewed. The doubly truncated normal inefficiency distribution generalizes the SFM in a way that allows for negative skewness as well. This implies that finding incorrect skewness does not necessarily indicate that the model is misspecified. A misspecification would arise, however, were the researcher to consider an incorrect distribution for the inefficiency process, which has a skewness that is not properly identified by the least squares residuals. The "wrong" skewness can be a finite sample artifact or a fact. Our study has considered the latter case and has shown that the normal/doubly truncated normal composed error SFM can still be valid with the "wrong" sign of the skewness statistic using a generalization of Waldman's (1982) proof. Moreover, "wrong" skewness in finite samples does not necessarily preclude its appearance in large samples under our specification. Our study thus provides a rationale for applied researchers to adopt an additional strategy in cases when this perceived empirical anomaly is found.

CHAPTER 4

Accounting for Heterogeneous Technologies in the Banking Industry: A time-varying Stochastic Frontier Model with Threshold Effects

4.1. Introduction

The U.S. commercial banking industry is characterized by its large number of heterogeneous institutions. Although the number of commercial banks currently operating in the banking sector has dropped to 6839 from the 14382 it was in 1984, the differences among banks are more profound. These differences could be largely attributed to the effects of deregulation and financial crises of the early 1980s and 1990s. Two acts have played an especially crucial role in forming the current landscape of the commercial banking industry: the Reigle-Neal Interstate Banking and Branching Efficiency Act, that was passed by Congress in 1994 and fully implemented in 1997, which allowed the interstate banking and branching; and the Financial Services Modernization Acts of 1999 that granted broad-based securities and insurance power to commercial banks. The first act allowed banks to geographically expand through acquisitions of other financial institutions and through opening new offices or branches within and outside of a particular state.

As a result, more than 6000 mergers occurred within less than two decades which created large national banks (mega banks) for the first time in the history of U.S. commercial banking. Certain banks had grown enormously in size and are characterized now as too-big-to fail by regulators as their failure would likely cause significant damage to the financial system and cause serious disruptions to the broader economy. Currently the asset sizes of the six largest mega banks correspond to 63% of the GDP, up from 17% in 1995. The second act had similar effects in leading commercial banks to grow in size by allowing them not only to enter insurance and securities companies' areas, but also directly to acquire these companies. In addition, technological innovations, such as the automated teller machine (ATM), credit card network and scoring, electronic payments, internet banking, and emergence of new financial instruments, such as mutual funds and derivatives led large banks to grow larger through various channels and increase the gap between these and smaller depository institutions. Figure 4.1 plots the size distributions (in natural logarithm scale) of commercial banks for four different quarters, which appear to be extremely skewed to the right with the skewness statistic increasing over time.

Heterogeneity could arise from different business opportunities that each financial institution faces, lending strategies, accessibility to the short-term money markets, risk exposures, expenditures on technology related innovations, and several other factors that are primarily associated with the size of these institutions.

Size of banks, which is typically measured by their total assets or deposits, is traditionally considered to affect the type of activities and the performance of banks. It is a standard practice adopted by regulators to analyze banks by splitting them into several size-categories. For example, Federal Deposit Insurance Corporation (FDIC) divides these financial institutions into four groups based on the market value of their total assets. The first group consists of banks with asset size under $100 million and the second group includes banks with asset size between $100 million and $1 billion. Banks from these two groups, which are usually characterized as small or community banks, mostly base their activities on retail and consumer banking and specialize in residential mortgages and individual loans. They have limited or no access to capital markets, such as federal funds market, and usually finance their activities with core deposits and/or equity. Their contribution is very important to the U.S. economy because of the personalized services they offer and their understanding of the communities they serve. The third group's assets range from $1 billion to $10 billion and the fourth group consists of very large banks with asset size that exceeds $10 billion.[1] Banks in groups three and four are considered large banks and typically engage in nontraditional banking and extend their activities both superregionally and nationally. Large banks have relatively easy access to purchased funds and the capital markets compared to the small banks, hold fewer core deposits, and are highly leveraged. Furthermore, these banks tend

[1]Often the fourth group is further divided to distinguish between top 10-25 banks and the rest banks in the group.

to hire expert personnel and pay higher salaries, as well as extensively engage in mergers and in investment in buildings and premises. The last two groups of banks are increasing both in number and importance, while the number of community banks and their assets as a percentage of the total industry assets are shrinking over time.

Based on our discussion so far it would not be appropriate to pool a highly heterogenous sample of banks into a single group. It also would be misleading to assign the banks to different groups based on arbitrary and ad hoc criteria when using statistical estimation and inference. Nevertheless, the majority of the received studies in the banking literature that estimate either the production or cost frontiers assume that banks are relatively homogeneous and have an access to the same best-practice technology and hence share the common frontier. When this assumption fails to hold it is more plausible to assume that the different types of banks employ different types of technology in their intermediation process. Mester (1994) and Wheelock and Wilson (2001) offer examples that divide the banks into classes based on the value of their asset size and treat the threshold values as known when estimating the technology parameters.

Asset size, while a common threshold variable in the banking literature to designate banks to certain groups, by no means exhausts the list of the variables that may be used to distinguish different types of banks. There are other criteria, such as riskiness, that can be used separately or along with bank's asset size to further segment the banking industry. These are typically employed in a cluster

analysis framework. Amel and Rhoades (1988, 1992), Brown and Glennon (2000), Tortosa-Ausina (2002), and Wang and Kumbhakar (2009) among others, use cluster analysis techniques to segment banks into distinct strategic groups in terms of their product mix and allocation of inputs. If asset size is considered as the only threshold variable, two questions still remain unanswered: how do we determine the appropriate cutoffs to split the banks in the sample? What does it mean for bank to be large or small? If the answer to these questions is not obvious then one needs to resort to data-driven methods, such as threshold effects, to deal with these particular issues.

These issues appear to be even more challenging and interesting in stochastic frontier literature, where recently researchers attempted to separate the firm-specific effects (differences) from the firm-specific efficiencies. Orea and Kumbhakar (2004), Greene (2005), El-Gamal and Inanoglu (2005) among others, employed the latent class specification to suggest heterogeneity in technology parameters and inefficiencies. Tsionas (2002), O'Donnel and Griffiths(2004), Tsionas and Kumbhakar (2004), and Huang (2004) employed the hierarchical Bayesian methods to address the same issue of heterogenous technologies in stochastic frontier models. Their findings pointed to the existence of considerable differences in technologies employed by firms and their efficiencies scores.

If bank differences are primary due to the heterogeneous technologies, then accounting for these while simultaneously estimating the bank-level efficiencies would be an appropriate solution to this problem. More specifically, our solution

to this problem involves applying the nondynamic panel threshold effects model of Hansen (1999, 2000a) to the sample of U.S. commercial banks modified to estimate time-varying inefficiencies in the spirit of Cornwell et al. (1990). In addition to the estimates of individual and group efficiency scores, we provide the estimates of returns to scale and measures of technological change. We show that pooling banks into a single class is clearly not justified by standard statistical techniques and produce estimates that have considerably different efficiency ranking than estimates based on the technology-specific effects model.

The remainder of the chapter is organized as follows. In section 2 the heterogeneity issue in stochastic frontier models and its importance is discussed. Section 3 outlines Hansen's (1999, 2000a) nondynamic panel model with threshold effects which is modified to account for time-varying individual effects. Section 3 lays out the baseline empirical model that describes the technology of U.S. commercial banking industry. Section 4 summarizes the estimation results from the threshold effects model for the panel data sample of banks for the period 1984-2009 and compares these to the full sample estimation results. Finally, section 5 concludes.

4.2. Heterogeneity in Stochastic Frontier Models

Parametric Stochastic Frontier Model (SFM) was first introduced by Aigner et al. (1977), Meeusen and van de Broek (1977), and Battese and Cora (1977) as a model that provides measures of the performance of individuals or firms. The technology in SFM is demonstrated through parametric functions, such as

Cobb-Douglas or translog functions, as opposed to the nonparametric alternative approaches of data envelopment analysis (DEA) proposed by Charnes et al. (1978) and the free disposal hull (FDH) of Deprins et al. (1984). The error term in SFM is assumed to be multiplicative and composed of two parts, a one-sided error term that captures the effects of inefficiencies relative to the stochastic frontier and a two-sided error term that captures random shocks, measurement error and other statistical noise, as well as allows for random variation of frontiers across firms.

The original model was developed in cross-sectional context and explicitly assumed that inefficiencies are independent from the regressors. This is a very strong assumption, the violation of which leads to inconsistent estimates of the model's parameters. Schmidt and Sickles (1984) introduced panel data model that provides consistent estimates by considering the inefficiencies as permanent fixed effects. That is, departing from the pure production frontier the model can be represented as:

$$\begin{aligned} y_{it} &= \alpha + X'_{it}\beta + \upsilon_{it} - u_i \\ &= \alpha_i + X'_{it}\beta + \upsilon_{it} \end{aligned}$$

where y_{it} is the output, X_{it} is a vector of inputs. υ_{it} is the noise component which is $iid\ N(0, \sigma_v^2)$, while $u_i > 0$ is the time invariant firm effect representing technical inefficiency and which may or may not be correlated with the regressors.

The time-invariance assumption of inefficiency term is also very restrictive and unreasonable for relatively long panels. Cornwell, Schmidt, and Sickles (1990) (CSS) instead formulated the individual effects as a time polynomial of degree two. That is

$$\alpha_{it} = \theta_{0i} + \theta_{1i}t + \theta_{2i}t^2$$

where $\theta's$ are firm-specific parameters. This quadratic specification allows technical efficiency to vary over time and across individual firms. If $\theta_{i2} = \theta_{i3} = 0$, then CSS model collapses to the fixed effects model of Schmidt and Sickles (1984). If we assume that the coefficients of time and the squared time are constant across firms then the model reduces to the fixed effects model with the linear and quadratic time term common to all producers, $\theta_2 t + \theta_3 t^2$. One interpretation of this restricted version of the model is that technical efficiency is producer-specific and varies through time in the same manner for all producers. An alternative interpretation is that technical efficiency is producer-specific and time-invariant, with the quadratic time term capturing the effects of technical change. It is not possible to distinguish between these two scenarios.[2] CSS describes several estimation strategies including the fixed-effects (within), random-effects (GLS), and the Hausman and Taylor (1981) efficient instrumental variable (IV) approach for their model. The fixed effects estimator does not assume independence of inefficiencies and the regressors. However, it does not allow for time-invariant regressors and requires $T \to \infty$ for consistent estimation of the effects. On the other hand, the random

[2]See Kumbhakar and Lovell (2000) for more discussion.

effects estimator is consistent as N goes to infinity provided that the effects are uncorrelated with the regressors, which is a testable assumption (Hausman-Wu test). For fixed T it is more efficient than the fixed effects estimator. The efficient IV provides the solution in case some regressors are correlated with the efficiencies. Inefficiencies in panel data model can also be treated as random effects drawn from some known distribution. They can be time-invariant (Pitt and Lee, 1981) or time-varying (Kumbhakar, 1990; Battese and Coeli, 1992).[3] Unfortunately, these random effects models do not allow for endogenous regressors.

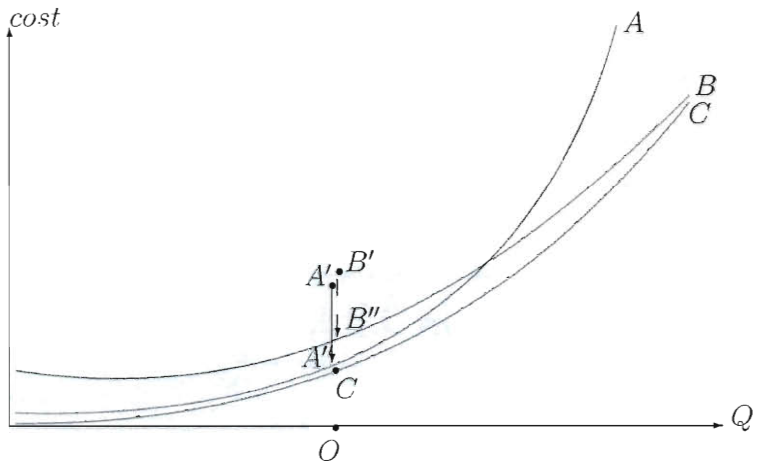
The common characteristic of all these models is that they do not account for the heterogeneity that might exist among individuals or firms. They assume common technology and/or inefficiency distribution parameters for all individuals and any unobserved differences are mixed with the inefficiency term. Greene (2007) defines two types of heterogeneity: the observable and the unobservable heterogeneity. The first type is controlled by considering some exogenous variables, such as ownership type, country of origin, etc., that can detect any differences among firms associated with these variables. These can be included in the kernel production or cost frontier and/or can enter the performance equation. The second type, however, cannot be easily detected and as such requires data-driven methods to

[3]Kumbhakar (1990) proposed $u_{it} = w(t)u_i$ specification for time-varying efficiency model, where $w(t) = 1/[1 + exp(\gamma t + \delta t^2)]$ and $u_i \sim^{iid} N^+(0, \sigma_u^2)$ and estimated γ and δ, along with the rest model parameters, by the maximum likelihood techniques. Whereas, Battese and Coeli (1992), on the other hand, define $w(t) = \exp(\eta(T-t))$ in their model. The latter model is more popular and widely employed in panel data SFMs due to the provision of the free software from the authors (Frontier 4.1 version).

account for it. In standard panel data models (with fully efficient units) u potentially represents these unobservable individual specific effects. These, however, cannot be incorporated and identified in a straightforward fashion in stochastic frontier models.

While estimating the efficiencies in SFM, why it is so important to account for heterogeneity/differences across firms? According to Tsionas (2002), a firm shouldn't be labeled as inefficient if it employs different technology (possibly old and inferior) than other firms in the industry do. It would stand to reason that it is not profitable for this firm to adopt a new technology, which of course does not come at no cost, and employing the old technology is the optimal for the time being. Hence, using a common frontier will pronounce this firm as inefficient although it may fully utilize its current technology. Ignoring the possibility of heterogenous technologies can also lead to erroneous efficiency ranking of firms and wrong conclusions about the measure of returns to scale at the individual and industry level. Consider an illustrated example (on the below graph) of two firms (or group of firms) A and B with cost frontiers labeled with the respective letters. The observed total costs of two firms are given by the point A' and B', respectively. If we assume common frontier (C) for these two firms, then the cost efficiency of firm A is given by the ratio of the minimum cost to the observed cost, $\text{CE}_A^{pooled} = \frac{OC}{OA'}$ and this of firm B by $\text{CE}_B^{pooled} = \frac{OC}{OB'}$. Under this assumption, we conclude that firm B is less efficient than firm A. However, if we consider each firm operating on its own frontier, then $\text{CE}_A^{ind} = \frac{OA''}{OA'}$ and $\text{CE}_B^{ind} = \frac{OB''}{OB'}$, where

ind stands for individual frontier. It is obvious that efficiency ranking of these two firms reverses since in this case firm A is pronounced to be less efficient than firm B.[4]



Appropriately defining outputs and inputs is still an elusive topic in the banking literature. The intermediation approach of Sealey and Lindley (1977), which is the standard in banking related research, considers the dollar volume of the banks' outputs and inputs instead of their physical units, which is obviously far from the ideal. Numbers and specific characteristic of loans are generally not available for the public use. As a result, as Berger and Mester (2002) also note, outputs produced by the larger banks can be substantially different from outputs produced by the smaller banks, requiring different monitoring and screening techniques. A

[4]Similar example can be found in El-Gamal and Inanoglu (2005).

large amount of the loan produced by a large bank and issued to a single borrower, for example, could have a significantly differential effect on its cost than if the same amount was lent in parts to smaller borrowers. Moreover, there is a significant gap in costs among banks with similar scale and product mix, which could be attributed to many factors (Berger and Humphrey, 1992). Two such factors are size and inefficiency. Ignoring the size effect, the full gap would be attributed solely to inefficiency and other uncontrollable factors. Thus, appropriately distinguishing between large and small banks could in part account for this difference by capturing the shifts in their production technologies.

How do we then account for both heterogeneity and inefficiencies in SFM? One way to do this is by introducing an individual specific intercept to the standard model.

$$y_{it} = \alpha_i + X'_{it}\beta + \upsilon_{it} - u_{it}$$

Notice that this requires that the inefficiency component is strictly time-varying in order to identify the individual effects. Recall also that in CSS model $\alpha_{it} = \theta_{0i} + \theta_{1i}t + \theta_{2i}t^2$, thus α_i cannot be separately identified from the θ_{0i}, unless we consider a specification for the inefficiency term without the intercept.

Greene (2005) considers the α_i to be either fixed parameter or random effects drawn from some known distribution. In the first case, he specifies what he calls the "true" fixed effects model and estimates it by the "brute force" techniques. The drawback of this model is that it induces the "incidental parameters" problem

and it also assumes that inefficiencies are uncorrelated with the regressors. In the second case, the "true" random effects model is specified which is estimated via simulated maximum likelihood (SML) method or quadrature. Both the effects and the inefficiencies are assumed to be uncorrelated with the regressors in this model.

Another way to model the individual heterogeneity in SFM is through the random parameters model in the spirit of Swamy and Tavlas (1995). This model assumes that both, the intercept term and the slope parameters are random in frontier function and can be estimated either by SML method (Greene, 2005) or Bayesian techniques (Tsionas, 2002; Huang, 2004). Orea and Kumbhakar (2004) and Greene (2005) instead estimate a latent class SFM, which assumes the existence of Q classes with each firm's membership in the specific class determined by its contribution to the log-likelihood function. Prior membership probabilities for each individual firm are specified, which may also be considered as functions of observed individual characteristics. El-Gamal and Inanoglu (2005) proposed a similar model which employes the estimation-classification method of El-Gamal and Grether (1995). By fixing the number of classes, the conventional likelihood ratio (LR) test is performed to investigate the parameter heterogeneity across classes in their model.

All the models described above assume that structural/technological parameters do not change over time. Firms are assigned to a particular class or adopt a specific technology that is not changed throughout the given sample. However, after observing the real world it can be seen that firms change their strategies and

production techniques in order to survive in a constantly changing and challenging environment. This at least suggests that firms need not to share the same technology parameters over time and need to be allowed to switch groups. This can be accomplished through the clustering analysis (Brown and Glennon, 2000; Tortosa-Ausina, 2002; Wang and Kumbhakar, 2009) which segments the industry based on certain strategic variables, such as output mix, etc. However, there is no consensus on the nature and the number of the strategic variables. We consider overcoming this limitation by employing structural break/threshold effects model.

4.3. The Threshold Effects Stochastic Frontier Model

Following Hansen (1999, 2000a), the CSS production frontier model with single threshold effects can be represented as

$$y_{it} = X'_{it} I(q_{it} \leq \gamma)\beta_1 + X'_{it} I(q_{it} > \gamma)\beta_2 + \upsilon_{it} - u_{it} \tag{4.1}$$

where $I(\bullet)$ denotes the indicator function, q_{it} is a continuous time-varying scalar representing the threshold variable, and γ is the threshold value that splits the sample into two technology-groups. The assumptions about the effect and noise term are maintained from previous section. The effects are assumed to be random which allows for identification of the intercept and time trend parameter (s) in the kernel regression function.

We can rewrite (4.1) in a more compact representation as

$$y_{it} = \beta' X_{it}(\gamma) + \upsilon_{it} - u_{it} \tag{4.2}$$

with

$$X_{it}(\gamma) = \begin{pmatrix} X_{it} I(q_{it} \leq \gamma) \\ X_{it} I(q_{it} > \gamma) \end{pmatrix} \tag{4.3}$$

where $X_{it}(\gamma) = \binom{X_{it} I(q_{it} \leq \gamma)}{X_{it} I(q_{it} > \gamma)}$ is the $NT \times (K+1)$ design matrix of regressors and $\beta = [\beta_1 : \beta_2]$ is a $(K+1) \times 1$ vector of structural parameters.

The above representation makes it clear that the model is capable of dividing the observations into two discrete class-regimes based on the threshold variable value. If we consider q_{it} as a proxy for firm size, then two regimes can represent two different technologies that each bank in the sample employs depending on its size. In other words, the main purpose of the threshold effects model is to investigate whether the independent variables have different impacts on different subgroups of the population. If the answer to the previous question is positive, then the estimates based on the full sample will be biased and will have little or no economic meaning.

Using the conventional stacking of the panel data models, the threshold effects stochastic frontier model (TSFM), can be rewritten in matrix notation as:

$$y = X(\gamma)\beta + Qu + \upsilon \tag{4.4}$$

where $Q_{NT\times 3N} = diag(W_i)$, $W_i = [1\ t\ t^2]$, and u is $3N \times 1$ iid random vector with zero mean and covariance matrix Δ.

Let $M_Q = I - Q(Q'Q)^{-1}Q'$ be the projection onto the null space of Q and $\Omega^{-1/2} = \frac{1}{\sigma}M_Q + F$ where

$$F = \frac{Q(Q'Q)^{-1/2}(Q'Q)^{-1/2}Q'}{\left[\sigma^2 I_{3N} + (Q'Q)^{1/2}\left(I_N \otimes \Delta\right)(Q'Q)^{1/2}\right]^{1/2}} \tag{4.5}$$

Then, while treating the value of the parameter γ as known,the generalized least squares (GLS) estimator of β is given by

$$\hat{\beta}(\gamma) = \left(\tilde{X}'(\gamma)\tilde{X}(\gamma)\right)^{-1}\tilde{X}'(\gamma)\tilde{y} \tag{4.6}$$

where $\tilde{y} = \Omega^{-1/2}Y$ and $\tilde{X}(\gamma) = \Omega^{-1/2}X(\gamma)$.

The estimation of γ parameters is performed by minimizing the concentrated sum of squared residuals using a grid search algorithm.[5] That is,

$$\hat{\gamma} = \arg\min_{\gamma} S(\gamma) \tag{4.7}$$

[5]The values of the threshold variable are sorted and the algorithm searches over all distinct values or certain quantiles of q. The value of q that minimizes the concentrated sum of squared errors is the solution of the optimization algorithm.

where

$$\begin{aligned} S(\gamma) &= \tilde{e}'(\gamma)\tilde{e}(\gamma) = (\tilde{Y} - \tilde{X}(\gamma)\hat{\beta}(\gamma))'(\tilde{y} - \tilde{X}(\gamma)\hat{\beta}(\gamma)) \\ &= (\tilde{Y} - \tilde{X}(\gamma)[\tilde{X}'(\gamma)\tilde{X}(\gamma)]^{-1}\tilde{X}'(\gamma)\tilde{Y})'(\tilde{Y} - \tilde{X}(\gamma)[\tilde{X}'(\gamma)\tilde{X}(\gamma)]^{-1}\tilde{X}'(\gamma)\tilde{Y}) \\ &= \tilde{Y}'\{I - \tilde{X}(\gamma)[\tilde{X}'(\gamma)\tilde{X}(\gamma)]^{-1}\tilde{X}'(\gamma)\}\tilde{Y} \end{aligned} \tag{4.8}$$

After the threshold parameter is estimated we can obtain the structural coefficients and the variance of within residuals from

$$\hat{\sigma}^2(\hat{\gamma}) = \frac{1}{N(T-1) - (K+1)} S(\hat{\gamma})$$

By regressing the residuals

$$e(\gamma) = Y - X(\gamma)\hat{\beta}(\gamma)$$

on Q, we can obtain the estimated individual effects $\hat{u}_{it} = \hat{u}_{0i} + \hat{u}_{1i}t + \hat{u}_{2i}t^2$. The group-specific efficiencies then are estimated as[6]

$$EFF_{it}(\gamma) = \exp(\hat{u}_{it} - \max_{1 \le j \le N} \hat{u}_{jt})$$

which is consistent for large N. This specification implies that in each period at least one producer produces on frontier and is 100% technically efficient.

[6] $EFF_{it} = \exp(\min_j \hat{u}_{jt} - \hat{u}_{it})$ for the cost frontier models.

The estimate of γ and consequently of $\beta(\gamma)$ are consistent. Chan (1993) and Hansen (1999) note that the estimate of the threshold can be treated as fixed in deriving the asymptotic distribution of $\beta(\gamma)$ in order to facilitate the inference. However, the distribution of γ is nonstandard which can complicate the inference. In particular, testing for the presence of the threshold becomes problematic, since γ is not identified under the null hypothesis of no threshold and conventional tests would have distributions that are also nonstandard (Davies' Problem, 1977). Hansen (1996, 1999) proposes a bootstrap method to simulate the asymptotic distribution of the classical LR test, which can be used in hypothesis testing. We use the same test to investigate the existence of heterogeneous technologies in our commercial banking application.

A model with a single threshold can be extended in a straightforward fashion to accommodate multiple thresholds. If we assume the presence of L thresholds then the model is represented as

$$\begin{aligned} y_{it} &= \beta_1' x_{it} I(q_{it} \leq \gamma_1) + \beta_2' x_{it} I(\gamma_1 < q_{it} \leq \gamma_2) + ... \\ +\beta_l' x_{it} I(\gamma_{l-1} &< q_{it} \leq \gamma_l) + \beta_{l+1}' x_{it} I(\gamma_l < q_{it}) - u_{it} + v_{it} \end{aligned}$$

with the restriction that $\gamma_1 < \gamma_2 < .... < \gamma_{l-1} < \gamma_l$ for $l = 1, 2, ..., L$.

Similarly to the single threshold model, the threshold parameters are estimated via a grid search algorithm on the threshold variable values and the rest model

parameters and their standard errors are consequently estimated. A slight complication, however, arises in estimating the threshold parameters in the multiple-threshold model. Joint estimation of the threshold parameters would require a grid search over an enormous number of points (even if we consider few quantiles of the q) which drastically increase with the number of break points. This calls for the sequential estimation of the threshold parameters which is consistent and is more than necessary especially when the sample size is large. However, this method yields asymptotically efficient estimates only of the last threshold parameter in the process. The previous estimates are contaminated by the presence of the neglected thresholds. Bai (1997) suggested a refinement estimation of the threshold parameters which amounts to re-estimating the threshold parameters backwards, each time holding the estimates of the previous thresholds fixed. The refinement estimator is shown to be asymptotically efficient.

4.4. Empirical Model and Data

The technology in the banking industry can be either demonstrated directly by the production frontier or indirectly by its dual cost function (Shephard, 1953). The cost frontier typically describes the minimum level of cost given a certain output level and input prices. The intermediation approach of Sealey and Lindley (1977) is the standard approach adopted in the banking literature according to which banks are viewed as intermediary multi-stage production units that collect loanable funds from depositors and investors to transform them into earning assets,

such as loans and securities (see for example Kaparakis et al., 1994, Berger and Mester,1997, 2002; Adams et al., 1999; Wheelock and Wilson, 2000, 2001; Kneip et al., 2011, among others). We adopt this approach in our empirical application. We abstain from including the cost share equations in the analysis due to the issues related to the allocative inefficiency ("Greene Problem"). Thus, the estimated overall efficiency represents the cost or economic inefficiency which can be due to technical or allocative inefficiency, or both.

Following Greene (2005), we specify five-output five-input stochastic cost frontier with flexible transcendental (translog) functional form[7]

$$
\begin{aligned}
\ln \frac{C_{it}}{w_{lit}} &= \alpha_{0+} \sum_{m=1}^{5} \alpha_m \ln y_{mit}(\gamma) + \sum_{\substack{k=1 \\ l \neq k}}^{4} \beta_k \ln \frac{w_{kit}}{w_{lit}}(\gamma) \\
&+ \frac{1}{2} \sum_{m=1}^{5} \sum_{j=1}^{5} \alpha_{mj} \ln y_{mit}(\gamma) \ln y_{jit}(\gamma) + \frac{1}{2} \sum_{\substack{k=1 \\ l \neq k}}^{5} \sum_{\substack{n=1 \\ l \neq n}}^{4} \beta_{kn} \ln \frac{w_{kit}}{w_{lit}}(\gamma) \ln \frac{w_{nit}}{w_{lit}}(\gamma) \\
&+ \sum_{m=1}^{5} \sum_{\substack{k=1 \\ l \neq k}}^{4} \delta_{mk} \ln y_{mit}(\gamma) \ln \frac{w_{kit}}{w_{lit}}(\gamma) + \eta \ln req_{it}(\gamma) + \xi \ln rmnpl_{it}(\gamma) \\
&+ \frac{1}{2} \zeta (\ln req_{it})^2 + \frac{1}{2} \psi (\ln rmnpl_{it})^2 + \theta_1 t + \frac{1}{2} \theta_2 t^2 \\
&+ \sum_{m=1}^{5} \lambda_{mt} \ln y_{mit}(\gamma) t + \sum_{\substack{k=1 \\ l \neq k}}^{4} \phi_{kt} \ln \frac{w_{kit}}{w_{lit}}(\gamma) t + \upsilon_{it} + u_{it}
\end{aligned}
$$

[7]Translog function provides the second-order Taylor series approximation to any arbitrary function at a single point. In addition, the returns to scale measures and factor demand elasticities are not required to be constant as in the Cobb-Douglas case.

where $C_{it} = \sum_k w_{kit} x_{kit}$ represents the observed total cost of the intermediation process for the individual bank in each time period t and y_m is the value of m^{th} output, $m = 1, .., 5$. Outputs are the real estate loans (y_1), commercial and industrial loans (y_2), loans to individuals (y_3), securities (y_4), and off-balance sheet items (y_5). $w's$ represent the inputs prices, which are interest-bearing deposits in total transaction accounts (x_1), interest-bearing deposits in total non-transaction accounts (x_2), labor (x_3), purchased funds (x_4), and capital (x_5). In addition, two variables are included to control for observable heterogeneity. The first variable is the ratio of the total equity to total assets (req) which typically measures the risk of insolvency of the bank. It reflects the ability of the bank to absorb the unexpected losses from their on- and off-balance sheet activities. Banks with lower values of this ratio are highly leveraged and thus should be considered riskier than banks with higher values, ceteris paribus. The second variable is the ratio of non-performing loans to total loans ($rmnpl$) which reflects the quality of loans made by the bank.[8] Quadratic time trend and time interaction with outputs and input prices are also included to account for non-neutral technological shifts of the cost frontier. The linear homogeneity in input prices restriction is imposed by normalizing the cost and the input prices by the price of capital. The symmetry restrictions ($\alpha_{mj} = \alpha_{jm}$ and $\beta_{kn} = \beta_{nk}$) are also imposed.

[8]Nonperfoming loans include the total loans and lease finance receivables that are nonacrual, past due 30-89 days and still accruing, and past due 90 days and still accruing.

Deposits can be determined as either inputs or outputs through the empirical test outlined in Huges and Mester (1993). That is, when quantity instead of the price of the deposits is included in the cost function and the increased quantity leads to the lower cost ($\ln C/\ln x < 0$), then the particular type of deposits is considered as an input. Should the deposits be considered as an output, then the opposite should hold. Increasing the output requires additional usage of inputs (none if the bank operates inefficiently), which subsequently should increase the cost. In the intermediation approach deposits are viewed as inputs, as opposed to the production and value-added approaches (Baltensperger, 1980; Berger and Humphrey, 1992) which treat the deposits as the bank output services. In the current application both types of deposits are determined as inputs by the above simple empirical test.

After the structural parameters are obtained we can estimate the class-specific scale economies and measures of technological change. The scale economies are defined as the degree to which a bank's total cost of producing financial services decreases as its output services increase proportionally and it is derived as the sum of the partial derivatives of the cost with respect to the outputs. That is,

$$\begin{aligned} Scale_{it}(\gamma) &= \sum_{m=1}^{5} \frac{\partial C_{it}(\gamma)}{\partial \ln y_{mit}(\gamma)} \\ &= \sum_{m=1}^{5} \left[\alpha_m + \sum_{j=1}^{5} \alpha_{mj} \ln y_{jit}(\gamma) + \sum_{k=1}^{5} \delta_{mk} \ln w_{kit}(\gamma) + \lambda_{mt} t \right] \end{aligned} \tag{4.9}$$

If this measure is less than one then there is a presence of economies of scale indicating that the bank is operating below the optimal scale level and can reduce the cost by expanding its output. On the other hand, if it is greater than one then the bank is experiencing diseconomies of scale and should reduce its output level to achieve optimal input combination. It is worthwhile to note that in the case of the Cobb-Douglas cost frontier specification the economies of scale is equal to $\sum_{m=1}^{5} \alpha_m$ and is common for all banks in the sample. In this special case threshold effects estimation can provide measures that are common for banks within the technology-group, but different across groups.

The data are extracted from quarterly Consolidated Reports of Condition and Income (Call Reports) for all U.S. commercial banks that are collected and administrated by the Federal Reserve Bank of Chicago and Federal Deposit Insurance Corporation (FDIC). The observed sample period is from 1984 to 2009 (third quarter). The initial unbalanced sample is comprised of 5,253 (2009.Q3) to 12,781 (1984.Q1) banks observed in each quarter with total of 861,420 observations, after dropping banks with zero costs, zero output and input levels, as well as those with obvious measurement errors and other data inconsistencies. This eliminated approximately 18.5% of observations of the entire population of all insured commercial banks. Finally, we randomly sample 2,500 banks with a total of 257,500 observations from this non-homogeneous pool of banks, which also constitutes the

estimation sample.[9] Summary statistics of this sample are reported in table 4.2 for four selected quarters.[10]

4.5. Empirical Results

The application of the threshold stochastic frontier model (TSFM) to the sample of commercial banks revealed the existence of six thresholds/cutoffs, which is translated into seven distinct technology-groups.[11] There was not enough evidence in the sample to suggest against the null hypothesis of six thresholds over the alternative hypothesis of seven thresholds based on bootstrap probability value. The estimated threshold parameters along with the group-specific average estimates of the cost efficiency (CE), returns to scale (RTS), and technical change (TC) are reported in table 4.1. Figures 4.3-4.5 plot the these estimates over time for each group, as well as for the full sample. The same table summarizes the accounting ratios, such as return on assets (ROA), return on equity (ROE), profit margin (PM), and asset utilization (AU), separately for each group.[12] These ratios are typically

[9]The reason for utilizing a balanced sample is to capture a stable technological behavior of the banks as they grow in size throughout the entire sample period and to film their switching the size-categories over time. This eliminates very small banks (those that failed or were acquired by other surviving banks) and the de novo banks (state member banks that have been in operation for five years or less), which could introduce a serious technological disruptions and bias the results. Another reason is that the threshold effects estimation method is computationally intensive, as it takes very long time even for computers with superior computing power which utilize multiple nodes.

[10]1984.Q1 (beginning of the sample), 1993. Q1 (prior to the introduction of the Reigle-Neal Act in 1994), 2000.Q1 (reference period), and 2009.Q3 (end of the sample).

[11]To determine the number of thresholds, a grid search was performed over 250 quantiles of the threshold variable in each of the 29 sequential estimation steps.

[12]ROA is defined as the ratio of the net income to the total assets and measures the profit earned relative to the bank's assets. ROE is defined as the ratio of the net income to total equity capital

Table 4.1. Estimation Results: Threshold Values, Cost Efficiency, Returns to Scale, Technical Change, Return on Assets, Return on Equity, Profit Margin, and Asset Utilization.

group	*CE*	*RTS*	*TC*	*ROA*	*ROE*	*PM*	*AU*
group 1	0.7683	1.486	4.5%	0.31%	2.6%	-2.7%	5.0%
$(0 < q \leq 19,876)$							
group 2	0.7680	1.516	3.6%	0.53%	4.9%	9.4%	5.0%
$(19,876 < q \leq 39,876)$							
group 3	0.7641	1.338	3.0%	0.61%	7.1%	12.4%	5.1%
$(39,876 < q \leq 110,178)$							
group 4	0.7960	1.211	3.6%	0.64%	5.1%	13.1%	5.1%
$(110,178 < q \leq 197,430)$							
group 5	0.8065	1.128	0.8%	0.63%	7.1%	13.0%	5.2%
$(197,430 < q \leq 361,946)$							
group 6	0.8184	1.120	6.2%	0.62%	5.1%	12.4%	5.2%
$(361,946 < q \leq 1,300,592)$							
group 7	0.7792	1.038	3.9%	0.58%	15.4%	11.5%	5.2%
$(q > 1,300,592)$							
full sample	0.7673	1.182	1.3%	0.57%	6.2%	11.1%	5.1%

used to evaluate the performance and profitability of financial institutions by managers, investors, and regulators. Parameter estimates from the translog stochastic cost frontier for each group are reported in table 4.3.

and measures the overall profitability of the bank per dollar of equity. PM is defined as the ratio of the net income to the total operating income and measures the bank's ability to pay expenses and generate net income from interest and non-interest income. AU is defined as the ratio of the total operating income to the total assets and measures the amount of the interest and non-interest income generated per dollar of the total assets.

In particular, we find that banks in groups 1,2, and 3, which are very small banks, appear to be less efficient compared to the banks in the other four groups. The difference is small, but statistically significant. On average, banks in these three groups are considerably scale inefficient for all sample periods and group 1 displays an upward efficiency trend for the most of the quarters. The number of the small banks is falling over time according to figure 4.2. On the other hand, group 6 is the most cost efficient group on average. However, the performance of this group is rapidly declining since 2000. Banks in this group have enough room to exploit their scale efficiencies. It is worthwhile to note, that pooling banks into a single class underestimates the efficiencies of banks in this group and reports them as less efficient than those in group 7. One possible explanation of the higher cost efficiencies of group 6 is that its member banks are able to adopt new technologies very quickly due to their manageable size. This also could be one of the reasons the regulatory authorities seek to place specific upper bounds on the size of very large banks ("too-big-to-fail" banks) to make them smaller, safer, and more manageable. Even a slight difference in the efficiencies of these large institutions could be translated in billions of dollars losses.

Figure 4.4 shows that the large banks in group 7 have already exhausted their potential scale economies. The hypothesis of constant returns to scale cannot be rejected for these banks after 1996. Previous studies in banking literature have shown that U.S. commercial banks operate at constant returns to scale at much lower output levels (McAllister and McManus, 1993; Wheelock and Wilson,

2001). In their recent study of the U.S. commercial banking industry, Wang and Kumbhakar (2009) found that a large proportion of very small banks, with asset sizes less than 25 millions, face decreasing returns to scale. Results of the current work fail to support this finding.

Overall the efficiency trends for all groups are consistent with those found in Almanidis et al. (2010). The efficiencies are increasing after the deregulation of the 1980s. Except for group 6, the efficiency levels are relatively stable for the period 1990-2005 and fall thereafter, possibly revealing the weaknesses of the banking industry and the seeds of the current financial distress. Similar patterns occur, although not to that extent, during previous recessionary periods of 1990 and 2001. In sum, we could informally conclude that the declining efficiencies of the U.S. commercial banking industry could serve as a predictor and indicator of a financial crisis.

Technological innovation has always been and still is the concern of all financial institutions. However, acquiring and adopting new technologies is not the same for banks of different sizes. Our results indicate that the small banks display technological progress which is decreasing over time. The cost frontier of the large banks in groups 6 and 7, on the other hand, is expanding at an increasing rate due to the high technology that these banks can afford. It is widely believed that small banks adopt new technologies with considerable lags, but at a lower cost. The technical change measure is constant at around 2% if the estimation is done without considering the possibility of the threshold effects.

Finally, according to the accounting ratios, on average larger banks appear to perform better than smaller banks. Simply looking at ROA suggests that, banks in group 7 are less profitable than banks in groups 3 and 4. However, they produce two to six times more profit per dollar of equity than banks in any other group. This is because they tend to hold less equity than other banks do, as they have relatively easy access to money and capital markets. Moreover, their profit margin, i.e., the ability to generate net income from interest and noninterest income, is low compared to this for banks in groups 3 to 6. The ability to generate noninterest income, however, is twice as high for these large banks (1.1% of total assets) compared to smaller banks (0.56% of total assets), because the former banks tend to engage in off-balance sheet activities more than their smaller peers are able to do.

4.6. Conclusions

In this chapter we investigated the existence of heterogeneous technologies in the U.S. commercial banking industry. we applied the threshold effects estimation technique with an exogenous threshold variable (total asset size) and determined seven distinct technology-groups. Pooling banks into a single class was clearly not justified by the result of the bootstrap test and produced distorted estimates and different efficiency ranking than estimates based on the technology-specific effects model. In addition, we provided estimates of individual and group efficiency scores,

as well as of those of returns to scale and measures of technological change. The average efficiencies were found to be time-varying whose level and slopes differed across groups. All groups displayed a consistent sharp decline in their average efficiencies during the financial crisis that was fired up in August of 2007. Results also have shown that the very large banks have already exploited their scale efficiencies and display technological progress which improves over time.

Table 4.2. Summary statistics for selected periods

variables	1984.$Q1$	1993.$Q1$	2000.$Q1$	2009.$Q3$
y_1	27.4 (251.6)	84.2 (720.5)	473.6 (6500)	695.2 (10000)
y_2	38.8 (642.7)	49.4 (630.8)	165.2 (2399)	221.7 (3466)
y_3	19.5 (149.8)	31.9 (366.5)	103.7 (1715)	134.9 (2099)
y_4	40.3 (178.6)	87.7 (779.2)	258.6 (4274)	377.4 (6850)
y_5	20.7 (318.6)	61.7 (996.1)	316.3 (6633)	2337 (71900)
w_1	0.011 (0.18)	0.006 (0.07)	0.006 (0.06)	0.003 (0.005)
w_2	0.018 (0.03)	0.009 (0.02)	0.002 (0.001)	0.007 (0.003)
w_3	8.80 (4.380)	9.25 (1.939)	10.7 (2.3010)	33.0 (7.1095)
w_4	0.023 (0.09)	0.012 (0.01)	0.033 (0.025)	0.075 (0.038)
w_5	0.090 (0.07)	0.091 (0.07)	0.080 (0.059)	0.215 (0.141)
req	0.091 (0.03)	0.102 (0.03)	0.104 (0.032)	0.108 (0.031)
rmnpl	0.026 (0.03)	0.011 (0.01)	0.024 (0.021)	0.038 (0.030)

Standard deviation in parenteses. Outputs and the price of the labor are expressed in thousands of U.S. dollars

y_1=real estate loans

y_2=commercial and industrial loans

y_3 =loans to individuals

y_4=securities

y_5=off-balance sheet items

w_1= average price of interest-bearing deposits in total transaction accounts

w_2=average price of interest-bearing deposits in total non-transaction accounts

w_3= average price of labor

w_4=average price of purchased funds

w_5=average price of capital

req= the ratio of the equity to total assets

rnmpl=the ratio of non-performing loans to total loans

Table 4.3. Technology-Group Estimation Results

	g1	g2	g3	g4	g5	g6	g7
y_1	-0.104**	-0.114***	0.063*	0.256***	0.335***	0.570***	0.496***
y_2	-0.088***	0.128***	0.232***	0.282***	0.517***	0.387***	-0.028
y_3	-0.008	-0.179***	-0.028	0.130***	0.115***	-0.220**	0.071**
y_4	-0.327***	-0.449***	-0.262***	0.203***	0.292***	0.781***	0.313***
y_5	-0.089***	-0.131***	-0.044*	-0.057**	-0.143***	-0.207***	0.076**
w_1	0.171***	0.116***	0.223***	0.316***	0.436***	0.336***	0.138***
w_2	0.618***	0.870***	0.873***	0.830***	0.471***	0.715***	0.646***
w_3	0.181*	-0.058	-0.052	0.123*	-0.094*	-1.407*	0.141*
w_4	-0.098***	0.037*	-0.006	-0.040*	0.131***	0.453**	0.130***
w_5	0.129*	0.035	-0.038	-0.228**	0.056	0.902	-0.056
w_1w_1	0.053***	0.052***	0.053***	0.044***	0.016***	0.018**	0.002
w_1w_2	-0.047***	-0.048***	-0.028***	-0.012***	0.006**	0.006	-0.006*
w_1w_3	-0.019***	-0.031***	-0.045***	-0.059***	-0.034***	-0.061**	-0.021***
w_1w_4	0.020***	0.031***	0.028***	0.035***	0.023***	0.023***	0.016***
w_1w_5	-0.006*	-0.004*	-0.007*	-0.008**	-0.010*	0.014	0.009
w_2w_2	0.153***	0.162***	0.119***	0.117***	0.093***	0.103***	0.088***
w_2w_3	-0.049***	-0.055***	-0.043***	-0.047***	-0.070***	-0.153***	-0.072***
w_2w_4	-0.046***	-0.053***	-0.040***	-0.048***	-0.022***	-0.002	0.005
w_2w_5	-0.106***	-0.114***	-0.073***	-0.081***	-0.088***	-0.071*	-0.077***
w_3w_3	0.056***	0.084***	0.061***	0.066***	0.070***	0.407**	0.053***
w_3w_4	0.006*	-0.001	0.007***	0.013***	0.015***	-0.034	0.022***
w_3w_5	-0.103***	-0.108***	-0.058***	-0.047***	-0.012*	0.088	-0.011
w_4w_4	0.013***	0.0190***	0.004**	0.001	-0.018***	-0.020***	-0.052***
w_4w_5	-0.077***	-0.086***	-0.063***	-0.058***	-0.024***	0.029	-0.037***
w_5w_5	-0.113***	-0.139***	-0.096***	-0.088***	-0.071***	-0.355*	-0.074**
w_1y_1	-0.008***	-0.004***	-0.005***	-0.004***	0.006***	0.004	-0.031***
w_1y_2	-0.002**	0.002***	0.002***	0.003**	0.003*	0.015***	0.036***
w_1y_3	0.018***	0.016***	0.013***	0.004***	-0.006***	-0.015***	0.009***
w_1y_4	0.004**	0.010***	0.010***	0.015***	-0.002	0.020***	0.013***

p*<0.1, p**<0.05, p***<0.01

Table 3: Cont'd

	g1	g2	g3	g4	g5	g6	g7
w_1y_5	-0.001*	-0.002***	-0.001**	0.001	-0.004***	-0.010***	-0.014***
w_2y_1	0.036***	0.011***	0.003	-0.001	0.007**	-0.016*	0.045***
w_2y_2	-0.011***	-0.010***	-0.014***	-0.005*	-0.009***	-0.003	-0.038***
w_2y_3	-0.012**	-0.024***	-0.016***	-0.015***	0.008***	-0.005	-0.004*
w_2y_4	0.023***	0.024***	0.015***	0.011***	0.026***	0.037***	0.010**
w_2y_5	-0.016***	-0.012***	-0.007***	-0.010***	-0.009***	0.011*	-0.011***
w_3y_1	-0.020***	-0.001	0.012***	0.007*	-0.005	0.088***	-0.031***
w_3y_2	0.013***	-0.004*	0.008***	-0.003	-0.006*	-0.073***	-0.014**
w_3y_3	-0.017***	0.005	-0.007***	-0.001	-0.006**	0.032	0.017***
w_3y_4	-0.016***	-0.022***	-0.020***	-0.032***	-0.009**	-0.074**	0.010*
w_3y_5	0.015***	0.012***	0.006***	0.011***	0.021***	0.011*	0.001
w_4y_1	-0.004**	-0.005***	0.001	-0.002	0.001	-0.001	0.016***
w_4y_2	0.002*	0.008***	0.005***	0.003***	0.011***	0.013**	0.010***
w_4y_3	0.010***	0.008***	0.008***	0.007***	0.006***	0.002	-0.006***
w_4y_4	0.002	-0.009***	-0.008***	0.001	-0.009***	-0.010*	-0.022***
w_4y_5	-0.001	0.002***	-0.002**	-0.002***	-0.011***	-0.013***	0.002
w_5y_1	-0.004	-0.001	-0.011**	-0.001	-0.009*	-0.075**	0.001
w_5y_2	-0.002	0.003	-0.001	0.003	0.001	0.048*	0.006
w_5y_3	0.001	-0.001	0.002	0.005	-0.002	-0.014	-0.016**
w_5y_4	-0.013*	-0.004	0.003	0.005	-0.006	0.026	-0.011
w_5y_5	0.003	-0.003	0.004*	0.001	0.003	0.001	0.023***
y_1y_1	0.063***	0.055***	0.079***	0.118***	0.131***	0.081***	0.153***
y_1y_2	-0.009***	-0.009***	-0.011***	-0.029***	-0.028***	-0.002	-0.037***
y_1y_3	0.006**	0.003*	-0.008***	-0.016***	-0.030***	-0.034***	-0.018***
y_1y_4	-0.014***	-0.018***	-0.049***	-0.081***	-0.087***	-0.113***	-0.104***
y_1y_5	0.001	0.002**	0.002	0.003*	0.005***	0.011*	0.001
y_2y_2	0.031***	0.035***	0.029***	0.043***	0.043***	0.055***	0.068***
y_2y_3	0.005***	0.004***	-0.002*	-0.002	-0.007***	-0.001	-0.003
y_2y_4	-0.007***	-0.025***	-0.030***	-0.022***	-0.032***	-0.042***	-0.016***

p*<0.1, p**<0.05, p***<0.01

Table 3: Cont'd

	g1	g2	g3	g4	g5	g6	g7
y_2y_5	-0.003**	-0.001*	0.001	0.002	-0.004***	0.001	0.010***
y_3y_3	0.041***	0.042***	0.048***	0.043***	0.042***	0.035***	0.028***
y_3y_4	-0.025***	-0.019***	-0.018***	-0.025***	-0.003	0.011*	-0.003
y_3y_5	0.005***	0.007***	0.005***	0.002*	0.003**	0.009***	-0.002
y_4y_4	0.122*	0.142	0.153***	0.148*	0.127	0.127	0.118
y_4y_5	-0.003*	0.001	-0.005***	-0.003*	0.004	-0.003	-0.002
y_5y_5	0.002*	-0.001	0.003***	-0.005	0.002*	-0.001	-0.009***
req	-1.72***	-1.970***	-1.604***	-1.168***	-1.033***	-2.624***	0.009
$rmnpl$	2.707*	3.431***	2.139***	0.522*	0.509	8.839***	-1.206
req^2	-0.092*	0.016	0.001	0.027	-0.245***	-0.699***	-1.213***
$rmnpl^2$	0.671*	0.157	-0.414*	0.512	-0.066	3.914***	5.938***
ty_1	0.002***	0.001***	0.001***	-0.001	0.001***	-0.001	0.002***
ty_2	-0.001	-0.003***	-0.001***	-0.002*	-0.004***	0.001*	-0.001***
ty_3	0.001**	0.003***	0.001*	0.001*	0.001***	-0.002	-0.005***
ty_4	0.001	0.001	0.001***	-0.001	-0.003*	0.002***	-0.002
ty_5	-0.001***	-0.003***	-0.002***	-0.005***	-0.001***	-0.001	-0.001***
tw_1	0.005***	0.001***	0.001***	0.001***	-0.004***	-0.003	0.005**
tw_2	0.002	0.001***	0.004**	0.002***	0.003***	0.006***	0.003***
tw_3	-0.001**	-0.001***	-0.001***	-0.002***	-0.001***	-0.009***	-0.001*
tw_4	0.005***	0.001	0.003***	-0.001***	-0.001***	-0.001	-0.002***
tw_5	-0.005	-0.005*	-0.004	-0.003*	0.001	0.004	-0.006
t	-0.032***	-0.015***	-0.009***	0.003	0.006**	0.005	-0.004
$\frac{1}{2}t^2$	0.002***	0.001***	0.001***	0.001***	-0.001	0.002***	0.0003*
C	8.981***	11.004***	8.044***	2.862***	1.343**	3.039*	1.834***

p*<0.1, p**<0.05, p***<0.01

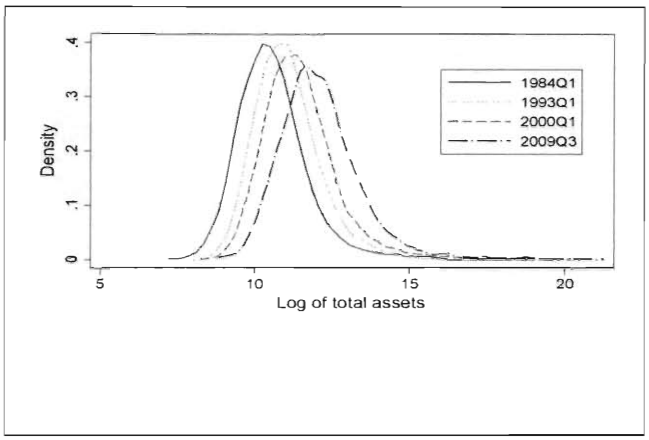


Figure 4.1. Distribution of log of Total Assets

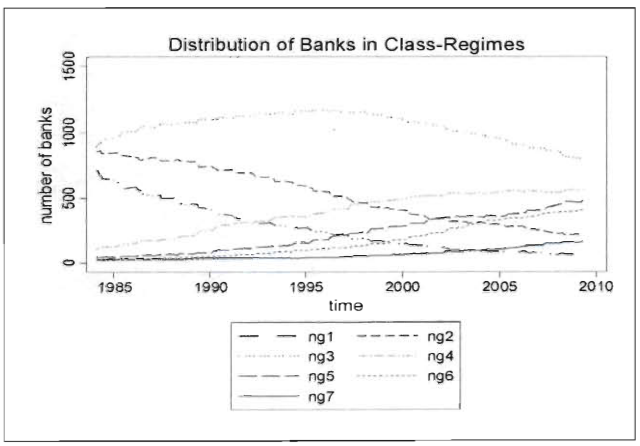


Figure 4.2. Distribution of banks in groups over time

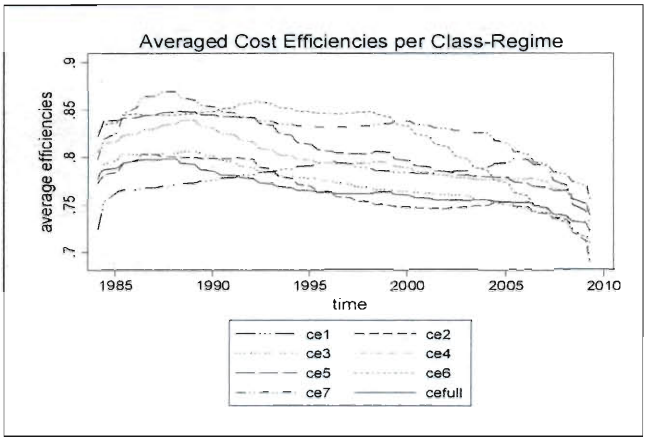


Figure 4.3. Averaged cost efficiencies for seven groups and the pooled sample

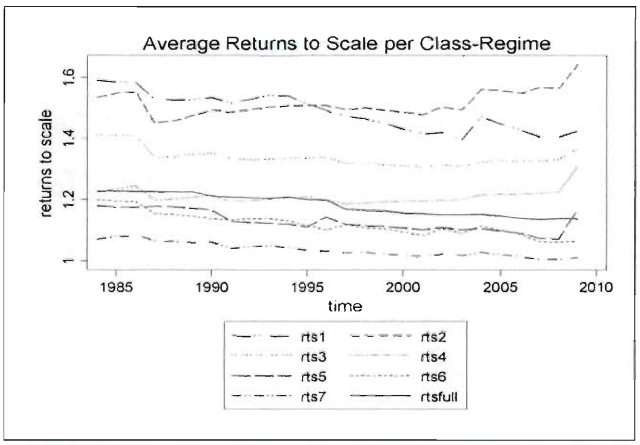


Figure 4.4. Average returns to scale measure for seven groups and the pooled sample

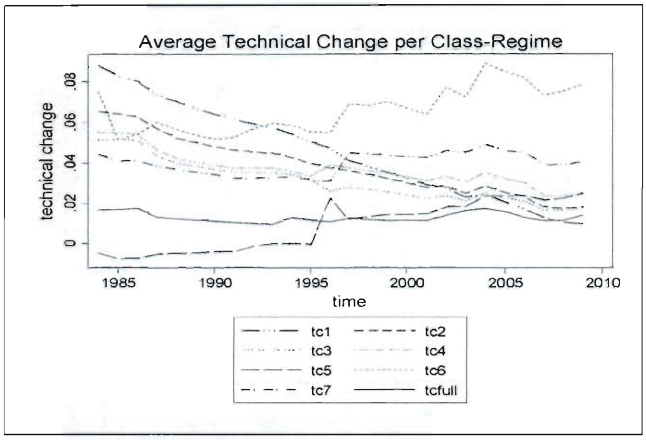


Figure 4.5. Average technical change for seven groups and the pooled sample

CHAPTER 5

Commercial Banking Data

The data on commercial banks used in this thesis are primarily extracted from the quarterly consolidated reports of condition (balance sheets) and income (income statements), also known as the Call Reports, that are collected and administrated by the Federal Deposit Insurance Corporation (FDIC) and are collectively available from the Federal Reserve Bank of Chicago's website.[1] Every national bank, state member bank, and insured state nonmember bank is required to file either a FFIEC 031 (Consolidated Reports of Condition and Income for a Bank with Domestic and Foreign Offices) or a FFIEC 041 (Consolidated Reports of Condition and Income for a Bank with Domestic Offices Only) form on a quarterly basis as of the last calendar day of March, June, September, and December.[2] Call Report contains detailed data on bank's various earning and non-earning assets, liability composition, capital structure, income, expenses, and other bank-specific structural and geographical characteristics. Office/branch-level data are available from the FDIC only on the amount of deposits, which are reported in Summary of Deposits each year (on June 30th) since 1994. Below, we provide a somewhat

[1] www.chicagofed.org

[2] Initially, only a subset of banks were required to file two of the four Call Reports on surprise dates, or "on call" and this is how its name derived.

detailed description of the main banks' on-balance and off-balance sheet elements, income and expenses, as well as of the regulatory capital and environmental variables. A particular care is provided to creating consistent time series of the relevant data, as the content and structure of the reports are frequently revised in light of global environmental and regulatory changes. Finally, following Kashyap and Stein (1995), Adams et al. (1999), Jayasuriya (2000), and Berger and Mester (2002), we detail the construction of the key variables used in this thesis. These mainly include output quantities, input prices and quantities, characteristics of banks and the regulatory environment in which they operate, as well as a number of measures of risk, asset quality, profitability, and performance. A comprehensive discussion of commercial banking industry and its functions can be found in Saunders and Cornett (2004).

We merge quarterly files obtained from the Federal Reserve Bank of Chicago, which contain all the variables reported on the Call Reports and the structure and geographical variables, between 1984 (first quarter)-2010 (second quarter).[3] This provided 1,057,545 observations after dropping reporting banks with non-positive total assets, total deposits, and total loans. All dollar values are converted to reflect 2000 (first quarter) prices using the consumer price index (CPI).[4] The majority of

[3]The files are available starting in 1976. However, the definitions of certain variables have considerably changed after 1984, because of the break in reporting forms, which makes it difficult to merge the files consistently from these two periods (see Kashyap and Stein, 1994, 1995 and "Notes on forming consistent time series" available at the Federal Reserve Bank of Chicago's website for more discussion on these issues).

[4]Quarterly series on CPI are available from the website of the Federal Reserve Bank of St. Louis (FRED) at http://research.stlouisfed.org

banks in this sample are very small banks (below 90th percentile by asset size). As it is shown in figure 5.1 the number of these banks is declining over time, while the medium and large-sized banks grow both in numbers and importance. As of current analysis, approximately 21.5% of the commercial banks are nationally chartered and 34% of banks hold the Federal Reserve membership.

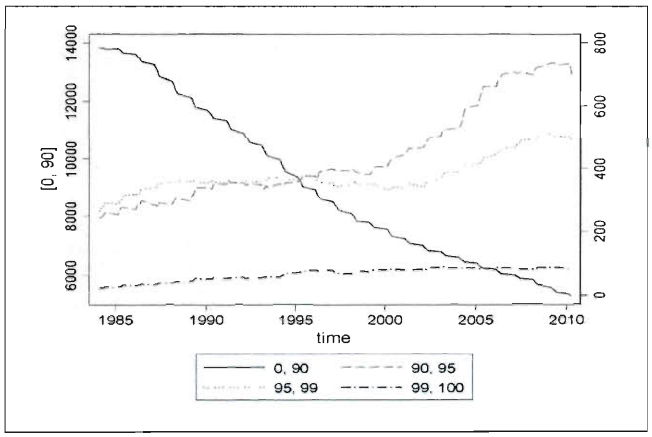


Figure 5.1. Number of banks in asset size percentiles

5.1. Balance Sheet Data

A balance sheet of a commercial bank, like of any other type of a business company, consists of two parts: total assets and total liability plus capital. Table 5.1 illustrates the balance sheet of a typical commercial bank and provides the Call Report item definition for each of its components. The structure and the composition of the assets and liabilities, however, is very different in case of financial institutions, which also varies considerably for institutions of different asset sizes.

Tables (5.3)-(5.10) report the representative balance sheets for four different size categories over four different periods.[5]

Total assets consist of cash, securities, total loans and leases, and other tangible and intangible assets. Cash involves noninterest-bearing balances, currency and coin held by the bank primarily to perform its daily activities and to meet withdrawals. Due to the recent advances in online and internet banking systems, as well as the broad usage of credit cards, banks tend to hold fewer cash (as a percentage of their total assets). Securities consist of items such as interest-bearing deposits due from depository institutions (DIs), federal funds sold and all securities purchased under agreements to resell, and investment securities (book value). Investment securities include U.S. Treasury securities, U.S. government agency and corporation obligations, securities issued by states and political subdivisions in the U.S., and other domestic and foreign debt and equity securities. Unlike the loans, securities are highly liquid, bear low default risk, and can be conventionally traded in the secondary markets. The majority of the smaller banks tend to hold significant amounts of securities, because of their inability to easily access short term money markets, such as the federal funds market. As a result of the recent liquidity crisis, banks (particularly large ones) have significantly increased their pools of investment securities to be able to meet unexpected liquidity needs.

[5]A bank is assigned to size categories whether its total asset falls into the following percentile ranges: [0,90], (90,95], (95,99], and (99,100]. The periods are: 1984 (first quarter), which is the beginning of the sample period, 1993 (first quarter) a year before Congress passed the interstate banking and branching law, 2000 (first quarter) which is the reference period, and 2010 (second quarter) which is the last period of observations.

Total loans fall into seven broad categories: real estate loans, commercial and industrial loans (C&I), agricultural or farm loans, lease financing receivables, loans to individuals, loans to depository institutions, and other loans. Real estate loans consist of loans to individuals and businesses that are secured by real estate. This type of loans constitute the largest component of the total loans for small banks. C&I loans include loans for commercial and industrial purposes to business enterprises (proprietorships, partnerships, and corporations), whether secured or unsecured, single payment or installment. Larger banks traditionally concentrate on wholesale banking and thus make fewer real estate loans and more C&I loans than smaller banks do. Recently large banks have substantially reduced the amount of the loans to corporations (as a percent of total assets) and as many banks in the industry were highly involved in the residential mortgage lending. Agricultural loans, which are almost exclusively provided by banks in the first category, include all loans issued to farmers for the purpose of financing agricultural production. Physical assets, such as buildings and vehicles, rented to a customer constitute bank's leases. These constitute a negligible percentage of bank's total assets and the vast bulk of industry leases are provided by large banks. Loans to individuals involve credit card, auto, student, and other miscellaneous personal loans. Larger banks consistently increase their share in the market for individual loans, especially the credit cards markets, over the smaller banks' share that traditionally used to have the advantage of providing personally based services to their clients. Loans to depository institutions and other loans include all loans to other banks

and depository institutions, nonbank associations and companies, as well as to state, local, and foreign governments. These type of loans loose their popularity in banks' loan compositions through time. Loans and leases are typically reported in dollar values, as their numbers are not available prior to 1993 (second quarter). Together with securities, loans and leases constitute bank's earning assets and are major components of bank's total assets. Net loans and leases are derived after deducting the unearned income on loans and the reserve for loan and lease loses from the total (gross) loans. Unearned income is the amount of income on loans earned but not yet reported on the bank's income statement. Whereas, the reserve for loan and lease losses reflect the management's assessment of the value of the defaulted assets in subsequent periods.

Other assets on banks' balance sheet consist of premises and fixed assets, other real estate owned, intangible assets and other miscellaneous assets. Premises and fixed assets, which include equipment, furniture, fixtures, and capitalized leases, are typically considered as a proxy for bank's capital. Other real estate owned primarily consists of real estate acquired for collateral and previously contracted debts. Intangible assets involve items such as goodwill, trademarks, etc. Finally, other miscellaneous assets represent bank assets that cannot be properly included in any of the preceding items. Other assets constitute a small fraction of banks' total assets.

On the liability and equity capital side the items are even more diverse. Although all banks are financed with deposits, purchased (borrowed) funds, and common equity, the structure and the compositions of these sources varies significantly across financial institutions. The primary source of funds, in particular for smaller banks, are the total deposits which can be classified as core and noncore deposits. Core deposits consist of demand deposits, negotiable order of withdrawal accounts (NOW accounts), money market deposit accounts (MMDAs), and retail certificate of deposits (CDs). Demand deposits include all noninterest-bearing transaction (checking) accounts. NOW accounts are interest-bearing checking accounts, which may or may not require prior notice before withdrawal. Large banks in the last two size-categories tend to keep fewer demand deposits and NOW accounts as opposed to their smaller peers. MMDAs, which lately is considered as an attractive source of funds by banks of all sizes, refer to interest-bearing checkable deposits with certain restrictions imposed on the minimum balance, number and the denomination of checks. Retail CDs include all nontransaction time certificates of deposit and open account time deposits with balances of less than $100,000, regardless of negotiability or transferability. Smaller banks heavily rely on retail CDs, which on average constitute about 70 percent of their total assets.

Purchased funds, as a second substantive source of funds, consist of wholesale CDs, federal funds purchased and all securities sold under agreements to repurchase, other borrowed money and notes issued to the U.S. Treasury, brokered

deposits, and subordinated notes and debentures. Wholesale CDs includes all outstanding time deposits of $100,000 or more and can be conventionally traded in the secondary market to meet the depositor's liquidity or other needs. Wholesale CDs and the core deposits form bank's total deposits and play equally significant role in financing banks assets.[6] Small banks in the first group have very limited access to capital markets and virtually do not participate in federal funds market. In contrast, larger banks have easier access to money markets and tend to have fewer core deposits. Similarly, small banks have limited ability to issue demand notes to the U.S. Treasury and to be involved in other type of borrowings, such as discount window borrowing provided from Federal Reserve bank. Finally, brokered deposits, which are wholesale CDs from broker agencies, as well as the subordinated notes and debentures, which typically have long duration and bear relatively low withdrawal risk, make up only a small fraction of banks' total liability and equity capital compared to the other components of purchased funds. Other liabilities, on the other hand, consist of owed by the bank funds that do not require interest payment.

The third source of funds is the total equity capital, which consists of common and preferred stocks, undivided profits, surplus, and capital reserve. The equity

[6]Alternatively, total deposits can be divided into transaction deposits (RCON2215), such as demand deposits and NOW accounts, and non-transaction deposits (RCON2385), such as retail or household savings and time deposits. Also, total deposits can be derived as the sum of interest-bearing deposits (RCON6636) and non-interest-bearing deposits (RCON2215).

capital is the most expensive source of funding, which drives larger banks make heavy use of purchased funds and hold less equity capital compared to small banks.

Not all of the banks' activities are reported on their balance sheet. Banks, especially recently, are highly engaged in fee generated activities off their balance sheet to stay in the business, as the interest income generated from balance sheet items has a declining trend. Off-balance sheet activities of a typically commercial bank involve derivative contracts (futures and forwards, swaps, and options), loan commitments, letters of credit, and investment-structured vehicles. In addition to generating income to financial institutions, derivative contracts are used to hedge interest rate, credit, and foreign exchange risks exposures.

5.2. Income Statement Data

Income statement for a particular financial institution reports the total operating income (interest income and non-interest income) gained and the expenses (interest and non-interest) paid for its on and off balance sheet items and activities. The difference between these two, after deducting the provision for loan and losses, taxes, and any extraordinary items, constitutes the net income for the bank. Table 5.11 summarizes the income statements for the representative bank in each size category through the time.

The interest income includes interest and fee income on bank's loans and leases, as well as interest received from securities, while the noninterest income is comprised of items such as income from fiduciary activities, service charges on deposit

accounts in domestic offices, trading gains (losses) and fees from foreign exchange transactions, other foreign transaction gains (losses), gains (losses) and fees from assets held in trading accounts, and other noninterest income from off-balance sheet activities. The latter income-generating source constitutes the substantial portion of larger banks' income and it becomes increasingly important as banks' ability of generating income from traditional lending activities is somehow limited. The interest expense consists mainly of interest on core deposits and interest on purchased funds. The non-interest expense, which is generally large relative to interest expense and non-interest income, includes the sum of salaries and employee benefits, expenses of premises and fixed assets, and other miscellaneous noninterest expenses that are not required to be reported by the bank.

Prices for output and input services are not directly reported in the Call Reports. By dividing the quarterly interest expenses and non-interest expenses of each subcategory item by its dollar amount, we can construct the respective prices similar to that used in the current thesis.[7] Similar prices for outputs and services are difficult to obtain, because the income data generally include interest income and fees which will contaminate the pure proxies. Table 5.2 details the construction of these prices from Call Report items.

[7]Prices are imputed under the assumption of uniform pricing, i.e., a bank pays the same price for its inputs in each of its operating markets. This assumption is somehow restrictive for banks with superregional activities. For example, the average wage paid by such bank to its employees in offices/branches located in different states or metropolitan statistical areas (MSA) are more likely to differ substantially. As Adams et al. (2007) state, this method will lead to some measurement errors which could be avoided if more disaggregated market-specific data used to proxy these prices.

Table 5.1. Balance Sheet Data

Assets	Call Report	Liabilities&Capital	Call Report
Total Assets	RCFD2170	Total deposits	RCFD2200
Cash	RCFD0080	Demand deposits	RCON2210
Interest-bearing balances	RCFD0070	NOW accounts	RCON2398
due from DIs		MMDAs	RCON6810
Federal Funds sold	RCFD1350	Retail CDs	RCON6648
and Repos			
Investment securities	RCFD0390		
Gross Loans and Leases	RCFD1400	Wholesale CDs	RCON2604
Real estate loans	RCFD1410	Federal Funds	RCFD2800
C&I loans	RCFD1766	purchased and Repos	
Agricultural loans	RCFD1590		
Loans to Individuals	RCFD1975	Other borrowed	RCFD2835
Lease financing receivables	RCFD2165	money and	
Loans to DIs	RCFD1489	U.S. Treasury notes	
Other loans	RCFD2080		
		Brokered deposits,	RCON2365
Net Loans and Leases	RCFD2122	subordinated notes	+RCON3200
Uil	RCFD2123	and debentures	
Rlll	RCFD3123		
		Other liabilities	RCFD2930
Premises and fixed assets	RCFD2145		
Other real estate owned	RCFD2150		
Intangible assets and other	RCFD2143	Equity Capital	RCFD3210
	+RCFD2160		

Uil=unearned income on loans

Rlll=reserve for loan and lease losses.

Table 5.2. Average Input Prices

Total interest-bearing deposits in total transaction accounts	Interest expense on transaction accounts / Total transaction accounts	RIAD4508 / RCON2215
Total interest-bearing deposits in total non-transaction accounts	Interest expense on non-transaction accounts / Total non-transaction accounts	RIAD4511* / RCON0352
NOW accounts	Interest expense on transaction accounts** / NOW accounts	RIAD4508 / RCFD2398
MMDAs	Interest expense on MMDAs / MMDAs	RIAD4509 / RCON6810
Retail CDs	Interest expense on Retail CDs / Retail CDs	RIADA518*** / RCON6648
Wholesale CDs	Interest expense on wholesale CDs / wholesale CDs	RIADA517**** / RCON2604
Federal funds purchased and Repos	Interest on federal funds purchased and Repos / U.S. Treasury notes and other borrowed money	RIAD4180 / RCFD2800
Notes issued to U.S. Treasury and other borrowed money	Interest on trading liabilities***** / U.S. Treasury notes and other borrowed money	RIAD4185 / RCON2835
Subordinated notes and debentures	Interest on subordinated notes and debentures / subordinated notes and debentures	RIAD4200 / RCON3200
Capital	Expenses of premises and fixed assets / Capital	RIAD4217 / RCFD2145
Labor******	Salaries and employee benefits / Labor	RIAD4135 / RIAD4150

*Alternatively can be defined as (RIAD4509+RIAD4511)/(RCON6810+RCON0352)
**Also includes interest on ATS accounts, and telephone and preauthorized transfer accounts
*** Prior to 1997Q1 the interest expense on retail CDs is derived as:
RIAD4170-RIAD4508-RIAD4511-RIADA517-RIAD4172
**** Prior to 1997Q1 the interest expense on wholesale CDs is given by RIAD4174 item
***** Includes Demand notes issued to the U.S. Treasury and other borrowed money
****** Labor is defined as the number of full time equivalent employees on the payroll of the bank and its consolidated subsidiaries at the end of the report period.

Table 5.3. 1984Q1 Balance Sheet-Asset Side data (in millions of U.S. dollars)

Asset size percentile	[0,90]	(90,95]	(95,99]	(99,100]
Cash	2.9 (5.5)	42 (5.7)	182 (5.8)	222 (5.7)
Securities				
-Interest-bearing balances due from DIs	1.4 (2.5)	25 (3.4)	209 (5.7)	369 (9.1)
-Fed funds sold and Repos	2.6 (6.0)	35 (4.7)	115 (4.2)	1,011 (2.5)
-Investment securities	15 (30)	140 (19)	404 (15)	2,258 (6.4)
Gross Loans and Leases	28 (52)	431 (58)	1,680 (57)	25,400 (63)
-Real estate loans	10 (18)	137 (18)	420 (15)	4,185 (12)
-C&I loans	7.5 (13)	95 (13)	484 (16)	1,210 (27)
-Agricultural loans	2.1 (7.5)	4.5 (0.6)	12 (0.4)	260 (0.6)
-Loans to individuals	6.4 (12)	83 (11)	24 (0.8)	2,003 (5.1)
-Leases	0.1 (0.1)	3.3 (0.4)	21 (0.6)	445 (1.1)
-Loans to DIs and other loans	1.6 (2.5)	36 (6.2)	244 (9.1)	5,541 (15)
Net Loans and Leases	27 (51)	423 (57)	1,665 (57)	25,200 (63)
-Uil	0.49 (1.0)	6.9 (0.9)	15 (0.5)	190 (0.5)
-Rlll	0.24 (0.5)	3.9 (0.5)	17 (0.5)	266 (0.7)
Other Assets				
-Premises and fixed assets	1.0 (2.0)	11 (1.5)	37 (1.2)	434 (1.1)
-Other real estate owned	0.17 (0.3)	1.6 (0.2)	4.9 (0.2)	61 (0.2)
-Intangible assets and other	0.87 (2.1)	12 (1.7)	53 (1.8)	1,061 (2.5)
Total Assets				
Mean	60	740	2,959	39,300
Median	31	693	2,233	24,300

Source: Author's calculations from the 1984.Q1 FDIC Call Reports. Percentage of total assets in parentheses.

Uil=unearned income on loans

Rlll=reserve for loan and lease losses.

Table 5.4. 1984Q1 Balance Sheet-Liability and Equity Capital Side data (in millions of U.S. dollars)

Asset size percentile	[0,90]	(90,95]	(95,99]	(99,100]
Total Deposits	45 (88)	558 (75)	1,942 (69)	14,500 (43)
Core Deposits				
-Demand deposits	8.2 (15)	106 (14)	412 (14)	4,237 (11)
-NOW accounts	NA	NA	NA	NA
-MMDAs*	6.5 (11)	70 (9.4)	239 (8.1)	2,493 (6.3)
-Retail CDs	24 (58)	271 (29)	921 (22)	6,694 (19)
Purchased Funds				
-Wholesale CDs	5.8 (10)	108 (14)	505 (18)	3,765 (12)
-Fed funds purchased and Repos	1.1 (1.1)	57 (7.7)	350 (11)	3,131 (8.6)
-Other borrowed money and notes issued to the U.S. Treasury	0.2 (0.2)	31 (4.4)	138 (4.9)	2,105 (5.6)
-Brokered deposits, subordinated notes and debentures	0.12 (0.2)	1.2 (0.2)	45 (1.4)	251 (1.1)
Other liabilities	0.64 (1.2)	8.1 (1.1)	44 (1.4)	1,420 (3.1)
Equity Capital	4.1 (9.2)	43 (5.8)	135 (4.5)	1,747 (4.3)
Number of Banks	13852	243	264	23

Source: Author's calculations from the 1984.Q1 FDIC Call Reports. Percentage of total assets in parentheses.

*MMDAs are available beginning 1984.Q3.

Table 5.5. 1993Q1 Balance Sheet-Asset Side data (in millions of U.S. dollars)

Asset size percentile	[0,90]	(90,95]	(95,99]	(99,100]
Cash	3.0 (4.3)	30 (4.0)	139 (3.8)	145 (4.6)
Securities				
-Interest-bearing balances due from DIs	0.7 (1.4)	5.2 (0.7)	40 (1.0)	1,297 (4.0)
-Fed funds sold and Repos	3.1 (5.1)	33 (4.4)	144 (3.9)	1,388 (5.7)
-Investment securities	24 (34)	162 (22)	684 (20)	4,277 (15)
Gross Loans and Leases	40 (52)	440 (59)	1,994 (56)	16,900 (52)
-Real estate loans	23 (28)	227 (30)	799 (24)	5,633 (18)
-C&I loans	7.0 (8.7)	101 (13)	556 (14)	5,876 (19)
-Agricultural loans	2.2 (5.5)	3.4 (0.4)	11 (0.3)	63 (0.2)
-Loans to individuals	6.8 (8.7)	74 (10)	41 (1.1)	2,059 (5.9)
-Leases	0.07 (0.1)	2.6 (0.4)	33 (0.8)	368 (1.1)
-Loans to DIs and other loans	0.78 (0.8)	11 (2.0)	92 (3.1)	2,317 (7.0)
Net Loans and Leases	39 (52)	438 (59)	1,989 (55)	16,870 (52)
-Uil	0.20 (0.3)	1.9 (0.2)	4.9 (0.1)	53 (0.1)
-Rlll	0.53 (0.9)	8.2 (1.1)	46 (1.2)	491 (1.3)
Other Assets				
-Premises and fixed assets	1.2 (1.6)	9.7 (1.3)	37 (1.4)	433 (1.2)
-Other real estate owned	0.44 (0.5)	5.0 (0.7)	20 (0.6)	243 (0.5)
-Intangible assets and other	0.87 (1.6)	12 (1.6)	70 (1.9)	1,484 (3.1)
Total Assets				
Mean	73	742	3,564	31,130
Median	49	709	2,622	19,800

Source: Author's calculations from the 1993.Q1 FDIC Call Reports. Percentage of total assets in parentheses.

Uil=unearned income on loans.

Rlll=reserve for loans and lease losses.

Table 5.6. 1993Q1 Balance Sheet-Liability and Equity Capital Side data (in millions of U.S. dollars)

Asset size percentile	[0,90]	(90,95]	(95,99]	(99,100]
Total Deposits	64 (88)	517 (72)	2,018 (59)	13,500 (46)
Core Deposits				
-Demand deposits	9.2 (12)	83 (11)	383 (10)	3,485 (12)
-NOW accounts	9.3 (13)	62 (8.4)	225 (6.5)	1,298 (4.7)
-MMDAs	9.0 (11)	89 (12)	390 (11)	3,215 (10)
-Retail CDs	45 (66)	421 (51)	1,736 (43)	9,600 (28)
Purchased Funds				
-Wholesale CDs	5.7 (7.6)	61 (8.0)	382 (10)	2,076 (8.6)
-Fed funds purchased and Repos	1.1 (0.9)	37 (4.7)	348 (8)	4,907 (10)
-Other borrowed money and notes issued to the U.S. Treasury	0.550 (0.5)	42 (6.0)	238 (7.4)	2,341 (6.3)
-Brokered deposits subordinated notes and debentures	0.23 (0.3)	3.8 (0.7)	65 (2.0)	291 (1.2)
Other liabilities	0.58 (0.7)	7.6 (1.0)	55 (1.4)	1,474 (2.7)
Equity Capital	6.6 (9.6)	51 (6.8)	222 (6.4)	1,874 (5.5)
Number of Banks	10475	346	357	55

Source: Author's calculations from the 1993.Q1 FDIC Call Reports. Percentage of total assets in parentheses.

Table 5.7. 2000Q1 Balance Sheet-Asset Side data (in millions of U.S. dollars)

Asset size percentile	[0,90]	(90,95]	(95,99]	(99,100]
Cash	3.4 (3.7)	22 (3.0)	99 (2.8)	165 (3.0)
Securities				
-Interest-bearing balances due from DIs	0.6 (1.1)	2.7 (0.4)	18 (0.5)	892 (1.3)
-Fed funds sold and Repos	3.1 (4.5)	22 (3.0)	146 (3.9)	2,674 (5.7)
-Investment securities	19 (20)	138 (19)	569 (17)	6,592 (14)
Gross Loans and Leases	59 (60)	459 (63)	2,099 (60)	27,200 (57)
-Real estate loans	36 (35)	286 (40)	978 (31)	9,595 (19)
-C&I loans	10 (11)	88 (12)	586 (15)	9,099 (18)
-Agricultural loans	3.3 (5.7)	5.7 (0.8)	14 (0.4)	110 (0.2)
-Loans to individuals	7.5 (8.1)	55 (7.6)	348 (9.1)	3,783 (11)
-Leases	0.14 (0.2)	4.4 (0.6)	39 (0.8)	1,445 (2.6)
-Loans to DIs and other loans	0.76 (0.7)	8.1 (1.4)	45 (1.7)	2,853 (3.5)
Net Loans and Leases	58 (60)	458 (63)	2,097 (60)	27,200 (57)
-Uil	0.08 (0.1)	0.59 (0.1)	2.0 (0.1)	18 (0.1)
-Rlll	0.57 (0.8)	6.0 (0.8)	32 (0.9)	431 (0.9)
Other Assets				
-Premises and fixed assets	1.8 (1.9)	10 (1.5)	34 (1.1)	475 (0.8)
-Other real estate owned	0.08 (0.1)	0.61 (0.1)	1.7 (0.1)	18 (0.1)
-Intangible assets and other	1.2 (1.8)	14 (1.9)	95 (2.4)	1,969 (3.3)
Total Assets				
Mean	93	725	3,574	36,700
Median	66	665	2,524	21,300

Source: Author's calculations from the 2000.Q1 FDIC Call Reports. Percentage of total assets in parentheses.

Uil=unearned income on loans.

Rlll=reserve for loan and lease losses.

Table 5.8. 2000Q1 Balance Sheet-Liability and Equity Capital Side data (in millions of U.S. dollars)

Asset size percentile	[0,90]	(90,95]	(95,99]	(99,100]
Total Deposits	78 (84)	5114 (71)	1,850 (56.9)	19,600 (41)
Core Deposits				
-Demand deposits	12 (13)	64 (8.8)	238 (6.8)	3,434 (6.9)
-NOW accounts	9.9 (11)	29 (4.2)	61 (2.0)	436 (1.0)
-MMDAs	11 (10)	114 (16)	467 (13.9)	6,242 (12)
-Retail CDs	56 (62)	431 (54)	1,794 (46)	17,700 (35)
Purchased Funds				
-Wholesale CDs	11 (12)	84 (12)	559 (15)	5,477 (17)
-Fed funds purchased and Repos	1.9 (1.4)	35 (4.6)	335 (8.0)	4,907 (10)
-Other borrowed money and notes issued to the U.S. Treasury	2.9 (2.4)	48 (6.5)	347 (8.9)	4,408 (11)
-Brokered deposits, subordinated notes and debentures	0.72 (0.6)	10 (1.7)	60 (2.4)	861 (3.3)
Other liabilities	0.75 (0.8)	8.5 (1.2)	61 (1.6)	1,367 (2.2)
Equity Capital	8.9 (11)	60 (8.3)	255 (7.5)	3,394 (6.4)
Number of Banks	7578	405	331	79

Source: Author's calculations from the 2000.Q1 FDIC Call Reports. Percentage of total assets in parentheses.

Table 5.9. 2010Q2 Balance Sheet-Asset Side data (in millions of U.S. dollars)

Asset size percentile	[0,90]	(90,95]	(95,99]	(99,100]
Cash	3.5 (3.2)	5.6 (3.1)	68 (2.8)	1,765 (2.9)
Securities				
-Interest-bearing balances due from DIs	7.0 (5.6)	30 (5.3)	124 (5.4)	7,795 (6.1)
-Fed funds sold and Repos*	NA	NA	NA	NA
-Investment securities	24 (18)	45 (18)	363 (17)	19,900 (17)
Gross Loans and Leases	88 (62)	386 (68)	1,443 (67)	59,800 (61)
-Real estate loans	64 (44)	299 (52)	1,043 (49)	21,800 (36)
-C&I loans	12 (8.9)	53 (9.2)	236 (10)	10,300 (11)
-Agricultural loans	5.0 (5.4)	9.4 (1.7)	18 (0.8)	185 (0.3)
-Loans to individuals	4.9 (4.1)	16 (2.8)	100 (3.7)	13,400 (11)
-Leases	0.20 (0.2)	1.2 (0.2)	9.0 (0.4)	1,108 (1.0)
-Loans to DIs and other loans	0.75 (0.5)	6.8 (1.2)	34 (1.4)	3,746 (2.4)
Net Loans and Leases	87 (62)	386 (67)	1,443 (66)	59,700 (61)
-Uil	0.02 (0.02)	0.19 (0.03)	0.61 (0.03)	17 (0.02)
-Rlll	0.90 (1.1)	7.7 (1.3)	35 (1.5)	2,386 (2.1)
Other Assets				
-Premises and fixed assets	2.6 (1.9)	10 (1.8)	33 (1.7)	835 (1.0)
-Other real estate owned	1.3 (0.8)	6.0 (1.1)	19 (1.0)	272 (0.3)
-Intangible assets and other	2.1 (2.5)	19 (3.2)	92 (3.8)	7,155 (5.3)
Total Assets				
Mean	134	727	3,328	110,000
Median	110	673	2,232	32,000

Source: Author's calculations from the 2010.Q2 FDIC Call Reports. Percentage of total assets in parentheses.

Uil=unearned income on loans.

Rlll=reserve for loan and lease losses.

* Beginning 3/31/2002, this item is no longer reported on the FFIEC 031 and 041 reports.

Table 5.10. 2010Q2 Balance Sheet-Liability and Equity Capital Side data (in millions of U.S. dollars)

Asset size percentile	[0,90]	(90,95]	(95,99]	(99,100]
Total Deposits	113 (84)	469 (82)	1,659 (79)	58,000 (64)
Core Deposits				
-Demand deposits	15 (12)	40 (7.0)	134 (6.2)	5,276 (4.9)
-NOW accounts	NA	NA	NA	NA
-MMDAs	19 (12)	105 (18)	534 (24)	30,900 (29)
-Retail CDs	88 (61)	409 (65)	1,425 (63)	50,700 (49)
Purchased Funds				
-Wholesale CDs	25 (17)	99 (17)	312 (16)	5,504 (7.4)
-Fed funds purchased and Repos	NA	NA	NA	NA
-Other borrowed money and notes issued to the U.S. Treasury	NA	NA	NA	NA
-Brokered deposits, subordinated notes and debentures	4.6 (2.8)	29 (5.2)	116 (5.5)	3,851 (6.0)
Other liabilities	0.83 (0.6)	4.1 (0.7)	23 (0.9)	3,252 (1.9)
Equity Capital	14 (11)	57 (9.9)	225 (9.9)	12,900 (11)
Number of Banks	5314	701	494	85

Source: Author's calculations from the 2010.Q2 FDIC Call Reports. Percentage of total assets in parentheses.

Table 5.11. Income Statement data (in millions of U.S. dollars)

Asset size percentile	[0,90]	(90,95]	(95,99]	(99,100]
1984.Q1				
-Interest income	1.320	14.058	58.508	860.008
-Non-interst income	0.099	1.580	8.003	95.410
-Interest expence	0.818	8.763	38.291	640.929
-Non-interest expence	0.410	4.947	20.610	214.048
Net Income	0.124	1.245	4.551	41.438
1993.Q1				
-Interest income	1.301	9.929	47.482	461.023
-Non-interst income	0.193	1.797	15.537	159.853
-Interest expence	0.513	3.692	17.768	230.900
-Non-interest expence	0.637	5.087	28.557	277.372
Net Income	0.235	1.7556	8.841	77.608
2000.Q1				
-Interest income	1.733	10.439	45.663	796.133
-Non-interst income	0.224	2.780	16.676	346.582
-Interest expence	0.772	4.689	21.230	418.034
-Non-interest expence	0.745	5.307	23.497	422.212
Net Income	0.273	1.931	9.484	157.617
2010.Q2				
-Interest income	3.218	13.181	47.245	2,143.966
-Non-interst income	0.509	2.409	11.567	1,111.319
-Interest expence	0.858	3.609	11.950	375.627
-Non-interest expence	2.109	8.351	31.245	1,573.607
Net Income	0.270	1.077	1.255	357.292

Source: Author's calculations from the quarterly FDIC Call Reports.

References

[1] Adams, R.M., Berger, A.N., and Sickles, R.C., 1999, Semiparametric approaches to stochastic panel frontiers with applications in the banking industry. Journal of Business and Economic Statistics 17, 349-358.

[2] Adams, R.M., Brevoort, K.P., and Kiser, E.K., 2007, Who Competes with Whom? The Case of Depository Institutions. Journal of Industrial Economics 55, 141-167.

[3] Afriat, S. N., 1972, Efficiency estimation of production functions. International Economic Review 13, 568-598.

[4] Aigner, D.J., and Chu, S., 1968, On Estimating the Industry Production Function. American Economic Review 58, 826–839.

[5] Aigner, D.J., Lovell , C.A.K., and Schmidt, P., 1977, Formulation and estimation of stochastic frontier models. Journal of Econometrics 6, 21-37.

[6] Alchian, A.A., 1950, Uncertainty, evolution, and economic theory. Journal of Political Economy 58, 211-21.

[7] Almanidis, P., 2010, Banking Crises, Early Warning Models, and Efficiency. Working Paper.

[8] Almanidis, P., and Sickles, R.C., 2009, Skewness problem in Stochastic Frontier Models: Fact or Fiction? Invited submission to Exploring Research Frontiers in Contemporary Statistics and Econometrics: A Festschrift in Honor of Leopold Simar, Ingrid Van Keilegom and Paul Wilson (eds.), Springer Publishing: New York, 2010.

[9] Almanidis, P., Qian, J., and Sickles, R.C., 2010, Bounded Stochastic Frontiers with an Application to the US Banking Industry: 1984-2009. Working paper.

[10] Altman, E.I., 1968, Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. Journal of Finance 23, 589–609.

[11] Amel, D.F., and Rhoades, S.A., 1988, Strategic groups in banking. Review of Economics and Statistics 70, 685–699.

[12] Amel, D.F., and Rhoades, S.A., 1992, The performance effects of strategic groups in banking. Antitrust Bulletin 37, no. 1: 171–86.

[13] Bai, J., 1997, Estimating multiple breaks one at a time. Econometric Theory 13, 315-352.

[14] Baltensperger, E., 1980, Alternative Approaches to the Theory of the Banking Firm. Journal of Monetary Economics 6, 1-37.

[15] Barr, R.S., and Siems, T.F., 1994, Predicting Bank Failure Using DEA to Quantify Management Quality. Financial Industry Studies 1-94, Federal Reserve Bank of Dallas.

[16] Battese, G. E., and Coelli, T.J., 1988, Prediction of Firm-level Technical Efficiencies with a Generalized Frontier Production Function and Panel Data. Journal of Econometrics 38, 387-399.

[17] Battese, G.E., and Coelli, T.J., 1992, Frontier production functions, technical efficiency and panel data, with application to paddy farmers in India. Journal of Productivity Analysis 3, 153-169.

[18] Battese, G.E., and Cora, G.S., 1977, Estimation of a production frontier model: with application to the pastoral zone of eastern Australia. Australian Journal of Agricultural Economics 21, 169-179.

[19] Bera, A.K., and Premaratne, G., 2001, Adjusting the tests for skewness and kurtosis for distributional misspecifications. UIUC-CBA Research Working Paper No. 01-0116.

[20] Berger, A. N. and Humphrey, D.B., 1992, Measurement and Efficiency Issues in Commercial Banking, in Output Measurement in the Service Sectors, edited by Z. Griliches, vol. 56, National Bureau of Economic Research, University of Chicago Press, Chicago.

[21] Berger, A.N. and Mester, L.J., 1997, Inside the black box: What explains differences in the efficiencies of financial institutions? Journal of Banking and Finance 21, 895-947.

[22] Berger, A.N. and Mester, L.J., 2002 Explaining the dramatic changes in performance of U.S. banks: Technological change , deregulation, and dynamic changes in competition. Working paper No. 01-06/R.

[23] Bover, O., Arellano, M., and Bentolila, S., 2002, Unemployment Duration, Benefit Duration and the Business Cycle. Economic Journal, Royal Economic Society 112 (479), 223-265.

[24] Breslow, N.E., 1972, Contribution to the discussion of paper by D.R. Cox (1972). Journal of the Royal Statistical Society, Series B 34, 216-217.

[25] Breslow, N.E., 1974, Covariance analysis of censored survival data. Biometrics 30, 89-99.

[26] Brown, J.A., and Glennon, D.C., 2000, Cost structures of banks grouped by strategic conduct. Applied Economics 32, 1591–605.

[27] Carree, M.A., 2002, Technological inefficiency and the skewness of the error component in stochastic frontier analysis, Economics Letters 77, 101-107.

[28] Chan, K.S., 1993, Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. The Annals of Statistics, 21, 520-533.

[29] Charnes, A., Cooper, W.W., and Rhodes, E.L., 1978, Measuring the efficiency of decision making units. European Journal of Operational Research 2, 429-444.

[30] Coelli, T., 1995, Estimators and hypothesis tests for a stochastic frontier function: A Monte Carlo analysis. Journal of Productivity Analysis 6, 247-268.

[31] Coelli, T., 1996, A guide to FRONTIER version 4.1: A computer program for stochastic frontier production and cost function estimation. CEPA working paper #96/07, Center for Efficiency and Productivity Analysis, University of New England, Arimidale, NSW 2351, Australia.

[32] Cole, R.A., and Gunther, J.W., 1995, Separating the likelihood and timing of bank failure. Journal of Banking and Finance 19, 1073-1089.

[33] 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(2), 103-117.

[34] Cole, R.A., and Wu, Q., 2009, Predicting bank failures using a simple dynamic hazard model. Working Paper.

[35] Cole, R.A., and Wu, Q., 2010, Is hazard or probit more accurate in predicting financial distress? Evidence from U.S. bank failures. MPRA Working Paper No. 24688.

[36] Cornwell, C., Schmidt, P., and Sickles, R.C., 1990, Production frontiers with cross-sectional and time series variation in efficiency levels. Journal of Econometrics 46, 185-200.

[37] Cox, D.R., 1972, Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187-220.

[38] Cox, D.R., and Oakes, D., 1984, Analysis of Survival Data. New York: Chapman & Hall.

[39] D'Agostino, R.B., and Pearson, E.S., 1973, Tests for departure from normality: Empirical results for the distribution of b_2 and $\sqrt{b_1}$. Biometrica 60, 613-622.

[40] Dasgupta, A., Steven, G., Selfb, and Das Gupta, S., 2007, Non-identifiable parametric probability models and reparameterization. Journal of Statistical Planning and Inference 137, 3380– 3393.

[41] Davidson, R., and MacKinnon, J.G., 1993, Estimation and Inference in Econometrics, New York, Oxford University Press.

[42] Davies, R.B., 1977, Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64, 247-254.

[43] Deakin, E., 1972, A Discriminant Analysis of Predictors of Business Failure, Journal of Accounting Research, 167-179.

[44] Debreu, G., 1951, The coefficient of resource utilisation. Econometrica 19, 273-292.

[45] Dempster, A.P., Laird, N.M., and Rubin, D.B., 1977, Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). Journal of the Royal Statistical Society, Series B 39, 1–38.

[46] Demsetz, H., 1973, Industry structure, market rivalry, and public policy, Journal of Law and Economics 16, 1-9.

[47] Deprins, D., Simar, L., and Tulkens, H., 1984, Measuring Labor Inefficiency in Post Offices. In the Performance of Public Enterprises: Concepts and Measurements. M. Marchand, P. Pestieau and H. Tulkens (eds.), Amsterdam, North-Holland, 243-267.

[48] DeYoung, R., 1999, Birth, Growth, and Life or Death of Newly Chartered Banks. Federal Reserve Bank of Chicago, Economic Perspectives 23, 18-35.

[49] DeYoung, R., 2003, The Failure of New Entrants in Commercial Banking Markets: A Split-Population Duration Analysis. Review of Financial Economics 12, 7-33.

[50] El-Gamal, M., and Grether, D., 1995, A Monte Carlo study of EC-estimation in panel data models with limited dependent variables and heterogeneity. In analysis of Panels and Limited Dependent Variable Models, Hsiao C, Lee LF, Lahiri K, Pesaran M.H. (eds). Cambridge University Press: Cambridge, 114-135.

[51] El-Gamal, M., and Inanoglou, H., 2005, Inefficiency and heterogeneity in Turkish banking: 1990-2000. Journal of Applied Econometrics 20, 641-664.

[52] Entani, T., Maeda Y., and Tanaka H., 2002, Dual models of interval DEA and its extension to interval data, European Journal of Operational Research 136, 32-45.

[53] Farewell, V.T., 1977, A model for a binary variable with time-censored observations. Biometrika 64, 43–46.

[54] Farewell, V.T., 1982, The use of mixture models for the analysis of survival data with long-term survivor. Biometrics 38, 1041-1046.

[55] Farrell, M., 1957, The Measurement of Productive Efficiency. Journal of the Royal Statistical Society A, General 120, 253-281.

[56] Gonzalez-Hermosillo, B., Pazarbasioglu, C., and Billings, R., 1997, Determinants of Banking System Fragility: A Case Study of Mexico. IMF Staff Papers 44, No 3.

[57] Greene, W.H., 1980a, Maximum likelihood estimation of econometric frontier functions. Journal of Econometrics 13, 27-56.

[58] Greene, W.H., 1980b, On the estimation of a flexible frontier production model. Journal of Econometrics 13, 101-115.

[59] Greene, W.H., 1990, A Gamma distributed stochastic frontier model. Journal of Econometrics 46, 141-164.

[60] Greene, W.H., 1997. Frontier Production Functions. Handbook of Applied Econometrics Vol. 2, Microeconomics, H. Pesaran and P. Schmidt, eds., Oxford University Press, Oxford.

[61] Greene, W.H., 2005, Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model. Journal of Econometrics 126, 269–303.

[62] Greene, W.H., 2007, LIMDEP Version 9.0 User's Manual, New York: Econometric Software, Inc.

[63] Greene, W.H., 2007, The Econometric Approach to Efficiency Analysis, in H. O. Fried, C. A. K. Lovell and S.S. Schmidt, eds., The Measurement of Productive Efficiency: Techniques and Applications. New York: Oxford University Press.

[64] Halling, M., and Hayden, E., 2006, Bank Failure Prediction: A Two-Step Survival Time Approach. Available at SSRN: http://ssrn.com/abstract.

[65] Hansen, B.E., 1996, Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64, 413-430.

[66] Hansen, B.E., 1999, Threshold effects in non-dynamic panels: Estimation, testing, and inference. Journal of Econometrics 93, 345-368.

[67] Hansen, B. E., 2000a, Sample splitting and threshold estimation. Econometrica 68, 575–603.

[68] Hausman J.A, and Taylor W. E.,1981, Panel data and unobservable individual effects. Econometrica 49, 1377-1398.

[69] Hicks, J. R., 1935, Annual survey of economic theory: the theory of monopoly. Econometrica 3, 1-20.

[70] Horrace, W., and Schmidt, P., 1996, Confidence Statements for Efficiency Estimates from Stochastic Frontier Models. Journal of Productivity Analysis 7, 257-282.

[71] Huang, R., 2004, Estimation of Technical Inefficiencies with Heterogeneous Technologies. Journal of Productivity Analysis 21, 277–296.

[72] Hughes, J.P. and Mester, L.J., 1993, A Quality and Risk-Adjusted Cost Function for Banks: Evidence on the "Too-Big-to-Fail" Doctrine. Journal of Productivity Analysis 4, 293-315.

[73] Jayasuriya, S.R., 2000, Essays on Structural Modeling Using Nonparametric and Parametric Methods with Applications in the U.S. Banking Industry. Unpublished Ph.D. dissertation, Rice University.

[74] Johansen, S., 1983, An Extension of Cox's Regression Model. International Statistical Review 51, 258–262.

[75] Johnson, N.L., Kotz, S., and Kemp, A.W., 1992, Univariate discrete distributions, second edition, New York: John Wiley & Sons.

[76] Johnson, N.L., Kotz, S., and Balakrishnan, N., 1994, Continuous Univariate distributions, Vol. 1, second edition, New York: John Wiley & Sons.

[77] Jondrow, J., Lovell, C.A.K., Materov, I.S., Schmidt, P., 1982, On the estimation of technical inefficiency in the stochastic frontier production function model. Journal of Econometrics 19, 233-238.

[78] Kalbfleisch, J. D., and Prentice, R.L., 2002, The Statistical Analysis of Failure Time Data. 2nd ed. New York: Wiley.

[79] Kaparakis, E.I., Miller, S.M., and Noulas, A.G., 1994, Short-run Cost Inefficiency of Commercial Banks: A Flexible Stochastic Frontier Approach. Journal of Money, Credit and Banking 26, 875-893.

[80] Kasa, K., and Spiegel, M.M., 2008, The Role of Relative Performance in Bank Closure Decisions. Economic Review, Federal Reserve Bank of San Francisco.

[81] Kashyap, A.K. and Stein, J.C., 1995, The impact of monetary policy on bank balance sheets. Carnegie-Rochester Conference Series on Public Policy 42, 151-195, North-Holland.

[82] Kim, Y., and Schmidt, P., 2000, A review and empirical comparison of Bayesian and classical approaches to inference on efficiency levels in stochastic frontier models with panel data. Journal of Productivity Analysis 14, 91-118.

[83] Klein, J.P., and Moeschberger, M.L., 2003, Survival Analysis. Techniques for Censored and Truncated Data, 2nd Edition, Springer.

[84] Kneip, A., Sickles, R.C., and Song W., 2011, A new panel data treatment for heterogeneity in time trends. Econometric Theory, forthcoming.

[85] Kuk, A.Y C., and Chen, C.H., 1992, A mixture model combining logistic regression with proportional hazards regression. Biometrika 79, 531-541.

[86] Kumbhakar, S.C., 1990, Production frontiers, panel data, and time-varying technical efficiency. Journal of Econometrics 46, 201-212.

[87] Kumbhakar, S.C., and Lovell, C.A.K., 2000, Stochastic Frontier Analysis. Cambridge University Press, Cambridge.

[88] Kutlu, L., and Sickles, R.C., 2010, Testing the Quiet Life and Efficient Structure Hypotheses: A Dynamic Model of Market Power. Rice University, Mimeo.

[89] Lancaster, T., 1990, The Econometric Analysis of Transition Data. Cambridge University Press, Cambridge.

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

[91] Leamer, E.E., 1978, Specification Searches: Ad hoc inference with nonexperimental data. New York, John Wiley and Sonc, Inc.

[92] Lee, L., 1993, Asymptotic distribution of the Maximum Likelihood Estimator for a Stochastic Frontier Function Model with a Singular Information Matrix. Econometric Theory 9, 413-430.

[93] Lee, Y.H., and Schmidt, P., 1993, A production frontier model with flexible temporal variation in technical efficiency. In: Fried, H.O, Lovell, C.A.K., Schmidt, P. (Ed.), The measurement of productive efficiency: Techniques and Applications, Oxford University Press.

[94] Lovell, C.A.K., Richardson, S., Travers, P., and Wood, L., 1994, Resources and Functionings: A new view of inequality in Australia. Models and Measurements of Welfare and Inequality, W. Eichhorn, Berlin: Springer-Verlag.

[95] Martin, D., 1977, Early Warning of Bank Failure: A Logit Regression Approach. Journal of Banking and Finance 1, 249-276.

[96] McAllister, P.H., and McManus, D.A., 1993, Resolving the scale efficiency puzzle in banking. Journal of Banking and Finance 17, 389-405.

[97] McFadden, D., 1974, Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (editor), Frontiers in Econometrics, New York: Academic Press.

[98] McFadden, D., 1981, Econometric Models of Probabilistic Choice. in C. Manski and D. McFadden (editors), Structural Analysis of Discrete Data with Econometric Applications, Cambridge, Mass.: M.I.T. Press.

[99] McLachlan, G., and Krishnan, T., 1996, The EM Algorithm and Extensions. John Wiley & Sons, New York.

[100] Meeusen, W., and van den Broeck, J., 1977, Efficiency estimation from Cobb-Douglas production functions with composed error. International Economic Review 18, 435-444.

[101] Mester, L.J., 1993, Efficiency in the savings and loan industry. Journal of Banking and Finance 17, 267-286.

[102] Mester, L.J., 1994, How efficient are third district banks? Business Review Federal Reserve Bank of Philadelphia.

[103] Meyer, B., 1990, Unemployment insurance and unemployment spells. Econometrica, 58, 757-782.

[104] Meyer, P.A., and Pifer, H.W., 1970, Prediction of Bank Failures, The Journal of Finance, 25, 853-868.

[105] O'Donnell, C., and Griffiths, W., 2004, Estimating State Contingent Production Frontiers. Working Paper Number 911, Department of Economics, University of Melbourne.

[106] Olson, J.A., Schmidt, P., and Waldman, D.M., 1980, A Monte Carlo study of estimators of the stochastic frontier production function. Journal of Econometrics 13, 67-82.

[107] Orea, C., and Kumbhakar, S.C., 2004, Efficiency Measurement Using a Latent Class Stochastic Frontier Model. Empirical Economics 29, 169–184.

[108] Peto, R., 1972, Contribution to the discussion of paper by D.R. Cox (1972). Journal of the Royal Statistical Society, Series B 34, 205-207.

[109] Pitt, M., and Lee, L., 1981,The measurement and sources of technical inefficiency in the Indonesian weaving industry. Journal of Development Economics 9 (1), 43-64.

[110] Qian, J., and Sickles, R.C., 2008, Stochastic frontiers with bounded inefficiency. Rice University. Mimeo.

[111] Rao, C.R., 1973, Linear Statistical Inference and its Applications, 2nd edition New York: Wiley.

[112] Richmond, J., 1974, Estimating the Efficiency of Production. International Economic Review 15, 515–521.

[113] Ritter, C., and Simar, L., 1997, Pitfalls of normal/Gamma Stochastic Frontier Models. Journal of Productivity Analysis 8, 167–182.

[114] Rothenberg, T.J., 1971, Identification in parametric models. Econometrica 39 (3), 577–591.

[115] Saunders A., and Cornett, M.M., 2004, Financial Markets and Institutions: A modern perspective. McGraw-Hill Companies, Inc. publications.

[116] Schmidt, P., and Lin, T., 1984, Simple tests of alternative specifications in stochastic frontier models. Journal of Econometrics 24 349-361.

[117] Schmidt, P., and Sickles, R.C., 1984, Production frontiers and panel data. Journal of Business and Economic Statistics 2, 367-374.

[118] Sealey, S., and Lindley, J.T., 1977, Inputs, outputs, and a theory of production and cost at depository financial institutions. Journal of Finance 32, 1251-1266.

[119] Sena V., 1999, Stochastic frontier estimation: A review of the software options. Journal of Applied Econometrics 14, 579-586.

[120] Shapiro, A., 1986. Asymptotic theory of overparameterized structural models. Journal of American Statistical Association 81 142–149.

[121] Shephard, R., 1953, Cost and Production Functions, Princeton University Press, Princeton, NJ.

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

[123] Simar, L., and Wilson, P.W., 2010, Estimation and Inference in Cross-Sectional, Stochastic Frontier Models. Econometric Reviews 29, 62-98.

[124] Stigler, G.S., 1958, The economies of scale. Journal of Law and Economics, 1, 54-71.

[125] Stevenson, R.E., 1980, Likelihood functions for generalized stochastic frontier estimation. Journal of Econometrics 13, 57-66.

[126] Swamy, P.A.V.B and Tavlas, G.S., 1995, Random Coefficient Models: Theory and Applications. Journal of Economic Surveys 9 (2), 165-96.

[127] Sy, L., and Taylor, J., 2000, Estimation in a Cox proportional hazards cure model. Biometrics 56, 227-236.

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

[129] Topaloglu, Z., and Yildirim, Y., 2009, Bankruptcy Prediction. Working Paper.

[130] Torna, G., 2010, Understanding Commercial Bank Failures in the Modern Banking Era. Available at: http://www.fma.org/NY/Papers/ModernBanking-GTORNA.pdf

[131] Tortosa-Ausina, E., 2002, Cost efficiency and product mix clusters across the Spanish banking industry. Review of Industrial Organization 20, 163–181.

[132] Train, K., 2003, Discrete Choice Methods with Simulation, Cambridge University Press.

[133] Tsionas, E.G., 2002, Stochastic frontier models with random coefficients. Journal of Applied Econometrics 17, 121–47.

[134] Tsionas, E.G., and Kumbhakar S.C., 2004, Markov Switching Stochastic Frontier Model. The Econometrics Journal 7, 1-28.

[135] Tsionas, E.G., and Papadogonas, T.A., 2006, Firm exit and technical inefficiency. Empirical Economics 31, 535-548.

[136] Waldman, D.M., 1982, A stationary point for the stochastic frontier likelihood. Journal of Econometrics 18, 275-279.

[137] Wang, D., and Kumbhakar, S.C., 2009, Strategic groups and heterogeneous technologies: an application to the US banking industry'.Macroeconomics and Finance in Emerging Market Economies 2:1,31-57.

[138] Wang, W.S., and Schmidt, P., 2008, On the distribution of estimated technical efficiency in stochastic frontier models. Journal of Econometrics 148, 36-45.

[139] Whalen, G., 1991, A proportional hazards model of bank failure: An examination of its usefulness as an early warning model tool. Economic Review, Federal Reserve Bank of Cleveland, 21-31.

[140] Wheelock, D.C., and Wilson, P., 1995, Explaining Bank Failures: Deposit Insurance, Regulation, and Efficiency. Review of Economics and Statistic, 77, 689-700.

[141] Wheelock, D.C., and Wilson, P., 2000, Why do banks disappear? The determinants of U.S. bank failures and acquisitions. Review of Economics and Statistics 82 (1), 127–138.

[142] Wheelock, D.C., and Wilson, P., 2001, New evidence on returns to scale and product mix among U.S. commercial banks. Journal of Monetary Economics 47, 653-674.

[143] Yildirim, Y., 2008, Estimating default probabilities of cmbs with clustering and heavy censoring. The Journal of Real Estate Finance and Economics 37 (2).