MONASH University

Non- and Semi-parametric Methods for Modelling Recovery Rates

A thesis submitted for the degree of

Doctor of Philosophy

by

Nithi Sopitpongstorn

Master of Applied Econometrics, Monash University, Australia

Master of Business Economics, Monash University, Australia

B.Econ., Kasetsart University, Thailand

Department of Econometrics and Business Statistics

Monash Business School

Monash University

Australia

March 2018

Contents

List of Figures

List of Tables

Copyright notice

Under the Copyright Act 1968, this thesis must be used only under the normal conditions of scholarly fair dealing. In particular no results or conclusions should be extracted from it, nor should it be copied or closely paraphrased in whole or in part without the written consent of the author. Proper written acknowledgement should be made for any assistance obtained from this thesis.

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

Abstract

The recovery rate (RR), which indicates the proportion of the loss from a particular defaulted loan that has been recovered, is essential to quantifying credit risk. Developments in RR modelling are driven by its application in credit risk management to control, monitor, and mitigate credit risk exposure. These developments enable creditors to enhance their economic and strategic financial decision making, which directly improves their competitive advantage. After the global financial crisis that is largely caused by credit crunch, the modelling and predicting RR have become very popular in the empirical finance literature. However, RR modelling has been found very challenging, mostly due to its non-standard empirical features, such as the RR lies within the unit interval [0,1], its distribution is bimodal with high concentrations at the boundaries, and nonlinearity and complex relationships between the RR and its covariates. Although there exists a vast literature on RR modelling, because of these atypical properties, the models studied in the literature have several limitations. They include the model estimates being biased due to *transformation and back transformation* of bounded RR data, predicted RR exceeding the boundaries, and the absence of model parameter estimates in order to conduct statistical inference on the model parameters due to *black-box* functional form.

The core objectives of this study are to: propose non- and semi-parametric methods for modelling the conditional mean regression and the conditional quantile regression, apply them to the widely studied Moodys recovery dataset, uncover the underlying nonlinear relationship between the RR and its covariates which include loan/borrower characteristics as well as the state of the economy, and improve the

accuracy of RR prediction. Moreover, this thesis compares and contrasts the predictive performances of proposed models with the existing ones in the literature. There are several significant contributions of the thesis to the recent literature of RR modelling. This is the first study to introduce non- and semi-parametric regression models for RR data and the aim is also to either eliminate completely or notably reduce the problem of predicted RR exceeding boundaries 0 and 1. Additionally, this is also the first study to estimate the response of recoveries to economic downturns in the context of both the conditional mean regression as well as the conditional quantile regression. The downturn RR is the crucial component in calculating mostly needed credit risk exposure during crises.

The proposed data driven local linear estimation of the conditional mean regression plays an important role in capturing the underlying nonlinear relationship between the RR and its determinants. The estimation of the marginal effects of covariates on RR becomes straightforward in this setting, and these results were utilised in specifying improved functional form for the proposed semi-parametric partial linear model for the conditional mean regression. Assessed by several model selection and prediction criteria, we find that the proposed models generate superior in-sample and out-of-sample RR predictions to those generated by existing models. Clearly, the proposed models do not only improve the predictions of defaulted loans recoveries relative to their parametric counterparts, they also enable reliable statistical inference of marginal and interaction effects of loan/borrower characteristics and economic conditions on RR.

This thesis takes a step forward from the recent literature and propose a nonparametric quantile regression for RR, in which the RR-covariate relationship, and the marginal and interaction effects were estimated at the various quantiles of the RR distribution. As such the heterogeneity in the marginal and interaction effects can be estimated in this framework. Furthermore, we also introduce partially linear additive quantile regression to estimate the overall effects of the covariates in order to generalise the interpretation of the results of nonparametric model. We provide evidence of the presence of heterogeneous effects at the various conditional quantiles. These findings improve our understanding of these complex relationship between RR and its covariates. In particular, we find that the results of the proposed quantile regressions can be used in designing strategic risk management by lenders, as well as downturn credit risk policies by regulators.

In the foregoing investigation into nonparametric and semiparametric methods for RR modelling, we find that the boundary issues are largely mitigated in the nonparametric and semiparametric settings, but there is no guarantee that the RR predictions generated by the local linear or partially linear models would fully lie within the unit interval. To ensure that the boundary problem is eliminated completely as well as improving other modelling aspects, we propose a flexible and robust nonparametric local logit regression for RR, and by its construction, the RR prediction would lie in [0,1]. This methodology was proposed by integrating two ideas, one is the QMLE regression for fractional response data and the other is the local logit model for the binary response variable.

The results highlight significant nonlinear marginal and interaction effects of conditioning variables on the recoveries of defaulted loans. Moreover, our analysis indicates that model specification, in particular the functional form plays an important role in improving the RR prediction. We assessed the merits of both nonparametric model and nonlinear parametric model whose functional form has been improved by exploiting the outcomes of the comprehensive marginal and interaction effects analysis of nonparametric regression. As such, we call this the calibrated model. The calibrated model performs as good as the proposed model in the in-sample and out-of-sample RR predictions. The calibrated model will be attractive to applied researchers and industry professionals working in the risk management area who are unfamiliar with nonparametric machinery.

Acknowledgement

First and foremost, I owe my heartfelt gratitude to my supervisors, Prof. Param Silvapulle and Prof. Jiti Gao for their steady guidance, continual support and encouragement. Param has kindly and thoughtfully supervised me from my Master's degree to this PhD, over the last 5 years. She always said, "don't get distracted and be focused", this helped me to rethink and be back on the right track throughout the process of completing this thesis. I also thank Jiti for dedicating his time. I have an impression that Jiti always has an answer or useful suggestion to my technical questions and confusions. They both are the best supervisors for any PhD students can wish for and I am very fortunate and feel very privileged to have them.

I would like to acknowledge PhD scholarships from Monash Business school and the department of Econometrics and business statistics to cover my tuition fee and living stipends. I would like to thank Australian Mathematical Sciences Institute (AMSI) and Australia and New Zealand banking group (ANZ), whole sale credit risk modelling department, who selected me to join 4 months internship as a credit risk modelling specialist at the bank. This was such a great/unique opportunity to have hands-on industrial experience, which helped me to improve the presentation of this thesis.

My appreciation is given to Prof. Oliver Linton for his very informative suggestions and comments on the nonparametric censored regression in Chapter 3 during his visit at Monash University. Also, the short course, provided by

Prof.Jeffrey Racine at Melbourne University, significantly boosted my learning curve and inspired me in nonparametric approaches, which played important role at the early stage of my work. This thesis has also benefited from various comments on Chapter 4 at the 29th Australasian Finance and Banking Conference 2016 at the University of New South Wales.

God bless all my family members who have shown so much faith in me throughout my challenging situation in life. I must heartedly thank my dad, Prayouth Sopitpongstorn, who has believed in my ability, even the time when I myself did not; and my mom, Poonsuk Sopitpongstorn, who has listened and supported me beyond measure. Without my loving wife, Dr. Pearpilai Jutasompakorn, who is staying by my side, inspiring me, sharing my hopes and dreams, I cannot be where I am today.

Finally, I would like to dedicate this thesis to my baby girl, Iris Sopitpongstorn.

Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma in any university or equivalent institution, and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

[redacted]

(Nithi Sopitpongstorn)

Chapter 1

Introduction

1.1 Background

The recovery rate (RR) of debt in the event of default is a crucial determinant of the default risk premium required by lenders and the regulatory capital needed to minimise exposure to losses. The Basel II Accord (2006) on capital adequacy requires internationally active banks to develop statistical models of credit risk and use them to determine the capital to be held against credit risk exposure. Scandizzo (2016) shows that the RR estimates have direct influence on the lender's capital adequacy than other risk parameters, as RR reflects the potential loss of the credit portfolio. Following the recent global financial crisis that is mostly caused by the credit crisis in the US, Basel Committee on Banking Supervision (2011) has reported that the regulatory capital requirements that the banks provided under advanced measurement approach were not sufficient enough to withstand the credit losses during the crisis, due to inadequate modelling of credit risk. As a consequence, this has been resulted in a notable increase in research into RR modelling, mostly for the purpose of recovery prediction by academics and industry professionals.

In the quest to find a model for RR relating to conditioning variables, several studies have observed that this modelling exercise presents some challenges due to several key empirical features of the RR: First, it is continuous, fractional, and bounded in [0,1]. Second, its empirical density is bimodal and asymmetric, with high proportions of recoveries at the boundaries zero and one. Third, in the presence of observations at 0 and 1, trimming and *transformation* as well as *back-transformation* of recoveries are needed for the use of valid statistical theory. Such transformation introduces bias in the model estimates, resulting in unreliable statistical inference. Despite the growing body of evidence of the presence of nonlinearity in the recovery-covariate relationship, there is little attempt that has been made in the literature to improve the specification of the widely used linear regression model.

The three aforementioned properties of RR lead to the problems and challenges intrinsic to building statistical models to account for RR-covariates relationships at the time of default and to capture the specific features of the recovery distribution. In analysing the RR-covariate relationship, the loan characteristics, including collateral status, types of loan, degrees of subordination, and debt cushion are considered in numbers of studies such as Altman and Kalotay (2014); Siao, Hwang, and Chu (2015); Chellathurai (2017) among others. Additionally, the macro-economic environment is also expected to have an effect on the loan's recovery, as the RR are generally found to be low during the economic downturns.

This thesis is the first study to propose data driven nonparametric and semiparametric regressions for RR modelling using kernel estimation methods to estimate the RR-covariate relationships in both conditional mean and quantiles of RR. The proposed models address two main limitations of the existing RR models. First, our models can capture and explain the underlying complex nonlinear RR-covariates relationship as well as the atypical RR properties in a flexible manner without the need to pre-specify the key assumptions, including data transformation, functional form and distribution assumptions. This in turn mitigates the misspecification issue, inappropriate assumptions and an additional bias in RR estimates of most parametric models. Second, we pay considerable attention to uncovering the underlying marginal effect of the covariates on RR, which has been largely unexplored in the existing studies.

Our study does not only discuss nonlinear marginal effect analysis, but also the other crucial relationships including interaction effects and heterogeneous effects at various conditional quantiles of RR. Although the machine learning algorithms for RR could be as flexible as our proposed models, the black-box problem remains one of their main limitations in order to estimate the transparent nonlinear relationships. Moreover, the proposed models for RR facilitate the estimation of downturn RR. Then, the performances of the proposed models are compared with those of various existing RR models by numbers of predictive accuracy criteria.

1.2 Thesis overview

As discussed, we build on insights from the findings of large-scale empirical research as well as studies documenting the merits of nonparametric and semi-parametric approaches for recovery predictions and marginal effect analysis. To this end, the research outcomes of the new flexible modelling framework that we propose would be useful in developing appropriate policies to mitigate underlying credit risk exposure, which in turn would improve risk management, risk monitoring, and credit risk pricing. In what follows, the outline in each chapter of the thesis are summarised.

Chapter 2 reviews the development of RR modelling and its applications in the literature on credit risk modelling. This chapter discusses the way in which the existing models attempts to address the specific properties of RR. There is a vast literature that proposes a wide range of estimation methods, ranging from standard parametric regressions to sophisticated machine learning algorithms. This literature review chapter begins by presenting an overall picture of RR modelling to provide background and address the key empirical features of the RR data. We then discuss detailed descriptions of recent developments in econometric techniques to deal with the specific properties of the RR. Subsequently, the key features of each model are discussed in order to compare their advantages and limitations. In this brief literature review, we identify the limitations of existing methods for modelling RR-covariate relationship, and a large gab in the literature. They provide motivation for the proposed models in the thesis to overcome the limitations.

In chapter 3, we propose both nonparametric and semiparametric conditional mean regression models for the RR, and estimate these models using the local linear and local constant methods. The nonparametric regression with the local linear estimation method facilitates a straightforward analysis of nonlinear marginal effects, which is useful in understanding the underlying relationships of recovery covariates. With the comprehensive marginal analysis of local linear model, we improve the functional form of our proposed semiparametric partially linear regression. This eases the computational difficulties associated with the fully nonparametric regression and avoids the misspecification problem in the semiparametric models. We are also interested in the predictive accuracy of the proposed model. The out-of-sample predictive performance of nonparametric and semiparametric regressions are evaluated against that of several widely studied parametric regression models such as inverse Gaussian regression, quasi maximum likelihood (QMLE) regression for fractional response variable, the Tobit model with two-sided censoring, the mixture distribution model proposed by Altman and Kalotay (2014), and the machine learning regression tree classification algorithm. Moreover, we propose a two-sided censored (at zero and one) nonparametric modelling framework, which is an extension of the one-sided censored nonparametric regression introduced by Lewbel and Linton (2002). A simulation study is conducted to assess the properties and practicality of this proposed model. This is followed by an empirical application of the model to RR data.

In chapter 4, we propose a nonparametric quantile regression for RR, in which the RR-covariate relationship including the marginal and interaction effects are estimated at the various quantiles of the RR distribution. This leads to further understanding of the heterogeneity of the RR-covariate relationship on the various parts of the RR conditional distribution, rather than that at the central tendency of RR as found in most existing studies. We also propose an improved two-stage estimation method for the partially linear additive quantile regression. This model provides an alternative way to estimate the overall linear, nonlinear and heterogeneous effects of the covariates on RR due to the flexible specification of the model. Subsequently, the relative performances of the proposed models are compared with those of their parametric counterpart in various aspects: goodness of fit, point prediction, distributional fit, and the RR Value-at -Risk evaluation.

In chapter 5, we propose a flexible and robust nonparametric local logit regression for RR, and by its construction, the RR prediction would lie within [0,1] interval. The proposed methodology integrates the analogous QMLE regression for fractional response data (QMLE-RFRV) by Papke and Wooldridge (1996), and the local logit model for the binary response variable introduced by Frölich (2006). This method guarantees the resolution of the boundary problem. In order to assess the model's performance as well as the behaviour of the model parameter estimates in the finite sample, we conduct a large-scale simulation study under various experimental designs. The local logit model is then applied to the RR data, which does not only provide the detailed analysis of nonlinear marginal effects, but also an additional information on the interaction effects of the covariates on

RR, which has not been addressed in the literature before. The results of these marginal and interaction effects of covariates on RR are also utilised to specify an improved nonlinear functional form of the parametric QMLE-RFRV by means of a “calibration” method. In addition, the models are robustly evaluated to compare their in-sample and out-of-sample predictive performances.

Chapter 7 concludes the thesis with an overview of the thesis’ main contributions, along with an outline of directions for future research.

Chapter 2

Literature review

2.1 Introduction

The largest and most obvious source of credit risk are loans that fail to meet their obligations in accordance with the agreed-upon lending terms. Measuring credit loss is crucial to maintaining credit exposure within acceptable levels in order to optimise a bank's risk-adjusted rate of return. Understanding the mechanism of credit loss using advances in statistical tools is a critical component of a comprehensive approach to risk management. Overall aim of this chapter is to: review the statistical models proposed for RR data, discuss advantages and disadvantages of them, and identify gaps and unanswered questions. In doing so, we identify potential research topic for the thesis.

The RR, which specifically indicates the proportion of the losses recovered in the event of default, is essential to quantifying credit risk, together with the probability of default and the exposure at default (Bohn & Stein, 2011). Developments in RR modelling are driven by practical needs in banking risk management to control, monitor, and mitigate credit risk exposure. These developments aid creditors in enhancing their economic and strategic financial decision making, which directly improves their competitive advantage (Gürtler & Hibbeln, 2013). Therefore, RR is one of the key components for provisioning credit losses, calculating risk capital, and determining fair pricing for credit risk obligations. The accuracy of the RR estimates is fundamental to calculating potential credit losses. Recognition of its importance has led to a growing number of studies in RR modelling, which will be discussed in this chapter.

Various methods have previously been proposed to deal with the specific properties of RR data. The central objective of most empirical studies is RR prediction in order to serve one of the practical needs. However, several trivial aspects remain challenging due to the empirical features of historical RR, including the bimodal [0,1] bounded response with high concentrations at the boundaries (Qi & Zhao, 2011; Schuermann, 2004).

In fact, non-negative response variables ($y \geq 0$) are very common in financial data such as percentages, proportions, and fractions. The parametric generalized linear model with any of the exponential family distributions is fitted to such one-sided response variables. For the two-sided bounded data, the beta regression has been recommended due to its support as well as the flexibility in its shapes (Kieschnick & McCullough, 2003). The model can take into account of the bounded data with bimodality. However, when the response variables have heavy tails at the boundaries zero and one, the model would not accommodate them. Although the support of the empirical density is [0,1], when applied beta regression, the support will shrink to (0,1) (Bayes & Valdivieso, 2016). As the probability masses at zero and one are clearly presented in the empirical density of RR, the standard beta distribution would not be suitable, which leads to applications of several non-standard regressions for RR modelling.

Econometric models that are able to handle the bounded continuous response variable are limited, and additional restrictions are normally required to address such properties. In addition, the underlying relationships between the RR and its covariates are complex and are expected to be non-linear. The conventional parametric linear method is unable to address such relationships. These concerns have led to the further introduction of various elaborated methods (Loterman, Brown, Martens, Mues, & Baesens, 2012; Altman & Kalotay, 2014). For example, Bastos (2010) and Qi and Zhao (2011) present applications of regression tree models, and Calabrese (2012) and Altman and Kalotay (2014) discuss parametric regression and mixture models (further details of these studies are provided in the sections that follow).

Plan of this chapter, we provide an overview of previous studies in RR modelling. Section 2.2 begins by presenting the definition, importance, and application of the RR in practice. This section also provides a discussion of some specific empirical properties of RR followed by the determinants of RR. Then, we turn to a discussion of the methodologies including *back transformation* regressions, conventional parametric regressions for bounded response variable, and data driven models in Section 2.3. The other statistical approaches for RR modelling are discussed in Section 2.4. Then, Section 2.5 provides the conclusion of the chapter which address the research gap and motivation for the forthcoming chapters of this thesis.

2.2 Credit risk and recovery rate modelling

Defaulted credit loss is inevitable for banks and financial institutions who recognise lending as one of their core businesses. The amount of the loss reflects the riskiness of the lenders, which severely affects their fundamentals. To address the importance of RR modelling, its role in credit risk management is discussed. This is followed by a discussion of the definition and the stylised facts of the empirical RR.

2.2.1 Credit risk modelling

The implementation of the Basel accord encourages banking supervisors globally to promote sound practices for managing credit risk (Risk Management Group the Basel Committee on Banking Supervision, 1999). In 2004, the Basel Committee presented the global standard prudential regulation of banks, Basel II, as an international standard. The role of the Basel accord is to ensure that banks hold a sufficient amount of capital as a buffer for periods of economic instability that may cause banks' insolvency. The regulatory capital requirement is calculated based on three main variables, which are the probability of default, the RR, and the exposure at default (BIS, 2004). Most importantly, the accord allows banks to model their credit risks through these risk parameters with two main options, namely foundation and advanced internal ratings-based approaches. The difference between these two approaches is the degree of the bank's involvement in estimating the credit risk variables.

Under the advanced approach, the bank is allowed to estimate all its own risk variables. On the contrary, some variables are provided by the regulator in the foundation approach. The flexibility to estimate the values tailored to their portfolio is likely to be a motivation for a bank to move from the foundation to the advanced approach (Schuermann, 2004). Estimating their own risk parameters is not only a better reflection of the risks involved in a particular bank, it also allows the bank itself to understand the risk structure and their borrower behaviours, which would lead to various internal applications.

Applications of credit risk models include determining the default risk premium and credit risk portfolio diversification, pricing default risk insurance, and the emergence of distressed debt as an investment class (Altman & Kalotay, 2014). These are essential for risk management to identify and quantify the risk related to credit exposures, which lead to the proper allocation of banks' capital as well as to an improvement of their strategic and economic decisions (Hartmann-Wendels, Miller, & Töws, 2014). These benefits are related initially to: the profitability of the individual bank, stability in the banking system, and efficient allocation in the economic resource. After the severe consequences of the recent global financial crisis, of which credit risk defaulted losses were one of the leading causes (Brunnermeier, 2009), banks can learn from past experiences by raising awareness and identifying, measuring, monitoring, and controlling the risks that might prevent a reoccurrence of the crisis.

2.2.2 The empirical recovery rate and its properties

Among the three key risk parameters, the probability of default has been studied in both academic and practical research for many decades (Saunders & Allen, 2010). On the other hand, after the implementation of Basel II in 2004, an extensive body of research in RRs arose, remedying the paucity of research in this area prior to the implementation (Schuermann, 2004). Recent studies have attempted to satisfy the criteria of the accord, which requires a robust system in place to validate the accuracy and consistency of rating systems, processes, and the estimation of all relevant risk components. A bank must demonstrate to its supervisory that the internal validation process enables it to assess the performance of its internal rating and risk estimation systems using an appropriate framework and to interpret the results meaningfully (BIS, 2004).

Furthermore, Basel committee on banking supervision (2001) strongly encourages banks to estimate their own RR rather than to rely on the foundation approach. They claim that the former option has the benefit of being able to recognise banks' own loss experiences, which can take into account lending standards and legal environments across markets and products. Furthermore, and perhaps more importantly, the RR is bank-specific, depending on the internal definitions of defaults and losses, the financial products offered, the lending policies, and procedures in the recovery process. If a bank performs its own estimation, these specific factors can be taken into account, which has a material impact on the recovery estimates. Consideration of these factors is limited in a ’one-size-fits-all’ approach.

Definition and empirical features

The RR is defined as the proportion of the loss that has been recovered from the exposure at default. It is fundamental to estimating the potential credit loss, which plays an important role in lending, investing, trading, or the pricing of loans. The accuracy of the estimates determines the efficiency in provisioning reserves for credit loss, calculating credit capital, and determining fair pricing for credit-risky obligations (Gupton & Stein, 2005). Scandizzo (2016) suggests that the role of RR estimates under the Basel accord has a more direct influence on the lender’s capital adequacy than the probability of default.

The RR is a continuous variable which lies in the unit interval [0,1]. One end of the boundary represents full recovery (1) and the other represents a total loss (0). This implies that lenders can neither recover nor lose an amount of the defaulted loss that exceeds the outstanding amount. In practice, some actual observations do fall outside of the boundaries, due to additional economic costs such as the additional collection cost and fees and penalties paid. However, these do not reflect the nature of the RR structure, so the recovery estimates are expected to be constrained to this interval (Tong, Mues, & Thomas, 2013).

The bimodal property is commonly observed in the empirical density of the RR, as the data normally has high concentrations at one followed by zero, with the remainder of the values dispersed across a broad range of losses (Altman & Kalotay, 2014; De Servigny & Renault, 2004). Intuitively, there is a clear incentive for lenders to try to recover the majority of their defaulted debts, which results in the high mass at the full RR. In a default event, banks can resolve the default by restructuring the terms of their debt contracts and renegotiating before the defaulter files for bankruptcy (Gilson, John, & Lang, 1990). An accurate recovery estimate and knowledge about the factors affecting the level of the recovery are useful in order to construct an appropriate restructuring program.

Recovery covariates

Recoveries after default intuitively result from restructured or cured exposures as well as bankrupt borrower's assets (Peter, 2011). One of the main factors determining the amount of recovery is collateral status. Banks favour granting loans with collateral, as it is crucial to the valuation of a potential loss. Collateral is a means of compensation for losses as the collateral of a defaulter is being liquidated after filed bankruptcy. Hence, the quality and value of the collateral influences the level of the RR (Araujo, Kubler, & Schommer, 2012). On the contrary, it is difficult to recover losses from uncollateralised loans.

In the past, the value of the collateral was taken as a proxy for the amount of recovered loss. Later, however, many studies suggested that considering only collateral value is inadequate. Jokivuolle and Peura (2003) argue that collateral does not provide as much protection as is normally expected. The value of the assets that serve as collateral depend on the overall business conditions, which are driven by common market factors. During an economic crisis, the value of borrower's total assets would be lower than the amount of its debts, and thus the value of the collateral would be so as well.

The capital structure of the defaulter is another important factor (Acharya, Bharath, & Srinivasan, 2007). For example, in the event of liquidating assets for different debt and shareholders, commercial loans have first seniority, followed by corporate bonds. Bruche and González-Aguado (2010) claim that the quality of the collateral is more important for senior than for junior secured debt, due to the priority order in which the assets are recouped. Both collateral status and seniority are firm-specific or idiosyncratic factors that are highly important from the lender's point of view to mitigate credit risk internally.

To adhere to the Basel accord, banks are required to demonstrate that the recovery estimate of each defaulted loan is reliable and consistent with their underwriting standards, risk profile, and available relevant data[1] (Basel committee on banking supervision, 2001). Importantly, estimations of the RR cannot rely solely on the collateral's estimated current market value: the estimates must take into account the effects of the macro-economic environment (Han & Jang, 2013; Thomas, Matuszyk, So, Mues, & Moore, 2016). The Basel accord requires banks to estimate the downturn recovery rate, which appropriately reflects the estimates during an economic downturn. Therefore, the relationships between RR and economic condition must be examined.

Several studies have shown that recovery tends to be high during an economic upturn and low during a downturn (Resti, 2002; Fong, 2006; Carey & Gordy, 2004; Bruche & González-Aguado, 2010; Altman & Kalotay, 2014; Koopman, Lucas, & Schwaab, 2012). Carey and Gordy (2004), for example, indicate that the distribution of the recovery shifts to the right during periods of good economic conditions compared to bad ones. However, how the macroeconomic variable affects recovery has not been extensively explored for the following reasons. First, recovery data might be unavailable for the full business cycle. This leads to limited information for the investigation of such relationships (Calabrese, 2014). Second, the macro-economic variables might affect the RR indirectly through the unobserved credit cycle (Frye, 2000; Altman, 2006; Carey & Gordy, 2004; Bruche & González-Aguado, 2010). Moreover, the relationship may be nonlinear. The macroeconomic variables may have different effects during economic upturn and downturn, whereas linear effects are generally assumed.

[1] The relevant recovery data under Basel accord refers to be either historic experience or comparable external data

2.3 Developments in recovery rate modelling

In this section, we review and discuss the developments in recovery modelling. The central aim is to show how the existing methodologies handle the specific features of the data and to compare and contrast their limitations. In what follows, we categorise the estimation methods into three different groups: conventional linear regression with back transformation technique, parametric regressions for [0,1] bounded data, and data-driven approaches.

2.3.1 Ordinary least square linear regression with back-transformation of [0,1] bounded response variable

Conventional OLS linear regression has been employed in many applied economic and financial studies, including RR modelling. The model provides a transparent interpretation with a simple implementation method (Dwyer & Korablev, 2009). Specifically, an implication of the first-order condition, marginal effect, is generally useful to analyse the economic insight. Although OLS tends to be preferred by practitioners, linear regression is inappropriate for the RR-covariates:

$$E(Y|x) = x'\beta,$$

where $Y \in [0,1]$ is recovery rate, x is a $k \times 1$ vector of explanatory variables, and β is a $k \times 1$ unknown parameter vector. As the RR is bounded in the unit interval, the unknown parameter β rarely provides the best description of $E(Y|x)$ due to the constant effect of any particular variable throughout the domain of x, unless the range of the variable is very limited (Papke & Wooldridge, 1996). In additional, the application of OLS is also theoretical misleading as the error term is assumed to be normally distributed, which is inappropriate for $Y \in [0,1]$. Given the limitations, however, the model has been used in several studies as a benchmark model in order to compare its predictive performance with that of alternative models (Qi & Zhao, 2011; Bellotti & Crook, 2012). These studies suggest that OLS has comparable predictive accuracy, although some predictions are not in the appropriate range.

Rather than directly estimating the RR with a linear model, Gupton and Stein (2005) suggest the application of OLS associated with data transformation to address the boundary issue. This is referred to the "back-transformation" technique. It involves a multi-step procedure, which is as follows:

- Step 1: The boundaries zero and one of the empirical RR which is denoted as Y are adjusted by an arbitrary value ν. The adjustment ensures that the boundary adjusted RR does not take the values of zero and one, which can be written as $Y^{(\nu)} \in (0,1)$, where $Y^{(\nu)}$ is the boundary adjusted RR;

- Step 2: $Y^{(\nu)}$ is transformed:

$$Y^{*} = \eta(Y^{(\nu)}), \tag{2.3.1}$$

  where $\eta(\cdot)$ is any invertible function transforming $Y^{(\nu)} \in (0,1)$ to $Y^{*} \in (-\infty, \infty)$;

- Step 3: Conventional OLS is applied by regressing Y^{*} on x to estimate the unknown parameter:

$$Y^{*} = x'\beta^{*} + e,$$

  where β^{*} is the unknown parameter representing the relationship between Y^{*} and the RR-covariates.

- Step 4: Back-transformation is applied to $x\hat{\beta}^{*}$ using an inverse function $\eta^{-1}(\cdot)$:

$$E(Y|x) = \eta^{-1}(x'\hat{\beta}^{*}).$$

  This ensures that the back transformed estimate lies within the unit interval.

In Step 1, several options of transformation function are available. Hu and Perraudin (2002) employ the inverse standard Gaussian distribution function as $\eta(\cdot)$ in (2.3.1). This is perhaps the function most commonly selected, as it has typically been applied as a benchmark model for the RR. The logit function can also be applied (Siao et al., 2015). Alternatively, Gupton and Stein (2005) indicate that the beta function is more appropriate, as it allows some flexibility to take the bimodality into account. They specify the conversion of the beta-distributed RR as:

$$Y^* = \Phi^{-1}[\mathcal{B}(Y, a, b)],$$

where a and b are the beta distribution centre and shape parameters, respectively, $\mathcal{B}(\cdot)$ is the beta distribution function, and $\Phi^{-1}(\cdot)$ is the standard inverse Gaussian distribution function. In particular, the transformation function in (2.3.1) is defined as: $\eta(\cdot) = \Phi^{-1}(\mathcal{B}(\cdot))$. The beta probability density function allows a flexibility that can reflect the bimodality through those two distributional parameters. Hence, many studies claim that the distribution is suitable for recovery data (Gupton & Stein, 2005; Ferrari & Cribari-Neto, 2004; Hlawatsch & Reichling, 2010).

The problems with back-transformation regression

There are three main statistical issues with the back-transformation technique. Firstly, in Step 1, the adjustment of the zero and one boundaries affects the outcome of the model estimation, as the observations at the boundaries commonly exhibit a large probability mass, as discussed above. Qi and Zhao (2011) report that the magnitudes of the arbitrary adjustment value strongly affect the predictability, while there are no theoretical criteria governing the selection of the appropriate adjustment. In practice, Altman (2006); Siao et al. (2015), for instance, select an arbitrary value that minimises the in-sample predictive error.

Secondly, the interpretation of the parameter estimates in Step 3 may be unclear. As, OLS regression is applied on the transformed RR, the estimators then reflect the relationships under the transformation. These do not necessarily reflect the effect of the particular explanatory variable on the RR.

Thirdly, the back-transformation in Step 4 adds a bias for conditional mean regression due to Jensen's inequality:

$$\eta(E(Y^{(\nu)}|x)) \neq E(\eta(Y^{(\nu)})|x),$$

unless $\eta(u) = u$. This inequality implies that back-transformation regression leads to bias as $E(Y^{(\nu)}|x)) \neq \eta^{-1}(x'\beta^{*})$. However, Tobback, Martens, Van Gestel, and Baesens (2014) suggest that the back-transformation regression performs comparably to the more sophisticated methods and can yield more comprehensible results. The technique are also applied on other statistical models which cannot accommodate [0,1] response variable (more details will be further discussed in Section 2.4). Most studies to date overlooked the model estimate bias introduced by the above transformation technique. One of the aims of this thesis is to eliminate the bias completely.

2.3.2 Quasi maximum likelihood and two-sided censored TOBIT regressions for [0,1] bounded data

To handle the bounded response data without using back-transformation, Wooldridge (2010) suggests two parametric linear regressions, which are quasi-maximum likelihood estimation regression for fractional response variables (QMLE-RFRV) and the two-sided censored Tobit model. These models have been applied for RR modelling in Bastos (2010); Calabrese (2012); Altman and Kalotay (2014). The main advantage of these models is that they are theoretically appropriate for [0,1] bounded RR data. As this thesis aims to mitigate or eliminate the boundary problem in chapters 3-5, the review of these two parametric models provide some guideline to accommodate the problem.

Quasi-maximum likelihood estimation regression for fractional response variable

QMLE-RFRV is one of the most common solutions for dealing with bounded dependent data, including RR. The model imposes plausible constraints on the conditional mean to ensure that the recovery estimate lies within the interval [0,1]:

$$E(Y|x) = \Lambda(x'\beta), \tag{2.3.2}$$

where Y is the continuous [0,1] bounded RR, x is the vector of k covariates indicating the loan specifics, which can be a mixture of continuous and discrete covariates, $\Lambda(\cdot)$ is the logit link function, $0 < \Lambda(\cdot) < 1$, and β is a vector of unknown parameters. The link function plays the role of controlling the boundary of the predicted RR. Any function that satisfies $0 < \Lambda(\cdot) < 1$ could be also applied, such as the log-log, the complementary log-log and probit functions (Bastos, 2010; Calabrese, 2012; Ramalho, Ramalho, & Murteira, 2011).

Papke and Wooldridge (1996) propose a quasi-maximum likelihood estimation method (QMLE) to estimate the unknown parameter of the fractional [0,1] response variable's covariate:

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} y_i \log(\Lambda(x_i'\beta)) + (1 - y_i)\log(1 - \Lambda(x_i'\beta)). \tag{2.3.3}$$

The vector of estimators defined in (2.3.3) is consistent and asymptotically normal regardless of the conditional distribution assumption. Although the QMLE-RFRV structure is identical to that of logistic regression, the variance based on the standard model is unreliable for fractional data. Therefore, Papke and Wooldridge (1996) propose fully robust sandwich standard errors and test statistics in order to achieve valid statistical inferences. The valid asymptotic inference of the estimators in (2.3.3) is useful for RR analysis, as Khieu, Mullineaux, and Yi (2012) employ the model to study the determinants of the RR.

The predictive power of this model has been investigated in several studies, which Bastos (2010) shows an evidence that QMLE-RFRV can outperform the alternative machine learning algorithm in some out-of-sample predictive evaluation criteria. This model is also one of the common benchmark models to compare the predictive power with other alternatives included in studies by Yao, Crook, and Andreeva (2015); Dermine and De Carvalho (2006); Calabrese and Zenga (2010); Chalupka and Kopecsni (2008); Yang and Tkachenko (2012).

We find that the structure of the model is appropriate and theoretically valid for RR modelling, as the boundary issue is completely eliminated. The model's structure motivate our research in chapter 5, which proposes the nonparametric local logit regression for [0,1] bounded response variable.

Two-sided censored Tobit model

The Tobit model with two-sided censoring at the values of zero and one is applied to address the recovery boundaries in Siao et al. (2015); Bellotti and Crook (2012); Jacobs Jr and Karagozoglu (2011); Tong et al. (2013); Gürtler and Hibbeln (2013); P. Li, Qi, Zhang, and Zhao (2016). Generally, Wooldridge (2010) suggests that the likelihood function of the two-sided censored Tobit model can handle the bounded dependent variable by combining two boundaries (upper bound and lower bound) and the continuous distribution in between. This likelihood suits the properties of the RR, since concentrations at both ends are empirically expected (Bellotti & Crook, 2012). The model employs maximum likelihood with the assumption of an underlying latent variable (Y^*), which assumes:

$$Y = \begin{cases} 0 \ \text{ if } Y^* \leq 0; \\ Y^* \ \text{ if } 0 < Y^* < 1; \\ 1 \ \ \text{ if } Y^* \geq 1, \end{cases} \tag{2.3.4}$$

where $Y^* = x'\beta + e$. Then, the log-likelihood function for the Tobit model is

$$\begin{aligned} ln\,[\mathfrak{L}(\theta|Y_1,..,Y_n)] = \sum_{i=1}^{n} \Big(& 1[Y_i = 0]\, ln[\Phi((-x_i\beta)/\sigma)] + 1[Y_1 = 1]\, ln[\Phi(-(1 - x_i\beta)/\sigma)] \\ & + 1[0 < Y_i < 1]\, ln[(1/\sigma)\phi((Y_i - x_i\beta)/\sigma)] \Big), \end{aligned} \tag{2.3.5}$$

where $\theta = \{\beta, \sigma\}$ is a matrix of unknown parameters, $1[\cdot]$ is the indication function which is equal to one when the condition in the bracket holds, $\Phi(\cdot)$ and $\phi(\cdot)$ are the normal cumulative and density distribution functions, respectively. This likelihood function is constructed by considering the probability of the dependent variable falling between the boundaries and also on each boundary separately throughout the indicator functions (Greene, 2003). Then, the conditional mean of the dependent variable is defined as:

$$\begin{aligned} E(Y_i|x_i) = \ & x_i'\beta\Big[\Phi[(1 - x_i\beta)/\sigma] - \Phi[-x_i\beta/\sigma]\Big] + \sigma\Big[\phi(-x_i\beta/\sigma) - \phi((1 - x_i\beta)/\sigma)\Big] \\ & + \Big[1 - \Phi[(1 - x_i\beta)/\sigma]\Big]. \end{aligned}$$

The parameters and the standard asymptotic inferences are estimated by means of the maximum likelihood estimation in (2.3.5). Gürtler and Hibbeln (2013) explain the censoring issue in the RR as length-biased sampling due to the duration of the workout process. They argue that most RR models use the information regarding the defaulted loan with a completed workout process, while there are some defaulted loans that may not yet have completed the process. This would cause interval censored data which results in bias and a less accurate prediction of the RR. Although the model structure is suitable for RR and addresses the key properties of RR data, Tong et al. (2013); Bellotti and Crook (2012) report that Tobit model has relatively poor predictive performance compared to standard OLS and back-transformation regression.

Sigrist and Stahel (2011) argue that the normal distribution assumption regarding the recovery rate's latent variable Y^* may not be appropriate, and thus they rather assume a gamma distribution. They then extend the censored gamma regression model to a two-tiered gamma model by allowing two different sets of parameters: one for the probability of 1-recovery rate being zero and the other for $0 < 1$-recovery ≤ 1. On the other hand, P. Li et al. (2016) simplify the Tobit model by proposing a two-step estimation: (i) using ordered logistic regression on the probability of the RR falling into three categories, namely $Pr(y_i = 0)$, $Pr(y_i = 1)$, and $Pr(0 < y_i < 1)$; and (ii) using OLS on the observations within the range (0,1). They found that the two-stage estimation outperformed Tobit and the censored gamma regression models.

One might consider the two-sided truncation model is similar to the two-sided censored model. Although, their definitions are somewhat similar, there are some crucial differences. In line with the literature, we assume RR a two-sided censored data. First, the two-sided truncation assumes that the variable beyond the boundaries cannot be observed. As such information is totally unobserved, suitable corrections to account for the observational bias are needed. In Section 2.2.2, we show that it is possible to observe RR that is greater than one or less than zero due to debt collection cost or financial fees. However, the values outside the [0,1] interval are not of interest in practice, as they are driven by banking fees and service charges, rather than the borrower characteristics. Therefore, RR > 1 is treated as one, whereas RR < 0 is treated as zero. By definition, this process leads to two-sided censoring.

Second, by definition, the data outside the boundaries are only partially observed for two-sided censored data, which are restricted to have the boundary values. This process also explains why high proportion of the data are observed at the boundary points. It can be clearly seen that the probability of censored data observed at each of the boundaries will be much higher than those for the truncated data, which may or may not have data at the boundaries. Clearly, the two-sided truncation is not an alternative to two-sided censoring.

2.3.3 Data-driven approaches

The main restrictions in the parametric models are the pre-specified assumptions, including functional form and distribution, which are assumed to be correct. As we find that the empirical RR data is non-standard, the parametric specification, such as normality and linearity, could be misleading assumptions. For example, the presence of nonlinear effects among covariates generally leads to inconsistency and biased estimates in linear models. In fact, these issues could be addressed in parametric models by including a sufficient number of interaction terms, imposing nonlinear parameterisations, and through the discretisation of some continuous variables. However, this remains a difficult task if prior knowledge of the functional form is limited. As the main focus of this thesis is to propose nonparametric regressions, the methods are robust to a non-normal distribution and nonlinearity, which is a special feature of RR data. We also fill the gap in RR model's functional form in chapters 3 and 5 by taking into account of the nonlinearity in the RR-covariate relationship.

Instead of relying heavily on pre-specified assumptions, as required in parametric models, a number of studies have recently introduced more sophisticated models to flexibly address the unique properties of the RR (Altman, 2006). Data-driven approaches require minimal restrictions, allowing the particular model to be shaped by the data rather than intuitions or prior assumptions. Such models may overcome the restrictions in the parametric framework. Moreover, these approaches are expected to have a higher predictive accuracy rate, which is highly attractive to practitioners.

Mixture distribution model

Altman and Kalotay (2014) propose a mixture of Gaussian distribution to estimate the distribution of defaulted debt recovery outcomes. This approach pays significant attention to the bimodal property of the recovery data. The study indicates the likelihood function as:

$$\mathfrak{L}(y_i^*|x_i,\theta,z_i) = \phi(y_i^*;\mu_1,\sigma_1)^{I(z_i\in(c_0,c_1))}\cdots\phi(y_i^*;\mu_m,\sigma_m)^{I(z_i\in(c_{m-1},c_m))}, \tag{2.3.6}$$

where y_i^* is the transformation of the boundary adjusted RR, as previously discussed in (2.3.1), using the inverse Gaussian function, m is total number of finite mixtures, $\theta = \{\mu,\sigma,c\}$, $\phi(\cdot)$ is the probability density function associated with the given mean $\mu \in \{\mu_1,..,\mu_m\}$ and variance $\sigma \in \{\sigma_1,...\sigma_m\}$, $c \in \{c_0,...,c_m\}$ is a cut-off points to assign z_i to a particular mixture component, and $z_i = x_i'\beta + e_i$ is data augmentation of the transformed RR. This study implements a model based on Koop, Poirier, and Tobias (2007), who apply a Gibbs sampling scheme together with the multinomial ordered probit model to estimate the unknown coefficients, the latent variable, and the unrestricted cut-off points.

As the model assumes mixtures of normal distributions on the transformed RR, it can take into account the bimodality property when $m > 1$. On the other hand, the model has several limitations that should be addressed. First, it relies on the back-transformation procedure, which leads to a bias in the model estimates, as discussed earlier. Second, the model may be sensitive to the number of mixtures[2], as this is the main assumption controlling the shape of the distribution. The

[2] Altman and Kalotay (2014) selects the optimum value of m that minimizes the in-sample predictive error

number of parameter estimates and predictive outcomes are dependent on the pre-specified value m. Lastly, interpretations of the estimators would be ambiguous, as they represent the effects of covariates x on z, the latent variable of the transformed RR. Inferring the effects of the conditional information is required which involves several steps (Altman & Kalotay, 2014).

A number of other models in mixture distribution and similar frameworks have been proposed and extended and can address the issues discussed above. Rather than pre-specifying the number of mixtures, Hartmann-Wendels et al. (2014) propose a hybrid finite mixture model, which partitions the data into several subclasses using k-nearest neighbours before applying linear regression on each subclass. They claim that this method can reproduce the multimodality of the recovery density and provide improved predictions. To avoid data transformation, Tanoue, Kawada, and Yamashita (2017) partition the data into two groups, based on whether the RR is zero or not. They then apply a logistic regression on each group. Huang and Oosterlee (2011) propose a generalised linear mixed model using beta distribution with additional normal distributed random effects in the linear predictor to allow for more flexibility than the traditional model does.

As the main advantage of applying mixture models is that they capture the multimodality of the RR density, Zhang and Thomas (2012) argue that the segmentation would be difficult and do not provide additional improvements compared to a single distribution assumption. Calabrese (2012) emphasises the intensive mass at the boundaries using multiple-steps estimation, which modifies the beta likelihood function by incorporating the weights at the boundaries using two logistic regressions. The estimators are expected to capture the negative and positive skewness more accurately than the standard beta regression does. In addition, separated regression models may be problematic when applied for making prediction. Bijak and Thomas (2015) propose a hierarchical model using a Bayesian framework to overcome this issue.

Machine learning algorithms

Machine learning algorithms are the preferable alternative methods for RR modelling and predictions, as they can overcome the misspecification problems in the parametric model. A number of machine learning algorithms in RR modelling have recently been proposed. The neural network (NN) was introduced by Qi and Zhao (2011) and Loterman et al. (2012), while the regression tree (RT), which is a recursively partitioned algorithm, was proposed by Bastos (2010).

(i) Neural networks algorithm

The neural network algorithm is defined as:

$$f(x) = \alpha_0 + \sum_{h=1}^{H} \alpha_h \Lambda(x'\beta_h) + e$$

where $h = 1,...,H$, H is the number of units in a *hidden-layer*, α_h represents a coefficient from the h^{th} hidden-layer unit, Λ is a logit function, and β_h is a $k \times 1$ vector of unknown parameters associated with h^{th} unit. The model's flexibility is mainly driven by the total number H units in the hidden-layer, as this increases the number of the over-imposed parameters by $(k+1)H$. The nonlinear effects can then be captured. Also, the error term has no distribution assumption and can be arbitrarily small if H is sufficiently large.

There are several drawbacks of the models to be concerned. The large H would cause an overfitting problem, in which case the in-sample errors are substantially small but the out-of-sample errors may be large. This is one of the main well-known issues encountered in the application of neural networks (Kalotay & Altman, 2016; Tobback et al., 2014). In addition, although the model has a linear structure, the estimators cannot be meaningfully interpreted. In particular, the parameters are associated with each unit in the hidden layer. This creates a highly complex structure between the covariates and the response variable, the interpretation of which is difficult. It is known as the *black box* problem.

(ii) Regression tree

The regression tree is a recursive partitioning method that uses a searching algorithm (Breiman, Friedman, Stone, & Olshen, 1984). Shalizi (2013) explains that the model sub-divides the data space into smaller subsets, which can interact in complicated and non-linear ways with the specific features of the data structure. This leads to a series of sequential logical if-then conditions to estimate the RR with a tree structure appearance as shown in Figure 2.1. In particular, the RR is estimated by a searching algorithm, which partitions the RR according to the loan characteristics (Bastos, 2010). There is no requirement for data transformation, as the predicted values are always in the unit interval.

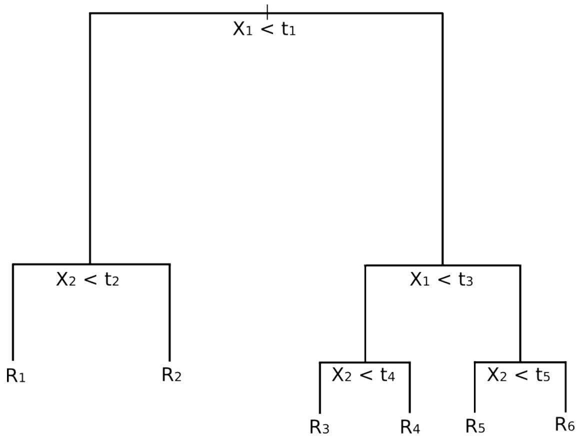


Figure 2.1: *Regression tree structure*

Figure 2.1[3] illustrates an example of the regression tree mechanism, in a case where there are only two continuous variables $\{X_1, X_2\}$. At the top of the tree chart there is a root node which is the full dataset before partitioned by whether $X_1 \leq t_1$, into two binary *daughter nodes*, where t_1 is the cut-off point. This split results from the searching algorithm assigning a condition which minimises the intra-subset variation of the recoveries in the daughter nodes. The conditions include the order

[3] The figure is downloaded from The Beginners Guide to Decision Trees for Supervised Machine Learning

of the explanatory variable and the cut-off values of each split. A number of the nodes are expanded until a further split cannot reduce the variation. The un-split nodes are defined as 'leaves', which determine the fixed levels of the RR, denoted as $R_1, ..., R_6$. These values are the average recovery of the total observations in each leaf, hence, the boundaries of zero and one are not violated.

This algorithm has been applied in many RR modelling studies in the last decade, and seems to be one of the most preferable predictive models. Bastos (2010); Calabrese and Zenga (2010); Hartmann-Wendels et al. (2014); Altman and Kalotay (2014); Kalotay and Altman (2016); Siao et al. (2015); Qi and Zhao (2011); Miller and Töws (2017) indicate that the model's predictive performance outperforms that of parametric models. On the other hand, Tobback et al. (2014) conclude that the regression tree's performance is not as good as that of the back-transformation regressions with high-dimensional covariates.

It is not only the predictive accuracy of the regression tree that attracts practitioners, but also the tree structure, which offers a simple way to infer the effect of RR-covariates through the condition of the binary split in each node. For example, x_1 would have a positive effect on RR if we assume the following assumptions $R_1 < R_2 < ... < R_6$ and $t_1 < t_3$ in Figure 2.1. However, this interpretation of the model may be restricted, as it cannot provide further detailed information regarding the effect of a particular covariate. For example, the nonlinear marginal effect analysis cannot be derived from the regression tree algorithm.

A number of other machine learning algorithms have recently been proposed to predict the RR. Yao et al. (2015) propose a least-squares support vector regression with different intercepts, which allows for the heterogeneity for different types of loans. Nazemi, Fatemipour, Heidenreich, and Fabozzi (2017) propose a fuzzy rule-based model to construct a fuzzy subspace structure. Their study also accommodates macroeconomic conditions by using principal components derived from 104 variables to improve the model's predictability. However, these two algorithms also suffer from the *black box* issue when it comes to valid statistical inference.

As we study nonparametric and semiparametric regression models, they allow us to conduct marginal and interaction effects analysis in all models proposed and implemented in this thesis.

2.3.4 Other statistical approaches for recovery rate modelling

As beta distribution can take into account bimodal density, Jacobs Jr and Karagozoglu (2011) introduce a beta-link generalised linear model (Ferrari & Cribari-Neto, 2004). Although the model has desirable properties relative to alternative linear parametric approaches, its predictability has not yet been investigated. However, Calabrese and Zenga (2010) and Renault and Scaillet (2004) estimate the recovery density using a beta kernel estimation method before conducting a hypothesis test of the appropriateness of theoretical beta distribution for RR modelling. Renault and Scaillet (2004) conclude that RR does not follow a beta distribution.

Given that RR can be calculated as:

$$1 - \frac{\text{Expected loss}}{\text{Exposure at default}},$$

Leow and Mues (2012) apply two OLS models to estimate the expected loss and the exposure at default separately. Tong et al. (2013) propose a zero inflated gamma model to estimate the expected loss. Their models can address the clustering of the full RRs.

Recently, a conventional quantile regression (Koenker & Bassett, 1978) has been applied for RR modelling. Siao et al. (2015) apply a linear quantile regression with a logit back-transformation technique for prediction. They estimate regressions at various quantiles of the RR distribution. Then, the quantile that yields the best in-sample predictive accuracy is employed to predict the out-of-sample recovery. On the other hand, Krüger and Rösch (2017) apply a standard linear quantile regression and allow the RR to be outside the boundaries zero and one. They find that the linear effects of the recovery covariates vary across the different quantiles. Furthermore, they introduce an alternative application of quantile regression to estimate the downturn recovery as well as the value at risk (VaR). In chapter 4, we propose the non- and semi-parametric quantile regressions to identify the presence of potential nonlinearity and heterogeneity in the effect of RR-covariates on RR at various quantiles, which has not been investigated before in the literature. We also employ the model selection criteria proposed in their paper, including the distribution fit measurement and its hypothesis testing as well as the VaR framework, to evaluate our proposed quantile regression models.

The dependency between the RR and the probability of default has been discussed in several papers (Altman, 2006; Frye, 2000; Bruche & González-Aguado, 2010; Fischer, Köstler, & Jakob, 2016). A concern arises because a negative relationship between these factors is expected, especially during an economic downturn, as both risk parameters are sensitive to turmoil in the adverse macroeconomic conditions (Dermine & De Carvalho, 2006; Acharya et al., 2007; Qi & Yang, 2009; Grunert & Weber, 2009). Specifically, the probability of default tends to increase while the RR seems to decrease during economic downturns. Rösch and Scheule (2014) suggest a joint estimation for forecasting probability of default and the RR. They propose a likelihood function that can take this issue into account. The study concludes that without the dependency of these two credit risk parameters, the capital requirement could be underestimated by 17%. Alternatively, Frontczak and Rostek (2015) assume that the RR depends on the actions taken immediately after default until resolution. They model the RR with an exponential Ornstein-Uhlenbeck diffusion in order to capture the stochastic process, which represents the workout recovery process.

Given the importance of the RR in credit risk management, a series of actuarial studies have discussed the tail risk as well as the limiting distribution of the RR (Yuan, 2016). Tang and Yuan (2013) estimate the random recoveries from the borrower at the time of default, and Wei and Yuan (2016) model the heavy tails underlying the risk factors of a low-default portfolio with weak contagion by employing a Sarmanov distribution with regularly varying tailed marginal distributions. On the other hand, Bonini and Caivano (2016) allow the workout experts' opinion to be embedded in a particular predictive model, which is common in actuarial modelling, to improve the accuracy rate of the estimates.

2.4 Conclusion

In the review of the literature on modelling RR, we have identified that although there are large number of proposed models to overcome problems associated with modelling RR-covariate relationships, the models have not adequately accommodated the special features of RR data. Moreover, lack of flexibility in conducting nonlinear marginal effects analysis, which are essential to design an appropriate treatment program and financial policy in order to mitigate the credit risk exposure, remains unsolved. The research gaps we identified together with a discussion of the limitation of the existing models provided motivations to the research endeavor of this thesis.

In particular, we found several limitations in the existing literature. Firstly, it was common to employ the *transformation and back transformation* technique to accommodate the boundary [0,1] of the RR, although it caused bias to the RR estimates. This technique was used in several studies, as parametric assumptions that could generate continuous bounded [0,1] respond variable estimate are limited. One of the main aims of thesis is to eliminate the bias completely. The proposed nonparametric and semiparametric regression models in the following chapters can address the stylized fact of the empirical RR in the data-driven manner using kernel estimation approach, which do not heavily rely on the *back-transformation* and other pre-determined parametric assumptions. The highly flexible features of the proposed models in chapters 3 and 4 do not require the restricted assumptions of most parametric models. Furthermore, chapter 5 proposes the nonparametric regression specifically for fractional response variable, which is suitable to accommodate the boundary property of the RR.

Secondly, RR model is expected to be nonlinear, but further analysis on the nonlinear relationships and nonlinear functional form of parametric models are overlooked. Most parametric models relied intensively on linear functional form, which might lead to misspecification issue if prior knowledge is not available. Chapters 3 and 5 aim to improve the functional form of RR models, which will mitigate the misspecification issue.

Thirdly, to accommodate nonlinearity in the RR-covariate relationship, the data driven machine learning algorithms were proposed in last decades. However, the unexplained *black box* issue remained a major concern of their applications, which also prevented the understanding of the complex effect of the covariates on RR. In this thesis, we pay considerable attention to explain the marginal and interaction effects on conditional mean of RR in chapter 3 and 5, and on various condition quantiles of RR distribution in chapter 4.

Chapter 3

Non- and Semi-parametric conditional mean regressions for recovery rate modelling

3.1 Introduction

In chapter 2, the stylized facts of empirical RR were outlined. Emphasis was placed on the [0,1] bounded RR with a bimodal density and nonlinearity in the effect of its covariates. Also, we found evidence suggesting that data-driven models, such as machine learning algorithms, were more able than linear parametric models to improve RR predictions. However, although machine learning algorithms offer better predictive performance, they struggle to provide a suitable explanation of how such a high performance is achieved, which is known as the black box problem. This leads to a lack of understanding of the nonlinear relationships between RR and its covariates. Moreover, despite the evidence indicating the presence of this nonlinear relationship, no consideration has been given to modelling such nonlinearity in the parametric settings (Gürtler & Hibbeln, 2013; Bastos, 2010; Hartmann-Wendels et al., 2014).

In this chapter, we propose a nonparametric regression with the local constant (LC) and local linear (LL) estimation methods, as well as a semiparametric partially linear (PL) regression, to model RR-covariate relationship. These data-driven models can capture the nonlinearity and, more importantly, transparently explain relationships between covariates and RR, which have not been much explored in the literature. We also address the boundary [0,1] requirement of the empirical RR by proposing a two-sided censored nonparametric regression using the local linear estimation method (LL2)[1], which is an extension of the one-sided censored nonparametric regression (Lewbel & Linton, 2002). We also conduct an extensive comparison of our proposed models and the existing models discussed in chapter 2, which include inverse Gaussian *back-transformation* regression (IG), quasi maximum likelihood regression for fractional response variables (QMLE-RFRV), the two-sided censored Tobit model (TOBIT), the mixture distribution model (MM) proposed by Altman and Kalotay (2014), and the machine learning regression tree algorithm (RT). The comparison is made in terms of in-sample and out-of-sample predictability of models.

The nonparametric regression is a full data-driven approach, which does not require any pre-specified assumptions. Our main findings highlight the application of LL method to analyse the nonparametric marginal effect estimates. These LL estimators are useful to uncover the underlying relationships between RR and its covariates. On the other hand, the semiparametric PL requires the user to specify the model's functional form, which is a combination of linear and nonlinear functions. This leads to a dimensional reduction, which eases the computational burden in the nonparametric regression. To specify the semiparametric model, we utilise the insight provided by the marginal effect analysis based on LL method

[1] The both simulation and empirical studies are conducted, which produce some promising results.

to improve the functional form of the semiparametric PL model. We find that the application of PL leads to outstanding out-of-sample predictive performance compared to that of nine other models.

This chapter is organised as follows: sections 3.2 and 3.3 discuss our proposed non- and semi-parametric regressions, respectively. This is followed by the data description and a preliminary analysis in section 3.4. Then, the empirical results are reported and analysed in section 3.5 before concluding.

3.2 Nonparametric regression

In this section, we introduce the local constant and local linear kernel estimation methods. These two methods do not require pre-determined distribution assumptions or functional forms. In addition, Racine and Li (2004) propose a nonparametric estimation in the presence of a combination of continuous and discrete covariates by assigning different types of kernel functions. These make the method more suited to model the RR-covariate relationship, which commonly has both continuous and discrete/categorical determinants.

Let us define

$$X_i = (X_i^c, X_i^d),$$

where the continuous regressors with p dimensions are $X_i^c \in \mathbf{R}^p$, and the remaining regressors X_i^d are a $q \times 1$ vector of categorical variables. For any t^{th} component of X_i^d as $t = 1, ..., q$, each component can take a discrete value as $X_{t,i}^d \in \{0, 1, ..., c_t - 1\}$, where $c_t \geq 2$ is a total category of $X_{t,i}^d$, such as $c_t = 2$ for a dummy variable. Hence, a nonparametric regression is given by

$$Y_i = g(X_i) + u_i, \quad i = 1, ..., n, \tag{3.2.1}$$

where Y_i is the RR as the dependent variable, and $g(\cdot)$ has an unknown functional form. We explore general relationships relying on the smoothing process using the kernel functions to estimate the conditional mean, which is defined as:

$$E(Y|X=x) = \frac{\int y \cdot f(x,y)dy}{f(x)} \equiv g(x), \tag{3.2.2}$$

where $f(x,y)$ is the joint density and $f(x)$ is the marginal density. To estimate this, we need to define the kernel functions and their bandwidth before applying local constant or local linear estimation methods.

3.2.1 Kernel functions

Our study employs three different kernel functions: a Gaussian kernel function for continuous variables and kernel functions for ordered and unordered categorical variables. These three functions are discussed in what follows before the general form of the kernel estimations is provided. Gaussian kernel function is the most common for estimating the semi- and non-parametric regression because its support is $-\infty < 0 < \infty$, while other kernel functions such as Rectangular, Epanechnikov, Biweight, Triangular have limited supports (DiNardo & Tobias, 2001). On the other hand, the kernel function for discrete variables are limited.

Gaussian kernel function

The standard Gaussian kernel function is employed for any continuous variable (X_i^c), which is denoted as:

$$\kappa_s\left(X_{s,i}^c, x_s^c, h_s\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{X_{s,i}^c - x_s^c}{h_s}\right)^2\right), \tag{3.2.3}$$

where $s = 1,..,p$, $\kappa(\cdot)$ is the Gaussian kernel function, and h_s is a bandwidth associated with the s^{th} continuous variable.

Kernel function for categorical variables

The choice of kernel function depends on whether the categorical variable has a natural ordering or not. First, the following kernel function proposed by Racine and Li (2004) is applied for any categorical variable with no intrinsic ordering of the categories:

$$\lambda_t(X^d_{t,i}, x^d_t, l_t) = \begin{cases} 1, & \text{if } X^d_{t,i} = x^d_t, \\ l_t, & \text{otherwise,} \end{cases} \tag{3.2.4}$$

where we assume that the t^{th} categorical variable is unodered, and l_t is the bandwidth associated with $\lambda_t(\cdot)$ which has a range of $[0,1]$.

On the other hand, if there is evidence indicating that the t^{th} categorical variable has a natural ordering, then we apply:

$$\lambda_t(X^d_{t,i}, x^d_t, l_t) = l_t^{|X^d_{t,i} - x^d_t|} \tag{3.2.5}$$

where we assume that the t^{th} variable has ordinal categories.

Kernel estimation of a probability density function

One of the main applications of these kernel functions is to estimate a probability density function. Given that the kernel function discussed is appropriately assigned to each variable, the joint density and marginal density in (3.2.2) can be estimated as follows:

$$\hat{f}(x,y) = \frac{1}{n \cdot h_0 \cdot h_1 \ldots h_p \cdot l_1 \ldots l_q} \sum_{i=1}^{n} \kappa_0(Y_i, y, h_0) \prod_{s=1}^{p} \kappa_s\left(X^c_{s,i}, x^c_s, h_t\right) \prod_{t=1}^{q} \lambda_t(X^d_{t,i}, x^d_t, l_t); \text{ and}$$
$$\hat{f}(x) = \frac{1}{n \cdot h_1 \ldots h_p \cdot l_1 \ldots l_q} \sum_{i=1}^{n} \prod_{s=1}^{p} \kappa_s\left(X^c_{s,i}, x^c_s, h_t\right) \prod_{t=1}^{q} \lambda_t(X^d_{t,i}, x^d_t, l_t). \tag{3.2.6}$$

3.2.2 Local constant estimation method

The conditional mean in (3.2.2) can be estimated by replacing the estimates in (3.2.6). Then, we can estimate $E(Y|X = x)$ by:

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{K}(X_i, x, H)}{\sum_{i=1}^{n} \mathbb{K}(X_i, x, H)} \tag{3.2.7}$$

where $\mathbb{K}(X_i, x, H) = \prod_{s=1}^{p} \kappa_s \left(X_{s,i}^c, x_s^c, h_t \right) \prod_{t=1}^{q} \lambda_t (X_{t,i}^d, x_t^d, l_t)$ is product of the kernel functions, and $H = \{h_1, ..., h_p, l_1, ..., l_q\}$ is a collection of bandwidths associated with each component in X_i. It is worth noting that as $\mathbb{K}(\cdot) > 0$, if the response variable is bounded between zero and one, then the local constant estimate in (3.2.7) will also lie within the same interval. This is an additional advantage of the method for RR modelling as it accommodates the boundary property of the empirical data.

In (3.2.7), the size of the bandwidths in H plays a critical role in estimating the $\hat{g}(\cdot)$. Our study employs the least-squares cross-validation method to select the appropriate bandwidth. We select the bandwidth H by minimising the following objective function:

$$CV_{LC}(H) = \sum_{i=1}^{n} (Y_i - \hat{g}_{-i}(X_i))^2 \tag{3.2.8}$$

where $\hat{g}_{-i}(X_i)$ is the leave-one-out kernel estimator of $g(X_i)$, which is defined as:

$$\hat{g}_{-i}(X_i) = \frac{\sum_{i \neq j, j=1}^{n} Y_j \mathbb{K}(X_i, X_j, H)}{\sum_{i \neq j, j=1}^{n} \mathbb{K}(X_i, X_j, H)}. \tag{3.2.9}$$

In particular, we estimate the unknown function $g(X_i)$ by utilising all information without the i^{th} observation. Then, the bandwidths are selected such that (3.2.8) is minimised.

Given the bandwidth H^*_{LC} selected from the least-squares cross-validation method, we can estimate the conditional mean by executing (3.2.7), which is equivalent to the solution of the following minimisation:

$$\min_{\alpha}(x) \sum_{i=1}^{n} (Y_i - \alpha(x))^2 \mathbb{K}(X_i, x, H^*). \tag{3.2.10}$$

Specifically, the constant $\alpha(x)$ is obtained to approximate the unknown $g(x)$ in the neighbourhood of x, as we use the local average of Y_i's to estimate $g(x)$, which is defined as the local constant kernel estimator.

In fact, it is true that there are numbers of bandwidth selection criteria available in the literature. They include rule-of-thumb, least squares cross-validation, likelihood cross-validation, biased cross-validation, plug-in approach, smoothed bootstrap, binning, among others (Jones, Marron, & Sheather, 1996; Zambom & Dias, 2012). The least squares cross validation has become popular because of its attractive feature that it automatically excludes irrelevant variables (Hall, Racine, & Li, 2004; Q. Li & Racine, 2007). It assigns large optimal bandwidths to the irrelevant variables, which over-smooth the variables towards the uniform distribution regardless of the respective marginal distributions. As a result, the irrelevant covariates are removed through the cross-validation method, which enhances the reliability of the empirical results.

3.2.3 Local linear estimation method

Although the LC estimation method can accommodate the [0,1] boundary property, it is difficult to carry out the marginal effect analysis of the LC estimate. This leads to the application of the LL estimation method, which provides the first derivative estimate of the unknown function $g(x)$ with respect to x or the marginal effect estimate. By directly estimating the marginal effect, the model can reveal the underlying relationships without the prior knowledge of the functional form.

This method may provide additional insights into RR structure which have not yet been documented.

The local linear estimator is obtained by minimising the following objective function:

$$\min_{\{a(x),b(x)\}} \sum_{i=1}^{n} (Y_i - a(x) - (X_i - x)'b(x))^2 \, \mathbb{K}(X_i, x, h), \tag{3.2.11}$$

where $\hat{b}(x)$ is the LL estimator, which is a consistent estimator of $\frac{\partial g(x)}{\partial x}$, and $\hat{a}(x)$ is an estimate of $g(x)$.

The solution of minimising (3.2.11) can be seen as local least squares in matrix form as:

$$\min_{\delta(x)} (\mathcal{Y} - \mathcal{X}\delta(x))'\mathcal{K}(x)(\mathcal{Y} - \mathcal{X}\delta(x)), \tag{3.2.12}$$

where $\delta(x) = (a(x), b(x)')'$, $\mathcal{Y}$ is the $n \times 1$ vector of i^{th} component Y_i, $\mathcal{X}$ is the $n \times (1 + (p + q))$ matrix, and $\mathcal{K}(x)$ is the $n \times n$ diagonal matrix having i^{th} diagonal element of $K(X_i, x, h)$. Thus, $\delta(x)$ can be estimated by standard generalised least squares:

$$\hat{\delta}(x) = (\mathcal{X}'\mathcal{K}(x)\mathcal{X})^{-1}(\mathcal{X}'\mathcal{K}(x)\mathcal{Y}),$$

which provides the consistent estimators and the asymptotic normality of $\hat{\delta}(x)$. As the LL method employs the local least squares method, there is no guarantee that $\hat{a}(x)$ will be bounded between zero and one, as found in the LC estimation in (3.2.7). We attempt to solve this boundary issue of LL method in section 3.2.4.

Similarly to the LC estimation, the bandwidth H plays a vital role. Hence, least-squares cross-validation is also applied to determine the appropriate bandwidths:

$$CV_{LL}(H) = \sum_{i=1}^{n} [Y_i - \hat{a}_{-i}(X_i)]^2, \tag{3.2.13}$$

where $\hat{a}_{-i}(X_i)$ is the leave-one-out[2] LL kernel estimator of $g(X_i)$ from (3.2.12). Finally, we can estimate the conditional mean and the marginal effects from (3.2.11), given the selected bandwidth H^*_{LL} from the solution of (3.2.13).

3.2.4 Nonparametric regression with two-sided censoring

We propose a nonparametric regression model with LL estimation method which allows two-sided censoring at zero and one. Before introducing this model, we examine the nonparametric regression with one-sided censoring proposed by Lewbel and Linton (2002). The estimation method is computationally convenient, requiring only two nonparametric regressions and a univariate integral approximation. We then extend the model to that for two-sided censored data to address the boundary issue in the LL method.

To introduce notations, we define:

$$Y_i = \begin{cases} 1, & if \;\; Y^* \geq 1 \\ Y^*, & if \;\; 0 < Y^* < 1 \\ 0, & if \;\; Y^* \leq 0, \end{cases} \tag{3.2.14}$$

where Y^* is an unobserved latent variable. The equation (3.2.14) is equivalent to:

$$Y_i = \max(0, \min(w(X_i) - e), 1), \tag{3.2.15}$$

where Y_i is the observed response variable, which can be written as $Y^* I(0 \leq Y^* < 1) + I(Y^* \geq 1)$, $I(\cdot)$ is the indicator function, $Y_i^* = w(X_i) - e_i$ is unobserved, X_i is the set of covariates which can be multidimensional and contain both discrete and continuous variables, the unknown function w is differentiable and has finite derivatives, the error e is independent of x with an absolutely continuous distribution function $F(e)$ and Lebesgue density function $f(e)$. We want to estimate

[2] defined in (3.2.9)

the latent function $w(X_i)$, and we can then ensure that Y_i is in the unit interval as shown in (3.2.14).

Let us define:

$$\begin{aligned} &Pr(Y=0|X=x) = Pr(w(x)-e \leq 0) = 1-F(w(x)), \\ &Pr(Y=1|X=x) = Pr(w(x)-e \geq 1) = F(w(x)-1), \\ \text{then,}\quad &Pr(0<Y<1|X=x) = F(w(x)) - F(w(x)-1). \end{aligned} \tag{3.2.16}$$

We also have:

$$\begin{aligned} E(Y|X=x) &= E[Y^{*}I(0 \leq Y^{*} < 1)] + E[I(Y^{*} \geq 1)] \\ &= \int_{w(x)-1}^{w(x)} (w(x)-e)dF(e) + F(w(x)-1) \\ &= w(x)\int_{w(x)-1}^{w(x)} dF(e) - \int_{w(x)-1}^{w(x)} edF(e) + F(w(x)-1) \\ &= w(x)[F(w(x)) - F(w(x)-1)] - [eF(e)]_{w(x)-1}^{w(x)} + \mathcal{F}(w(x)) \\ &\quad - \mathcal{F}(w(x)-1) \\ &= \mathcal{F}(w(x)) - \mathcal{F}(w(x)-1), \end{aligned} \tag{3.2.17}$$

where $\mathcal{F}(w) = \int_{-\infty}^{w} F(e)de$ is the integrated cumulative density function. Given (3.2.16) and (3.2.17), we define $r(x) = E(Y|X=x)$ and $q(r(x)) = E(I(0<Y<1|r(X)=r))$. Let

$$\mathcal{G}(w) = \mathcal{F}(w) - \mathcal{F}(w-1),$$

where $\mathcal{G}$ is a monotonically increasing function and invertible with $\mathcal{G}'(w) = F(w) - F(w-1) \geq 0$ for all w and with strict inequality for some ranges. Then we can define

$$\begin{aligned} &q(r(x)) = \mathcal{G}'(\mathcal{G}^{-1}(r(x))), \\ \text{and}\quad &r(x) = \mathcal{G}(w). \end{aligned} \tag{3.2.18}$$

Given these assumptions, we can follow the theorem 2 in Lewbel and Linton (2002), which shows that:

$$w(x)+k=\gamma_0-\int_{r(x)}^{\gamma_0}\frac{1}{q(r)}dr, \tag{3.2.19}$$

for some location constant $k(\gamma_0)$, where γ_0 is any nonnegative constant. If $\gamma_0 < r(x)$ then integrals of the form $\int_{r(x)}^{\gamma_0}$ above are to be interpreted as $-\int_{\gamma_0}^{r(x)}$. The equation (3.2.19) can be proved by using the change of variable $r=\mathcal{G}(w)$, $dr=\mathcal{G}'(w)d(w)$, and $q(r)=\mathcal{G}'(\mathcal{G}^{-1}[\mathcal{G}(w)])=\mathcal{G}'(w)$. Then,

$$\begin{aligned}\gamma_0-\int_{r(x)}^{\gamma_0}\frac{1}{q(r)}dr &= \gamma_0-\int_{\mathcal{G}^{-1}(\mathcal{G}[w(x)])}^{\mathcal{G}^{-1}(\gamma_0)}\frac{1}{\mathcal{G}'(w)}\mathcal{G}'(w)dw\\ &= \gamma_0-\int_{w(x)}^{\mathcal{G}^{-1}(\gamma_0)}1dw\\ &= \gamma_0-\mathcal{G}^{-1}(\gamma_0)+w(x),\end{aligned} \tag{3.2.20}$$

where $k=\gamma_0-\mathcal{G}^{-1}(\gamma_0)$, and the equation (3.2.19) holds. If we choose γ_0 such that $q(\gamma_0)=1$, then $\gamma_0=\mathcal{G}^{-1}(\gamma_0)$ (Lewbel & Linton, 2002).

We estimate (3.2.19) as the following steps:

(i) estimate $\hat{r}(x)$ using nonparametric regression with LL estimation method of Y on X as discussed in section 3.2.3,

(ii) estimate $\hat{q}(r)$ using one-dimensional nonparametric regression of $I(0<Y<1)$ on $\hat{r}(X)$,

(iii) estimate $\hat{w}(x)$ given the fact that

$$\hat{w}(x)=\gamma_0-\int_{\hat{r}(x)}^{\gamma_0}[\frac{1}{\hat{q}(r)}]dr,$$

for some constant γ_0. We also apply the Trapezoidal integral approximation in step (iii) to estimate the unknown function $w(x)$,

(iv) Then, the prediction is generated as $\max(0, \min(\hat{w}(x), 1))$

In Appendix A, we provide the simulation result of the proposed model, which shows that the model works well to estimate the unknown function $w(\cdot)$.

3.3 Semiparametric partially linear regression

The semiparametric partially linear model allows a combination of parametric linear and non-parametric components. The model partitions the vector of dependent variables X_i as follows: Z_i^b is a vector of independent variables assumed to have linear functional form, and Z_i^m are the remaining regressors with unspecified functional form (Robinson, 1988; Härdle, Liang, & Gao, 2012). Then, the general form of the model is written as:

$$Y_i = Z_i^{b'}\beta + m(Z_i^m) + u_i, \quad i = 1, \ldots, n \tag{3.3.1}$$

where Z_i^b and Z_i^m can be vectors containing both continuous and categorical variables, β is an unknown parameter vector associated with Z_i^b, and $m(\cdot)$ is an unknown function of Z_i^m. The model specification in (3.3.1) requires pre-specified functional form to allocate the covariates into either linear function in the first component or an unknown function in the second component. The choice between a nonlinear and linear functions of an independent variable in the model largely depends on a prior knowledge, theory and/or empirical findings. The inclusion of the variables in the correct functional forms improves the PL model specification; see (Härdle et al., 2012) for some empirical examples.

Estimation of parametric component

The unknown parameters β can be estimated by taking the conditional expectation with respect to Z_i^m on (3.3.1) as $E(Y_i|Z_i^m) = E(Z_i^b\beta|Z_i^m)$, then subtracting (3.3.1):

$$Y_i - E(Y_i|Z_i^m) = (Z_i^b - E(Z_i^b|Z_i^m))'\beta + u_i$$

By employing a least-squares method, we estimate:

$$\hat{\beta} = \left[\sum_{i=1}^{n} \tilde{Z}_i\tilde{Z}_i'\right]^{-1} \sum_{i=1}^{n} \tilde{Z}_i\tilde{Y}_i, \tag{3.3.2}$$

where $\tilde{Z}_i = Z_i^b - E(Z_i^b|Z_i^m)$ and $\tilde{Y}_i = Y_i - E(Y_i|Z_i^m)$. Importantly, due to the identification problem, the intercept term should not be included in the parametric component.

Since the conditional expectation $(E(\cdot|Z_i^m))$ in (3.3.2) is unknown, it can be consistently estimated by a kernel method:

$$\begin{aligned} \hat{E}(Y_i|Z_i^m) &= \frac{\sum_{j=1}^{n} Y_j \mathbb{K}(Z_i^m, Z_j^m, H)}{\sum_{j=1}^{n} \mathbb{K}(Z_i^m, Z_j^m, H)}, \\ \hat{E}(Z_i^b|Z_i^m) &= \frac{\sum_{j=1}^{n} Z_j^b \mathbb{K}(Z_i^m, Z_j^m, H)}{\sum_{j=1}^{n} \mathbb{K}(Z_i^m, Z_j^m, H)}, \end{aligned} \tag{3.3.3}$$

where $\mathbb{K}(Z_i^m, Z_j^m, H)$ is a product of the kernel functions[3] defined in (3.2.7) associated with Z_i^m, and the least-squares cross-validation in (3.2.8) is applied to select the bandwidths for each conditional expectation estimate in (3.3.3). Since the estimation process in (3.3.3) involves a random denominator, a trimming function

[3] Each variable in the vector Z_i^m could be either continuous or discrete, then the kernel function should be properly assigned as discussed in Section 3.2.1

is suggested to construct an asymptotic distribution:

$$\hat{\beta} = \left(\sum_i (Z_i^b - \hat{E}(Z_i^b|Z_i^m))(Z_i^b - \hat{E}(Z_i^b|Z_i^m))' \right)^{-1} \sum_i (Z_i^b - \hat{E}(Z_i^b|Z_i^m))(Y_i - \hat{E}(Y_i|Z_i^m))1_i \tag{3.3.4}$$

where $1_i = 1$ if $\hat{f}(Z_i^m) = \sum_{j=1}^{n} \mathbb{K}(Z_i^m, Z_j^m, H)$ is less or equal to ϵ, and $1_i = 0$ otherwise, and $\epsilon = \epsilon_n \geq 0$ is a trimming parameter which satisfies $\epsilon_n \to 0$ as $n \to \infty$. By using the trimming function, the small $\hat{f}(Z_i^m)$, which can cause technical difficulties, is removed.

Estimation of nonparametric component

Given the parameter estimate in (3.3.4), the unknown function $m(\cdot)$ can be nonparametrically estimated. The nonparametric component is denoted by rearranging (3.3) as;

$$m(Z_i^m) = E(Y_i - Z_i^{b'}\beta|Z_i^m)$$

As the unknown parameter in the linear component is estimated in (3.3.4), it allows us to consistently estimate the nonparametric component (Gao, Liu, & Racine, 2015), which is given by:

$$\hat{m}(z^m) = \frac{\sum_{i=1}^{n} (Y_i - Z_i^{b'}\hat{\beta})\mathbb{K}(Z_i^m, z^m, H)}{\sum_{j=1}^{n} \mathbb{K}(Z_i^m, z^m, H)} \tag{3.3.5}$$

Then, the smooth parameter estimators are chosen to minimise use of the least-squares cross-validation:

$$CV_{PL}(H) = \sum_{i=1}^{n} \left(Y_i - Z_i^b\hat{\beta} - \hat{m}_{-i}(Z_i^m) \right)^2 \tag{3.3.6}$$

where $\hat{m}_{-i}(Z_i^m)$ is a leave-one-out kernel estimator of $m(Z_i^m)$.

The two-step estimation of PL regression eases the computational burden, as it reduces the dimensions for the nonparametric estimation so that they are less than $p+q$. This overcomes the computational difficulty in nonparametric regressions discussed in the previous section. It also mitigates the curse of high dimensional problem encountered in nonparametric regressions, which require sufficiently large number of observations for the estimation of the model.

3.4 Data and preliminary analysis

In this section, we summarise the empirical RR data which are employed in the thesis. This includes a preliminary data summary, which provides definitions of the variables included as well as the intuitive and expected effects on the RR. This provides an overview of the overall picture and highlights the specific stylised nature of the data.

3.4.1 Data description and summary statistics

The dataset on realised RR was obtained from the Moody's Ultimate Recovery Database, which has been used in several studies, such as Qi and Zhao (2011), Altman and Kalotay (2014) and Siao et al. (2015), among others. Our data has 3,573 cross-sectional recovery rates from US corporate loans that were defaulted on and bankrupted between 1994 and 2012. The rates are discounted nominal rates by the date that the last interest rate was paid. Moody's also provides the debt characteristics prior to default, including the debt cushion[4] (DC), the instrumental rank (Rank) in capital structure, the types of the defaulted loan (Type), and collateral status (Col).

[4] Moody's defines debt cushion as the ratio of the face value of a claim to the total debt below it. The high DC reflects the low outstanding debt in the company capital structure.

To take into account of economic risk, we obtained the St. Louis Fed Financial Stress Index (SI) provided by the US federal reserve bank of St. Louis[5]. It measures the degree of financial stress in the markets, which moves according to the economy. The average value of the index is designed to be zero in late 1993, where positive stress suggests above-average financial market stress. Stress index is estimated using 17 key indices, such as the federal funds rate, corporate credit risk spread, interest rate, and inflation (Kliesen, Smith, et al., 2010). As Moody's provides the date of default of each loan, the stress index is matched with this date to reveal the economic condition at each loan's time of default.

[———— Insert [Figure 3.1] here ————]

Overall, the RRs are bounded between zero and one, representing a permanent loss and a full recovery, respectively. High percentages at both ends are observed: 6% at zero and 31% at one. This forms the [0,1] bounded data with bimodal density at the boundaries of zero and one. The sample density is shown in Figure 3.1. The average and median RR are 0.55 and 0.58, respectively.

In terms of the RR covariates, there are five variables in total, which include two continuous variables: DC and SI, and three categorical variables: Type, Rank, and Col. For the continuous variables, Figure 3.2a shows the density of DC, where the average DC is 0.24, and 46% of the total observed defaulters have a DC of zero.

Figure 3.2b represents the density of SI, reflecting the economic conditions at the time of default. We observe that SI is mostly between -1 and 1, while $SI > 1$ reflects the extremely stressed economy observed during financial crises, such as GFC. Furthermore, we observe that SI $\geq$ 1.5 is observed only during the GFC period. Figure 3.3 shows the movement of the SI between 1994 and 2012, which indicates several economic events such as the dot-com crisis (1999-2003), economic expansion (2004-2006), as well as the GFC (2007-2010). Figure 3.4

[5] Federal Reserve Bank of St. Louis, St. Louis Fed Financial Stress Index [STLFSI], retrieved from FRED, Federal Reserve Bank of St. Louis: https://fred.stlouisfed.org/series/STLFSI

compares the movements of the annual averages of SI and RR from 1994 to 2012. It shows that the average RR increases when SI is negative and decreases during positive SI periods. We find that the sample RR average is 0.56, while the averages during the dot-com and GFC crises are 0.50 and 0.53, respectively. This implies that RR is likely to be lower during periods of deeper financial stress. Therefore, a negative effect of SI on RR is expected.

As we include three categorical variables: first, there are four instrumental ranks (Rank = {1,2,3,4}) where the lower Rank represents the higher priority in the capital structure of a particular defaulted borrower. Second, we consider six types of loan which include two commercial loans including term and revolving loans ($Type^{(1)}$, and $Type^{(2)}$, respectively), and four corporate bonds: senior secured bond ($Type^{(3)}$), senior subordinate bond ($Type^{(4)}$), senior unsecured bond ($Type^{(5)}$), junior and subordinate bonds ($Type^{(6)}$). Lastly, the collateral status is a dummy variable, where $Col = 0$, and 1 represent a loan with and without collateral, respectively.

[————— Insert [Figures 3.2 and 3.3] here—————]

3.4.2 Preliminary analysis of recovery rates

The preliminary analysis of the effects of all five covariates on RR is provided in Table 3.1 as a contingency table. Specifically, the RR is dissected based on the information provided in the first column. We then report the sample average and quantiles[6] of each conditional RR in columns 5-9.

[————— Insert [Table 3.1] here —————]

Panel A(i) of Table 3.1, type of loan, shows that the data are 42% commercial loans and 58% bonds. Considering the average RR of each type, revolving loans have the highest average RR, while junior and subordinate bonds are the most

[6] We consider 0.05, 0.25, 0.5, 0.75 and 0.95 quantiles

risky loan types, with the lowest average RR of 0.24. The sample RR densities of the commercial loans and senior secured bonds tend to have more negative skewness than the remaining loans, due to the relatively high medians and high masses at the upper quantiles.

Panel A(ii), the instrumental rank indicates the repayment priority in the capital structure. If the collateral was liquidated to repay the borrower's defaulted debts, Rank 1 loans would be repaid first, followed by other ranks in ascending order. Therefore, we find that the average RRs decrease as Rank increases. Also, as most commercial loans generally have Rank 1, the table shows that the RR densities of Rank 1 are similar to those of the commercial loans.

Lastly, Panel A(iii), collateral is one of the main sources of funds for repaying outstanding defaulted debt. Thus, collateralised loans are intuitively less risky than uncollateralised loans. Table 3.1 shows that collateralised loans have substantially higher average RRs than uncollateralised loans. Fifty percent of the uncollateralised loans recover less than 20% of the total loss, while more than half of collateralised loans recover more than 90%.

For the continuous variables, Panel B(i) of Table 3.1 partitions the empirical RR into three subsamples based on levels of DC: DC = 0, $0 < \text{DC} \leq 0.5$, and DC > 0.5. The results show that the RR of zero-DC defaulted loans, which make up 46% of the total observations, has an average of 0.4. The average increases to 0.88 for loans with DC > 0.5, where more than half have a full recovery rate, as the median RR is 1. Our findings suggest a positive effect of DC on RR.

For the effect of SI, RR is partitioned into three subsamples according to the levels of SI in Panel B(ii) of Table 3.1. The levels are denoted as negative SI, $0 < \text{SI} < 1$, and $\text{SI} \geq 1$, which represent low-, high-, and substantially high-stress periods, respectively. Recovery rate during the low-stress period is lowest at 0.7, while the rates are similar at approximately 0.5 for high- and substantially high-stress periods. During good economic conditions, more than half of the defaulted loans can be recovered to more than 80% of the total loss, compared to 40% for the other periods. We then expect a negative effect of SI on RR. It can also be noted that the densities of RR during high and extremely high economic stress periods are similar, as the RRs at all given quantiles of both periods are approximately equivalent in the columns 5-9 of Panel B(ii).

3.5 Empirical results

The proposed non- and semi-parametric regressions are applied to RR data with its five determinants discussed in the previous section. The estimation results focusing on the marginal effect analysis are discussed. This is followed by a comparison of the predictive accuracy of the proposed models and that of the existing models, namely IG, QMLE-RFRV, TOBIT, MM, and RT (these abbreviations are defined in Section 3.1).

3.5.1 Estimation results and marginal effect analysis

In this section, the results of nonparametric regression with the LL estimation method and an analysis of the marginal effect of each covariate on RR are firstly discussed. This analysis informatively identifies both linear and nonlinear effects of the covariates on RR. This finding positively enhances the model's specification of semiparametric PL model in order to correctly identify and allocate the independent variables to either linear or nonparametric components. Then, the marginal effect analysis of the PL model estimates is conducted. Lastly, we apply the nonparametric regression with the LC method and discuss the result thereof. In what follows, the detail of our estimation result and marginal effect analysis of each model is discussed.

Nonparametric regression with local linear estimation method

The LL estimation method is applied to the full sample of 3,573 defaulted loans with five given RR covariates. The kernel functions are assigned according to the variables discussed in Section 3.4, where the Gaussian kernel is assigned for DC and SI, the kernel for unordered discrete variables for Type and Col, and the kernel for ordered discrete variables for Rank. We treat Rank as an ordered categorical variable, as the defaulted loan at the top of the structure (Rank = 1) is commonly paid off first and is followed by debt in the next-highest rank. Hence, the RR of the lower rank is supposed to be lower than that of the higher rank, as discussed in the preliminary analysis. In addition, we denote $Y = \text{RR}$, $X = (X^c, X^d)$, $X^c = (DC, SI)$, and $X^d = (Type, Rank, Col)$.

The leave-one-out least-squares cross-validation for bandwidth selection in (3.2.13) is employed. It yields the result of

$$H^*_{LL} \in \{0.1065, 1.3609, 0.4704, 0.1054, 0.1122\}, \tag{3.5.1}$$

where H^*_{LL} is a vector of selected bandwidth minimising (3.2.13) for DC, SI, Type, Rank, and Col, respectively. Given the optimal bandwidth, we immediately employ the local least squares method in (3.2.11) to estimate the marginal effect of each variable $\hat{b}(x)$.

[———————— Insert [Figure 3.5] here————————]

Figure 3.5a illustrates the LL estimate of DC, which clearly shows nonlinear behaviour. Although the overall effect of DC on RR is positive, as expected, the effects are insignificant when DC is less than 0.3. This implies that an increase in DC does not necessarily lead to an increase in RR. Specifically, the RR does not respond effectively to a change in DC < 0.3. A significant positive effect with increasing rate of DC is observed when DC is approximately between 0.3 and 0.5. The effect remains positive while decreasing in strength as DC ranges between 0.5

and 0.8, before having almost zero effect with a constant rate. This finding would benefit lenders' strategies for managing default exposure. In order to increase RR, lenders should pay extensive attention to stimulating the DC of the defaulted loans that have a DC between 0.3 and 0.5, rather than those with substantially low or high levels of DC. This could be a more efficient strategy than equal treatment across all defaulted loans, as nonlinearity is not taken into account in such a strategy.

The effects of change in SI are shown in Figure 3.5b. The result shows that an increase in stress level causes a deterioration of the RR. Our finding also suggests that the negative effects are approximately linear, as the effects remain constant over almost the full range of SI. However, the negative effect of SI decreases in strength for SI > 3 and is not significant at these levels. This implies that during substantially high-stress periods, such as during the GFC, the additional stress level does not affect the RR.

[———— Insert [Figure 3.6] here————]

In terms of discrete variables, the LL estimators are illustrated in Figure 3.6. The estimators of Type in Figure 3.6a can be interpreted as similar to those of the parametric model. Each estimator represents a difference compared to the reference group, which is a term loan ($\text{Type}^{(1)}$). Figure 3.6a shows that among six types of loan, revolvers ($\text{Type}^{(2)}$) are expected to have the highest RR, followed by $\text{Type}^{(1)}$, secured and unsecured senior bonds ($\text{Type}^{(3)}$ and $\text{Type}^{(5)}$), and subordinate bonds ($\text{Type}^{(4)}$ and $\text{Type}^{(6)}$), respectively. This finding is consistent with the preliminary analysis in Table 3.1 of the previous section as well as with intuition. According to the confidence intervals, the estimator of the revolving loans is significantly higher than that of $\text{Type}^{(1)}$ and that of all bonds with the exception of $\text{Type}^{(3)}$.

Figure 3.6c shows that loans with collateral are expected to have significantly higher RRs than those without collateral. Lastly, Figure 3.6b shows the estimators of Rank, which we assume an ordered property. The negative estimators of all Ranks indicate that defaulted loans with lower ranks have higher RRs than those with higher ranks. In particular, loans with Rank$^{(2)}$, Rank$^{(3)}$, and Rank$^{(4)}$, as expected, have lower RRs than Rank$^{(1)}$ by 0.15, 0.27, and 0.37, respectively. Overall, we find that our estimators for collateral status and instrumental rank are consistent with what we observe in the preliminary analysis.

Semiparametric partially linear regression

As discussed in the methodology section, PL regression requires a pre-assumption regarding functional form. The results of the nonparametric regression with the LL method enable us to identify the appropriate choices of the nonparametric and parametric components for a particular variable in (3.3.1). Hence, we specify:

$$Y_i = Z_i^b \beta + m(DC_i) + u,$$

where $Z^m = \{DC\}$, and $Z^b = \{SI, Type, Rank, Col\}$. Specifically, as DC has a non-linear effect on the RR (see Figure 3.5a), the effect will be nonparametrically estimated through $\hat{m}(\cdot)$. On the other hand, the effect of SI is observed to be approximately linear (see Figure 3.5b). Therefore, we specify the variable in the linear component together with all discrete variables to estimate the unknown linear estimators as $\hat{\beta}$. This leads to a dimensional reduction in the nonparametric estimation for DC from multi-dimension to single dimension. This also eases the computational burden and complexity brought about by high-dimensional issues. The model requires a two-step estimation process to estimate the parametric and nonparametric components, respectively.

[———— Insert [Table 3.2] here————]

Table 3.2, second column, shows the selected bandwidths required for the first-step estimation in order to obtain $\tilde{Y}_i$ and $\tilde{Z}_i$ in (3.3.3)[7]. Then, given the bandwidths, we employ the least-square method in (3.3.2) to estimate the parametric

[7] This involves the kernel estimations of the conditional mean for $E(Y_i|DC_i)$ and $E(Z_i^p|DC_i)$

coefficients of SI, Type, Rank, and Col, which are reported in table 3.2, third column. These estimates provide as transparent an interpretation as that of OLS, since they are the partial effects of the variables on the RR.

In table 3.2, the coefficient estimates have signs similar to those of the LL estimators of nonparametric regression. Significant negative effects are observed in SI and Rank, which are in line with intuition. Regarding the effect of the change in SI, the parametric estimate suggests that if we consider the movement from the least-stressed economy (SI = -1) to a relatively highly stressed economy (SI = 1), lenders can expect the RR to be approximately 0.03 lower. We also find that the coefficient estimates of Rank are larger as Rank increases from Rank$^{(2)}$ to Rank$^{(4)}$. This would show that the defaulted loan with a higher priority in the capital structure expects to have higher the level of RR. For example, a defaulted loan with Rank 1 is expected to have a higher RR than a loan with Rank 4 by 0.28. On the other hand, positive effects are found in Type$^{(2)}$ and Col, which are consistent with the nonparametric results in both direction and magnitude.

[————— Insert [Figures 3.6] here—————]

In the second-step estimation, we estimate the unknown function $\hat{m}(DC)$ using (3.3.5). The nonlinear relationship estimates between DC and the RR are shown in Figure 3.7 through the nonparametric component estimate $\hat{m}(DC)$, where the selected bandwidth is 0.03. We find that an increase in DC does not effectively increase the recovery rate when DC is at a relatively low level, DC < 0.2. A positive effect of DC is observed when DC ranges between 0.2 and 0.6. However, a further increase in DC > 0.6 does not affect the level of RR. As the slope of this nonlinear function estimates is consistent to what we discussed in the results of the local linear estimation method.

[————— Insert [Figures 3.7] here—————]

Nonparametric regression with local constant estimation method

Similar to the LL method, the local constant (LC) estimation method requires the least-squares cross-validation criteria to select the appropriate bandwidths, as discussed in (3.2.8). The selected bandwidths are:

$$H^*_{LC} = \{0.0942, 0.4120, 0.2956, 0.14464, 0.1090\}, \tag{3.5.2}$$

where H^*_{LC} is a vector of selected bandwidths for the local constant method corresponding to {DC, SI, Type, Rank, Col}, respectively. The given bandwidth is then employed to estimate the unknown function, denoted as the LC estimator $\hat{a}(x)$ described in (3.2.2) and (3.2.10).

Unlike the LL estimators, the direct marginal effect estimate is unavailable in the local constant estimators. To illustrate the effects of the RR covariates on the LC estimates, we provide a graphical explanation in figure 3.8, which illustrates the effect of a particular RR determinant on the RR of a defaulted loan with given fixed and pre-specified characteristics. Specifically, to show the effect of a particular variable, we hold other variables constant at their sample means[8] and estimate the conditional RR at the various values of the variable of interest.

[———— Insert [Figures 3.8] here————]

Figure 3.8a shows the nonlinear effect of DC, which is consistent with the LL estimation and the partially linear regression. Defaulted loans with DC < 0.2 are expected to have similar RR of 0.5, the positive effect of DC on RR is observed for an increase in DC from 0.2. An increase in DC between 0.2 and 0.6 has the largest effect, which is consistent to our previous findings. We also observe the negative impact of SI on the conditional RR, especially when SI increases from -1 to 1 in figure 3.8b. However, the figure shows some unexpected positive effects in some ranges of SI, when SI is between 1.5 and 2.3, or greater than 3.5.

[8] The fixed value of each variable is DC = 0.24, SI=0.55 (neutral economic condition), Type = 2 (revolving loan), Rank = 2, Col = 0 (no collateral)

For the categorical variables, Figure 3.8c shows that revolving loans and senior secured loans have the highest RR, followed by term loans and other types of corporate bond. Figure 3.8d shows that defaulted loans with Rank 1 consistently have the lowest RR, followed by Ranks 2 to 4, respectively. Lastly, regarding collateral status, Figure 3.8e shows that loans with collateral have a higher RR than loans without collateral by approximately 0.1.

Other alternative parametric regressions

For comparison purposes, we provide the parametric coefficient estimates from QMLE-RFRV, two-sided censored Tobit, inverse Gaussian *back-transformation* regression (IG), and the mixture model (MM), where these models assume standard linear functional form, in Table 3.3. We find that all estimates are mostly in line with our expectations regarding the signs of the estimates across the four models. Positive signs are consistently observed in DC, revolving loan (Type = 2), and Col, and negative signs are found in SI and Rank. We also find that the coefficient estimates of all instrumental ranks increase as Rank increases from 1 to 4. These findings are consistent with the previous results of our proposed models.

[——————— Insert [Table 3.3] here———————]

3.5.2 Predictive performance of models

In this section, the predictive performances of the proposed models are evaluated and then compared with the performances of the existing models, which include QMLE-RFRV, two-sided censored Tobit, IG, MM, and the regression tree algorithm (RT). We employ the standard mean squared error (MSE) to measure the predictive accuracy of full sample, in-sample, and out-of-sample data.

Importantly, to imitate the application of RR prediction in practice, we use intertemporal data partitioning to define the in-sample and out-of-sample data (Kalotay & Altman, 2016; Gupton & Stein, 2005). In particular, we evaluate 12 windows of in- and out-of-sample subsamples, where we take the first in-sample window to be the loans defaulted on from 1994 to 2000, and out-of-sample to be the remaining observations between 2001 and 2012. Then, the in-sample windows roll until the last in-sample window is the observation between 1994 and 2011, while the out-of-sample RR is the loans defaulted on in 2012.

This data partitioning method ensures that there is no overlapping information between in-sample and out-of-sample data, as we observe that: (i) one borrower might have multiple defaulted loans; but (ii) all loans from the borrower will be defaulted in the same year. The predictive evaluation based on these in- and out-of-sample data partitioning is more robust than a conventional way[9] which allows the overlapping information between two subsamples.

Full sample prediction

Given the specifications and the estimates discussed in Section 3.5.1, their full sample predictive performances are reported in Table 3.4.

[————— Insert [Table 3.4] here—————]

The results show that the nonparametric regressions yield high accuracy rates. The model with the LC estimation method yields the most precise predictions, with MSE = 0.06 and MAE = 0.2. This is followed closely by the LL estimation method, with MSE = 0.07 and MAE = 0.21. On the other hand, the performance of the proposed PL model is comparable to that of the machine learning RT. In particular, the MSE and MAE of the semiparametric PL model are 0.088 and 0.238, respectively, compared to 0.083 and 0.224 for the RT. Our result is consistent with the finding in the existing literature which suggests that nonlinear models commonly offer better predictive performance than linear models. Furthermore, although the linear regressions performed poorly in generating RR predictions,

[9] For example, the full sample is randomly partitioned to form in- and out-of-sample with 70:30 ratio

QMLE-RFRV performs best among the other parametric models, followed by Tobit.

In-sample predictions

[———— Insert [Table 3.5] here————]

Table 3.5 reports the in-sample predictive performances across 12 windows. On average, the proposed models, with the exception of the LL model, outperform the existing models, and nonparametric regression with the LC method still provides the best prediction. On the other hand, the in-sample predictive errors of LL are high for the first nine windows, before decreasing sharply in the last four windows. We observe that these are the consequences of the boundary issue in LL, as the predictions may lie outside of the [0,1] boundary, especially with a small sample size. We find that this issue is mitigated in the full sample estimation, where the sample size is large.

To overcome the boundary issue in the LL method, we apply the LL estimation for nonparametric regression with two-sided censoring introduced in Section 3.2.4, denoted as LL2 in Table 3.5. The result shows a vast improvement in MSE from LL method to LL2 method, where the MSE in the first in-sample window is reduced from 0.18 to 0.06, respectively.

A similar boundary problem is found in PL, as the model specification does not guarantee that the PL predictions will be restricted to within the bounded [0,1]. However, we observe that only 0.5% of the total in-sample predictions exceed the boundary with small magnitudes[10] compared to 5% for LL. This could be due to the role of dimension reduction in the PL's functional form as well as the nonlinearity captured by the nonparametric component, which may mitigate the boundary problems of PL.

[10] The maximum predicted recovery is 1.014 while the minimum is -0.001

Given the average MSEs in Table 3.5, the best model in terms of in-sample predictive accuracy is LC, followed by LL2 method, where their MSEs are 0.069 and 0.073, respectively. These two models also consistently have the lowest MSEs in all 12 rolling windows. Interestingly, then, we find that the parametric QMLE-RFRV offers a relatively low average MSE of 0.078, which is slightly lower than the predictions of PL with an MSE of 0.079. This could be due to the fact that QMLE-RFRV is theoretically valid for the [0,1] bounded recovery rate, and it can accommodate some degree of nonlinearity through the logit link function. However, the performances of PL and QMLE-RFRV are similar when we consider each window in Table 3.5. The yearly MSEs of QMLE-RFRV do not consistently outperform the predictive performances of PL, and the differences between these two models are small. In addition, although RT performs excellently in full sample prediction, its performance is poorer than that of PL and QMLE-RFRV for in-sample predictions, with an average MSE of 0.084 followed by Tobit, IG, and MM, respectively.

Out-of-sample predictions

Table 3.6 evaluates the out-of-sample predictive errors of the given the models estimated by the 12 rolling in-sample windows discussed previously. On average, the table reports that PL provides the most accurate predictions, followed by LC, RT, and LL2, respectively. We also find that the performance of the nonparametric regression with LL method has a high MSE, as we observed in its in-sample prediction performance.

[————— Insert [Table 3.6] here—————]

The results in Table 3.6 further reveal that the predictive accuracies of QMLE-RFRV, Tobit, IG, and MM are low during the 2005 to 2009 windows, where the 2008 window yields the highest MSE of these models. This poor performance might be due to the GFC, as we have shown that the economy experienced the highest recorded SI in 2008 in the preliminary analysis of Figures 3.3 and 3.4. In addition, the SI reached its lowest level in 2005, with the highest average RR. Given the observed conditions in 2005-2009 together with the predictive performances of QMLE-RFRV, Tobit, IG, and MM models, our results indicate that these four models might not be able to accommodate and immediately respond to sudden changes in macroeconomic conditions, especially during GFC. The models would be restricted by the linear functional form. On the other hand, the data-driven approach, including PL, LC, and RT, which do not rely on the linear functional form can pick up the changes in the economic conditions more responsively. These models provide flexibility, as the restricted linear functional form is not required, but is rather estimated nonparametrically. This may yield the superior predictive power of the non- and semi-parametric regressions compared to that of the linear models. This finding highlights an advantage of our proposed PL and LC, as well as the existing RT to estimate the downturn RR.

A common observation for a flexible model is the presence of low MSE in in-sample prediction and high MSE in out-of-sample prediction. Hence, Table 3.6 also evaluates the degree of difference between the in-sample and out-of-sample predictions in terms of MSE ratio[11]. If the ratio is closer to one, it indicates that the predictions in both subsamples are similar. The results in Table 3.6 show that PL has the highest ratio of 0.90, followed by IG with 0.88 and Tobit with 0.87. On the other hand, LC and LL2 methods show relatively low ratios, with 0.73 and 0.68, respectively.

The PL model's performance seems to be robust, as it shows relatively low MSE variation across 12 rolling out-of-sample windows, as well as its relative MSE ratio. The proposed nonparametric regression with LC method also yields a similarly low MSE, but this model seems to have a large variation of MSE and the relative MSE ratio than PL model. On the other hand, the proposed nonparametric

[11] Relative MSE ratio = $\frac{MSE_{in-sample}}{MSE_{out-of-sample}}$

regression with the LL estimation method as well as the model with two-sided censoring assumption provide a predictive performance as good as that of the alternative linear regressions.

The outstanding performance of the PL model is expected due to its functional form. Assigning DC to the nonparametric component and the remaining variables to the linear parametric component leads to two main advantages over the other models included. Firstly, the dimension in the nonparametric component is reduced significantly, which overcomes the computational difficulty in the nonparametric regressions. Secondly, the nonlinear effects can be taken into account, which mitigates the misspecification problem in the linear model.

Additionally, we also apply the Receiver Operating Characteristic (ROC)[12] curve to determine how well the predictive model can differentiate between high and low RRs (Gupton & Stein, 2005; Siao et al., 2015). In our study, if the RR is greater than 0.5, we define it as high-RR, and low-RR otherwise. Models that yield an area under the ROC curve (AUC) that is closer to 1 have better discriminatory power.

[———— Insert [Table 3.7] here————]

Table 3.7 suggests that PL has the highest discrimination rate of 0.82 on average, while LC, LL2, QMLE-RFRV, Tobit, and IG provide an almost identical AUC of 0.80. Considering the yearly breakdown performances in Table 3.7, all parametric QMLE-RFRV, IG and TOBIT models show that their AUCs reach the lowest rate in 2008. The results are consistent with their out-of-sample MSEs performances, which suggest that the performance of the parametric models are poor during the GFC period.

[12] This criterion is common in binary choice model to measure the discriminatory power of the model, see Hanley and McNeil (1982)

3.6 Conclusion

In this chapter, we proposed a nonparametric regression with the local constant (LC) and local linear (LL) estimation methods and a semiparametric partially linear (PL) model for the recoveries of defaulted loans. We have addressed two important issues associated with these models: clarifying the nonlinear marginal effects in the RR-covariate relationship and improving the prediction of RR.

First, attention was directed towards the marginal effect estimates using the LL estimation method. The results indicate a nonlinear effect of debt cushion (DC), which is a characteristic variable of loans, and an approximately linear effect of economic stress index (SI). Moreover, the LL method's findings enable us to improve the function form of the PL model by specifying only DC in the nonparametric component and the remaining variables, SI and categorical loan-specific variables in the parametric component. In the PL model, the marginal effect estimates are found to be similar to those of the LL estimates. Non- and semi-parametric models and the findings together make a significant contribution to the RR modelling literature.

Second, the predictive performances of the proposed models were compared with those the existing several approaches, including QMLE regression for fractional response variable, two-sided censored Tobit, inverse Gaussian *back-transformation* regression, the mixture distribution model, and the regression tree algorithm. Overall, we find that the partially linear regression consistently outperforms all the other models included in this study, while the nonparametric regression with the local constant method has slightly weaker predictive performance. In addition, the partially linear model's superior predictability tends to be more robust and stable than that of the alternative parametric regressions, especially during the financial crisis period.

Furthermore, we observe that there is still the boundary problem in the LL method, as high proportions of RR predictions exceed the [0,1] boundary. Therefore, we additionally introduced the nonparametric regression with two-sided censoring using LL method. We then applied the model to the empirical RR data and find that this method substantially improve the predictability of the LL estimation method.

The methods introduced and applied to study RR-covariate relationship in this chapter have not fully accommodated the RR distributional properties such as asymmetry and bimodality. Therefore, in the next chapter, we will introduce quantile regression which estimates the RR-covariate relationship over the various quantiles of the conditional distribution. This allows not only the nonlinear marginal effect analysis on the conditional mean of RR, but it will also capture the heterogeneity of the effects on the various conditional quantiles of RR. Moreover, although the boundary issue is vastly mitigated in this chapter, it has not been completely eliminated. In chapter 5, we, therefore, will propose and study models that will completely eliminate the boundary problem.

3.7 Appendix A: Simulation study of the nonparametric regression with two-sided censoring

To provide evidence on the finite sample performance of the proposed nonparametric regression with two-sided censoring, we conduct a simulation study. The data-generation process assumptions are the same as in Lewbel and Linton (2002). We then further impose the two-sided censoring condition by assuming fixed two-sided censoring assumptions at -0.5 and 0.2. Then, we generate the simulated data as follows:

$$Y = \max(-0.5, \min(w(X) - e, 0.2)),$$

where $Y^* = w(x) - e$, $w(X) = X^3$, $X \sim \text{Uniform}[-1,1]$, and $e \sim N(0, 0.25)$. The sample size is n = 200 for 1,000 iterations. Figure 3.9 shows that the empirical density of Y is similar to the empirical bimodal density of the RR in Figure 3.1.

[————————[Figures 3.9 and 3.10] here ————————]

Figure 3.10 illustrates the estimates of $w(\cdot)$ over 1,000 iterations. Compared to the cube function (solid line), the estimates can recover the true latent function relatively well, as shown in the grey area which represents the variations of the estimates over the iterations. It clearly shows that the median of the estimates is a good approximation of the cube function. Furthermore, our result is consistent and comparable to the main finding in one-sided censoring simulation study results provided by Lewbel and Linton (2002). The boundary condition can then be addressed, as we employ $\hat{y} = \max(-0.5, \min(\hat{w}(x), 0.2))$.

Given the simulation result[13],we can apply the proposed model by assuming two-sided censoring at zero and one. This will ensure that the RR prediction will lie strictly within the [0,1] boundary.

[13] The further study of this method including the detailed finite and asymptotic properties will be one of our future research directions

3.8 Appendix B: Tables and figures

			Recovery rate at q-quantile					
Variables	Frequency	%	mean	5%	25%	50%	75%	95%
Recovery rate	3,573	100%	0.5570	0.0000	0.1849	0.5888	1.0000	1.0000
Panel A: Discrete Variables								
(i) Type of loans								
Term loans (Type$^{(1)}$)	746	21%	0.7054	0.0222	0.4343	0.8638	1.0000	1.0000
Revolving loans (Type$^{(2)}$)	738	21%	0.8251	0.2147	0.6934	1.0000	1.0000	1.0000
Senior secured bonds (Type$^{(3)}$)	446	12%	0.5896	0.1079	0.2093	0.5682	1.0000	1.0000
Senior subordinated bonds (Type$^{(4)}$)	355	10%	0.2429	0.0000	0.0080	0.1005	0.3779	0.9225
Senior unsecured bonds (Type$^{(5)}$)	1,061	30%	0.4263	0.0000	0.1027	0.3578	0.7257	1.0000
Junior bond (Type$^{(6)}$)	227	6%	0.2352	0.0000	0.0000	0.0968	0.3491	1.0000
(ii) Instrument rank								
Rank$^{(1)}$	1,711	48%	0.7476	0.1221	0.5151	1.0000	1.0000	1.0000
Rank$^{(2)}$	1,258	35%	0.4294	0.0000	0.1159	0.3170	0.7422	1.0000
Rank$^{(3)}$	393	11%	0.2994	0.0000	0.0041	0.1531	0.5255	1.0000
Rank$^{(4)}$	211	6%	0.2514	0.0000	0.0010	0.1027	0.3615	0.9122
(iii) Collateral								
Uncollaterized loans	1,712	48%	0.3685	0.0000	0.0396	0.2372	0.6638	1.0000
Collaterized loans	1,861	52%	0.7303	0.1240	0.4545	0.9622	1.0000	1.0000
Panel B: Continuous variables								
(i) Debt Cushion (DC)								
DC = 0	1,631	46%	0.3969	0.0000	0.0738	0.2840	0.6959	1.0000
0 < DC < 0.5	1,049	29%	0.5345	0.0043	0.2092	0.5476	0.8961	1.0000
0.5 < DC < 1	893	25%	0.8766	0.2386	0.9435	1.0000	1.0000	1.0000
(ii) Stress index (SI)								
SI ≤ 0	853	24%	0.7058	0.0022	0.4728	0.8585	1.0000	1.0000
0 < SI < 1	2,272	63%	0.5102	0.0000	0.1533	0.4654	1.0000	1.0000
SI ≥ 1	448	12%	0.5105	0.0000	0.1344	0.4617	1.000	1.0000

Table 3.1: *Preliminary analysis of the empirical recovery rate data*

Note: We provide the contingency table of recovery rates in the column 4-9, where we partition the recovery rate conditional on the information provided in the first column.

	Partially linear model		
Variables	Bandwidth	Coefficients	
Dependent variable			
Recovery rate	0.0178	N/A	
Independent variables			
Debt cushion	0.0343	N/A	
Stress index	0.0055	-0.0308	*
		(0.0062)	
Revolving loan	0.0791	0.0591	*
		(0.0180)	
Senior secured bond	0.2547	0.0061	
		(0.0248)	
Senior subordinate bond	0.0494	-0.0693	
		(0.0353)	
Junior secured bond	0.0377	-0.0019	
		(0.0327)	
Subordinate bond	0.0131	-0.0888	*
		(0.0375)	
Rank 2	0.0074	-0.1400	*
		(0.0190)	
Rank 3	0.0566	-0.1954	*
		(0.0250)	
Rank 4	0.0782	-0.2827	*
		(0.0334)	
Collateral status	0.0574	0.1138	*
		(0.0299)	

Table 3.2: *Estimates of the partially linear regression*

Note: The table reports the estimates using two-step estimation method discussed in section 3.3 which we assign all covariates except DC in a linear component. In the first-step estimation, bandwidths of all variables except DC are selected as discussed in (3.3.3) and reported in column 2. Given these bandwidths, their parametric coefficients are estimated using (3.3.4) and reported in column 3. To estimate the nonparametric component in the second-step estimation, the bandwidth of DC is selected by minimizing (3.3.6). (*) indicates that the estimator is significant at 5% level of significance.

Variables	Linear models			
	QMLE-RFRV	Tobit	IG	MM
Debt cushion	2.5590 *	0.7932 *	1.8066 *	1.8593 *
	(0.1603)	(0.0359)	(0.0827)	(0.0052)
Stress index	-0.1928 *	-0.0578 *	-0.1058 *	-0.1854 *
	(0.0353)	(0.0078)	(0.0195)	(0.0003)
Revolving loan	0.4388 *	0.1612 *	0.3302 *	0.3069 *
	(0.1130)	(0.0260)	(0.0606)	(0.0080)
Senior secured bond	0.0539	-0.0155	-0.0480	-0.0260 *
	(0.1387)	(0.0291)	(0.0735)	(0.0041)
Senior subordinate bond	-0.1731	-0.0876	-0.3827 *	-0.2297 *
	(0.1915)	(0.0467)	(0.1205)	(0.0110)
Junior secured bond	0.1143	0.0767	0.1106	0.1937 *
	(0.1748)	(0.0414)	(0.1065)	(0.0086)
Subordinate bond	-0.2367	-0.1256 *	-0.5057 *	-0.1857 *
	(0.2050)	(0.0495)	(0.1270)	(0.0123)
Rank 2	-0.6270 *	-0.1307 *	-0.3205 *	-0.3780 *
	(0.0999)	(0.0217)	(0.0536)	(0.0024)
Rank 3	-0.9109 *	-0.2384 *	-0.6974 *	-0.5739 *
	(0.1356)	(0.0302)	(0.0788)	(0.0047)
Rank 4	-1.4636 *	-0.2834 *	-0.8888 *	-0.6339 *
	(0.2107)	(0.0367)	(0.0958)	(0.0069)
Collateral status	0.5057 *	0.1459 *	0.3532 *	0.3335 *
	(0.1577)	(0.0379)	(0.0980)	(0.0073)

Table 3.3: *Estimation results of four alternative existing linear models linear parametric estimators of alternative parametric models*

Note: The table reports the coefficient estimates for QMLE regression for fractional response data (QMLE-RFRV), Tobit regression with two-sided censoring at zero and one (TOBIT), linear regression on invert Gaussian transformed recovery rate (IG), and three mixture-distribution model on transformed recovery rate (MM). (*) indicates significance at 5% level

Models	MSE	MAE
Proposed models		
Local constant method	0.0677	0.2033
Local linear method	0.0747	0.2164
Partially linear regression	0.0886	0.2384
Alternative parametric regressions		
QMLE-RFRV	0.0892	0.2408
Tobit	0.1107	0.2944
IG	0.1271	0.2444
MM	0.1319	0.2464
RT	0.0813	0.2244

Table 3.4: *Full sample predictive accuracy*

Note: The table reports the full sample prediction of the proposed models and other five alternative models: QMLE regression for fractional response data (QMLE-RFRV), Tobit regression with two-sided censoring at zero and one (TOBIT), linear regression on invert Gaussian transformed recovery rate (IG), three mixture-distribution model on transformed recovery rate (MM), and regression tree algorithm (RT).

Year of default	Proposed models				Alternative models				
	LC	LL2	LL	PL	QMLE-RFRV	RT	Tobit	IG	MM
In-sample period									
1994-2000	0.0552	0.0565	0.1780	0.0658	0.0665	0.0942	0.0754	0.0790	0.0914
1994-2001	0.0749	0.0756	0.1703	0.0755	0.0755	0.0871	0.1026	0.1061	0.1168
1994-2002	0.0725	0.0714	0.1405	0.0768	0.0772	0.0836	0.0946	0.0998	0.1225
1994-2003	0.0732	0.0769	0.1300	0.0801	0.0799	0.0819	0.0981	0.1024	0.1234
1994-2004	0.0707	0.0754	0.1118	0.0791	0.0778	0.0823	0.0960	0.1000	0.1188
1994-2005	0.0692	0.0746	0.1066	0.0796	0.0781	0.0820	0.0950	0.0997	0.1146
1994-2006	0.0688	0.0739	0.1052	0.0798	0.0778	0.0820	0.0949	0.0995	0.1163
1994-2007	0.0687	0.0738	0.1026	0.0800	0.0776	0.0815	0.0945	0.0990	0.1315
1994-2008	0.0690	0.0740	0.0957	0.0825	0.0821	0.0823	0.1004	0.1050	0.1382
1994-2009	0.0678	0.0748	0.0790	0.0840	0.0835	0.0837	0.0994	0.1042	0.1326
1994-2010	0.0677	0.0746	0.0750	0.0830	0.0826	0.0829	0.0985	0.1032	0.1305
1994-2011	0.0678	0.0748	0.0748	0.0829	0.0825	0.0828	0.0985	0.1031	0.1319
Average	0.0688	0.0730	0.1141	0.0791	0.0784	0.0839	0.0957	0.1001	0.1224

Table 3.5: *In-sample predictive accuracy*

Note: The proposed models are a nonparametric regression with three estimation methods: local constant(LC), local linear (LL), and local linear with two-sided censoring (LL2), and (ii) the semiparametric partially linear regression (PL). The alternative models are described in Table 3.4

Year of default	Proposed models				Alternative models				
	LC	LL2	LL	PL	QMLE-RFRV	RT	Tobit	IG	MM
Out-of-sample period									
2001-2012	0.1059	0.1123	0.1554	0.0976	0.1071	0.1271	0.1215	0.1266	0.1492
2002-2012	0.0998	0.1256	0.2382	0.0944	0.1035	0.1035	0.0997	0.1045	0.1228
2003-2012	0.1106	0.1535	0.3085	0.1060	0.1104	0.1159	0.1030	0.1095	0.1561
2004-2012	0.0940	0.1148	0.1862	0.0997	0.1148	0.0960	0.1170	0.1226	0.1572
2005-2012	0.0891	0.1184	0.1851	0.1011	0.1344	0.0975	0.1338	0.1336	0.1716
2006-2012	0.0911	0.1203	0.1945	0.0961	0.1387	0.0924	0.1473	0.1420	0.1810
2007-2012	0.0929	0.1255	0.2026	0.0962	0.1495	0.0927	0.1558	0.1490	0.1925
2008-2012	0.0978	0.1309	0.2186	0.0945	0.1559	0.0931	0.1637	0.1545	0.2276
2009-2012	0.0936	0.0901	0.0901	0.0831	0.0859	0.0827	0.0853	0.0877	0.1256
2010-2012	0.0834	0.0812	0.0812	0.0611	0.0630	0.0735	0.0673	0.0709	0.1021
2011-2012	0.0863	0.0884	0.0884	0.0706	0.0751	0.0882	0.0765	0.0831	0.1542
2012	0.0747	0.0568	0.0568	0.0562	0.0702	0.1035	0.0522	0.0673	0.1597
Average	0.0933	0.1098	0.1671	0.0881	0.1091	0.0972	0.1103	0.1126	0.1583
Variation	0.0001	0.0006	0.0052	0.0003	0.0009	0.0002	0.0012	0.0008	0.0010
Relative MSE ratio	0.7375	0.6651	0.6828	0.8982	0.7192	0.8629	0.8675	0.8888	0.7731

Table 3.6: *Out-of-sample predictive accuracy*

Note: The relative MSE is computed by the average in-sample MSE over the out-of-sample MSE. If the ratio is one, it indicates that there is no difference between in-sample and out-of-sample predictive performance.

Year of default	Proposed models			Alternative models				
	LC	LL2	PL	QMLE-RFRV	RT	Tobit	IG	MM
Out-of-sample period								
2001	0.7712	0.7692	0.7801	0.7811	0.6893	0.7768	0.7768	0.7556
2002	0.7846	0.7754	0.7918	0.7911	0.7444	0.7947	0.7947	0.7498
2003	0.7685	0.7636	0.7987	0.7944	0.7474	0.7993	0.7993	0.7516
2004	0.7984	0.7789	0.8215	0.8044	0.8090	0.8057	0.8057	0.7648
2005	0.8098	0.7795	0.8160	0.7855	0.7840	0.7909	0.7909	0.7506
2006	0.8201	0.7807	0.8227	0.7809	0.8080	0.7785	0.7785	0.7388
2007	0.8083	0.7685	0.8194	0.7707	0.8021	0.7723	0.7723	0.7186
2008	0.8054	0.7607	0.8219	0.7590	0.8066	0.7597	0.7597	0.6841
2009	0.8212	0.8255	0.8535	0.8455	0.8335	0.8452	0.8452	0.7958
2010	0.9119	0.9015	0.9085	0.9034	0.8803	0.9012	0.9012	0.8354
2011	0.7832	0.7797	0.8182	0.8182	0.7378	0.8182	0.8182	0.7308
2012	0.7600	0.8200	0.8200	0.8200	0.7500	0.8200	0.8200	0.7500
Average	0.8035	0.7919	0.8227	0.8045	0.7827	0.8052	0.8052	0.7522
Variation	0.0013	0.0014	0.0009	0.0013	0.0022	0.0012	0.0012	0.0012

Table 3.7: *Out-of-sample Area under Receiver Operator Curve (AUC)*

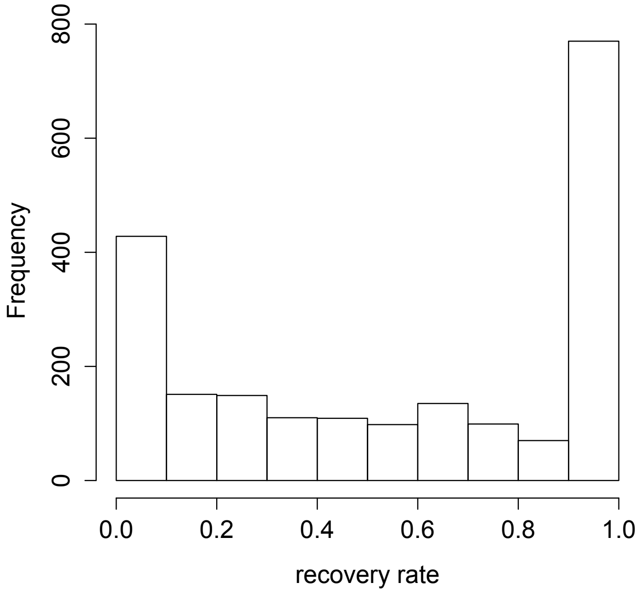
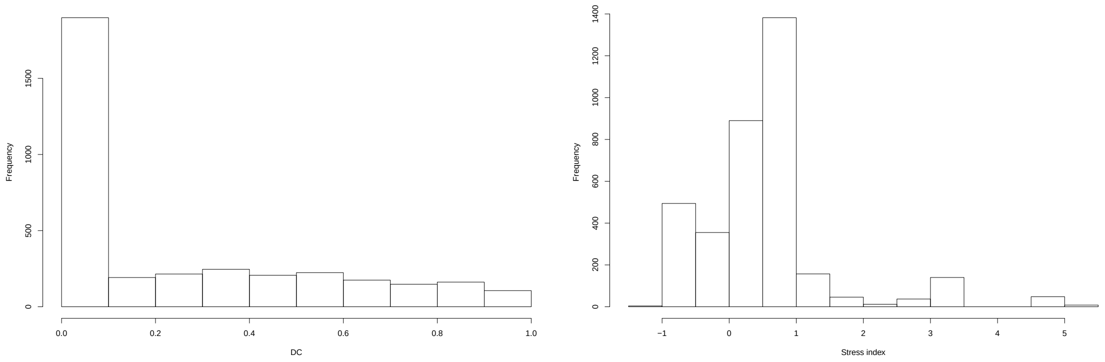


Figure 3.1: *Density of the empirical recovery rate*



(a) *Debt cushion*

(b) *Stress index*

Figure 3.2: *Densities of debt cushion and stress index*

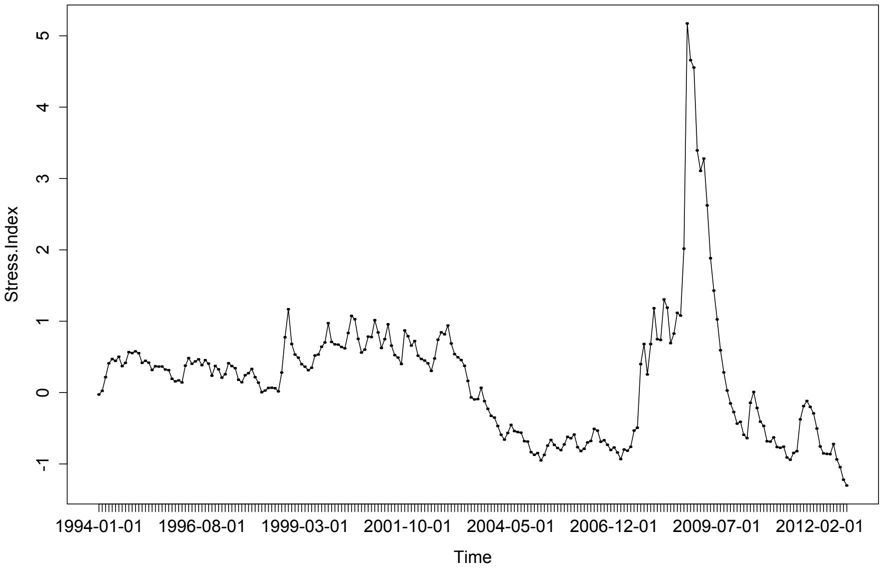


Figure 3.3: *The movement of Stress index between 1994 and 2012*

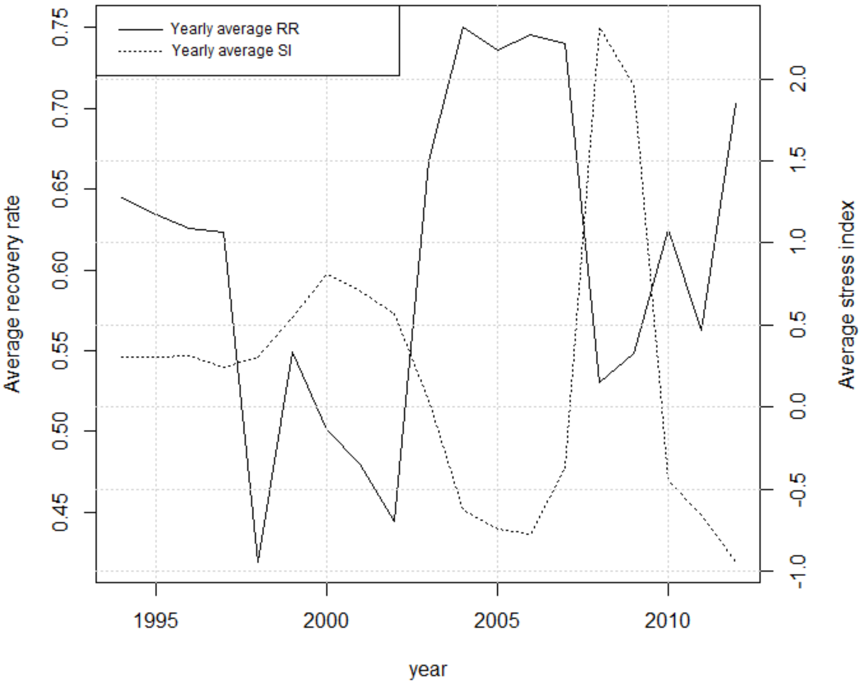
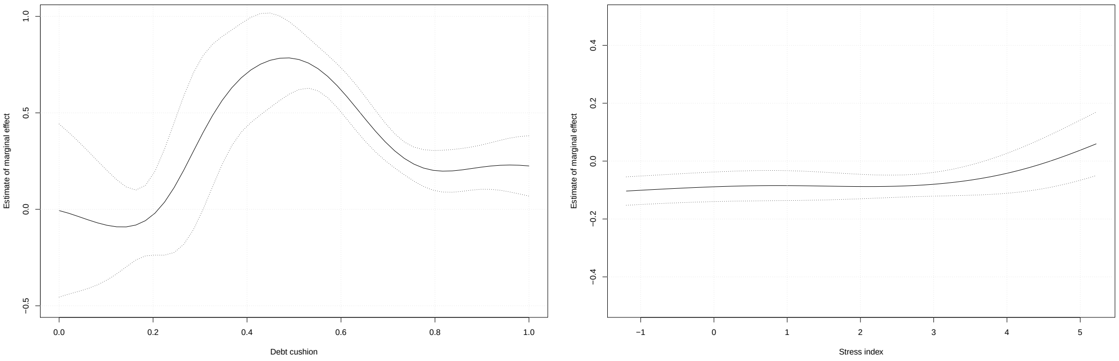


Figure 3.4: *Annual averages of recovery rate and stress index*

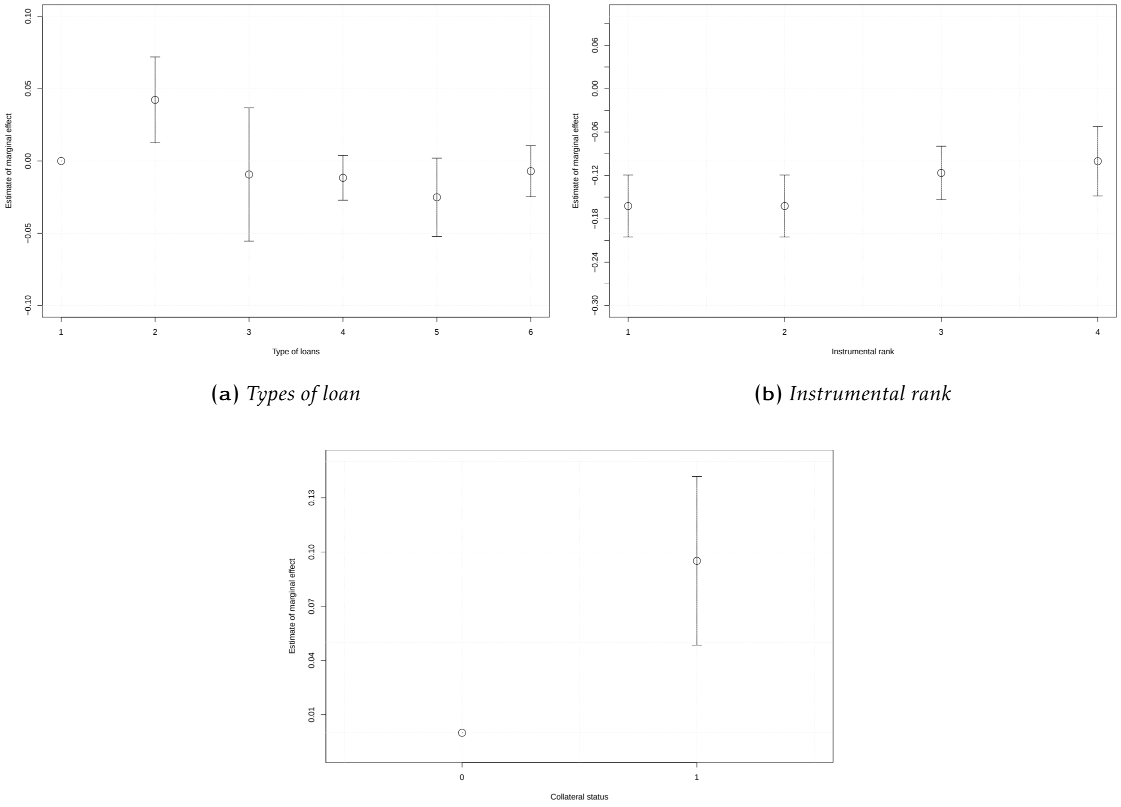


(a) *Debt cushion*

(b) *Stress index*

Figure 3.5: *Marginal effect of debt cushion and stress index on RR, estimated by local linear estimation method*

Note: The dark solid lines are the marginal effect estimates using local linear method for the given continuous variables as described below each figure. The dotted lines represent the bootstrapping confident bounds at 5% level of significance, where we employ 1,000 iterations



(a) *Types of loan*

(b) *Instrumental rank*

(c) *Collateral status*

Figure 3.6: *Marginal effects of types of loan, instrumental rank, and collateral status on RR, estimated by local linear estimation method*

Note: The figures represent the box-plot of the marginal effect estimates using local linear estimators for categorical variables described below each figure. We also provide the bootstrapped confident interval of each estimate.

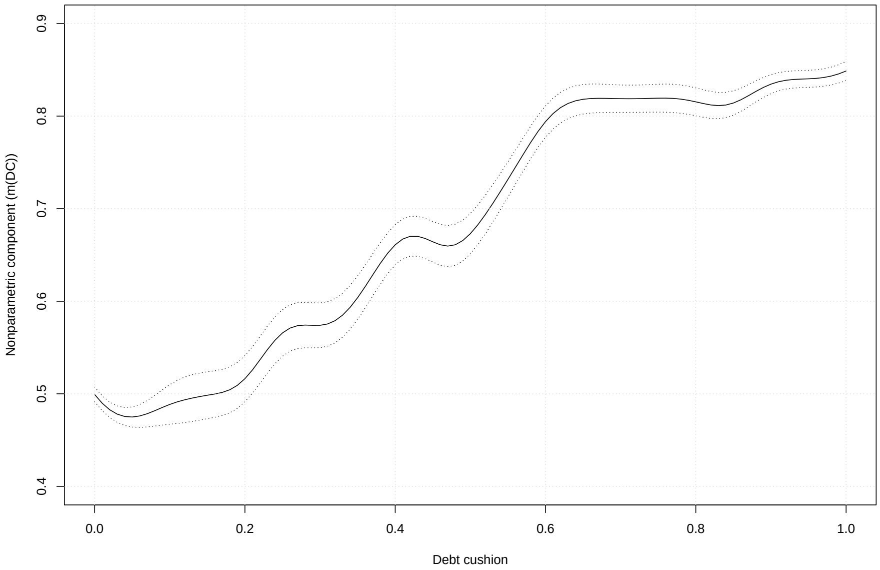
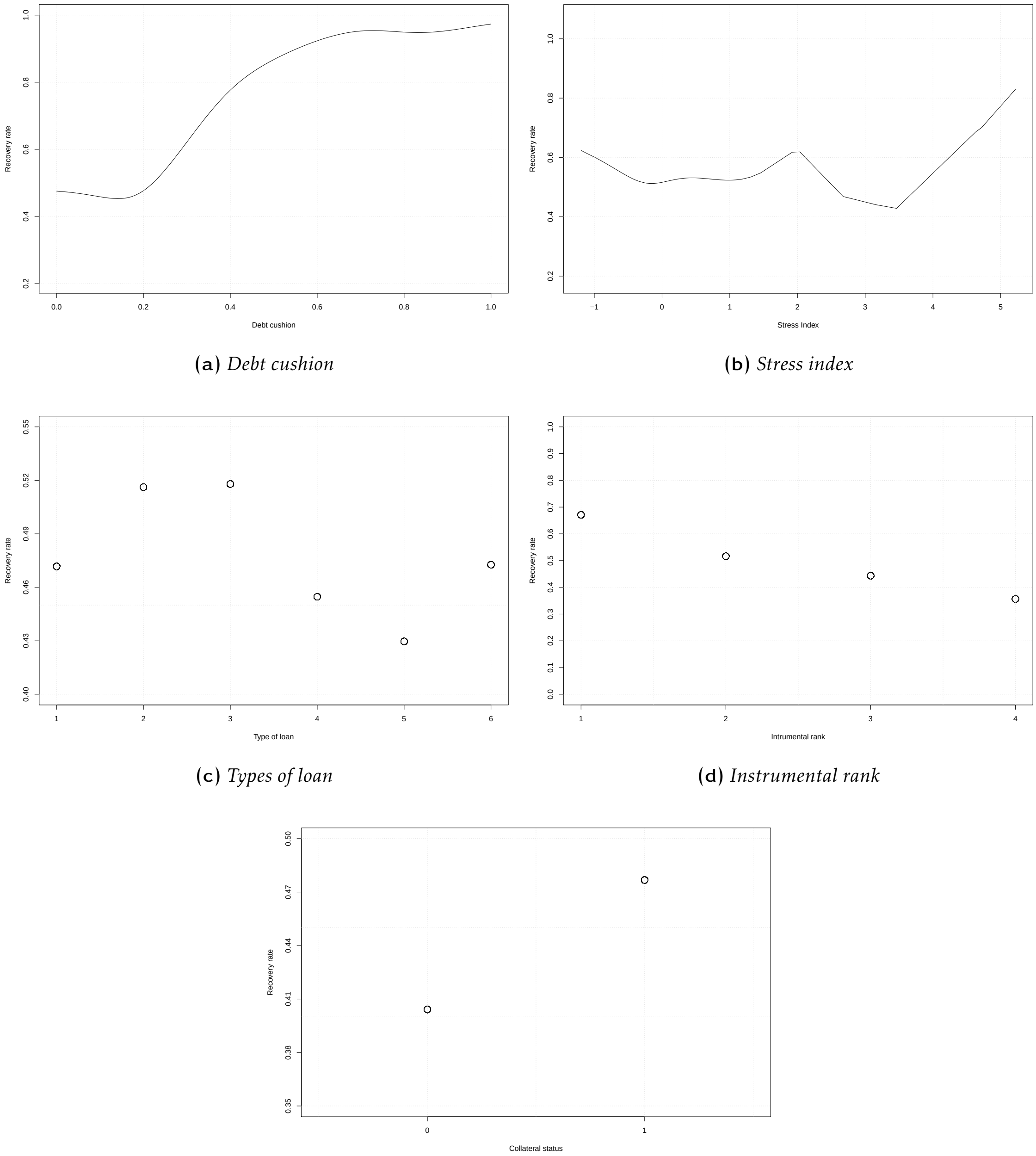


Figure 3.7: *Nonlinear effect estimates of debt cushion using partially linear regression*

Note: As DC is assigned in the nonparametric component in PL model specification, this figure represents the estimates of the unknown function of DC, $\hat{m}(DC)$, with its bootstrapped confident interval with 5% level of significance.



(a) *Debt cushion*

(b) *Stress index*

(c) *Types of loan*

(d) *Instrumental rank*

(e) *Collateral status*

Figure 3.8: *Effects of RR-covariates using local constant estimation method*

Note: To obtain the functional plots, we predict the recovery rate by given any variable of interest varying within its support range while fixed other remaining variables fixed at their means. Then we plot the predicted recovery as a function of the variable of interest described below each figure.

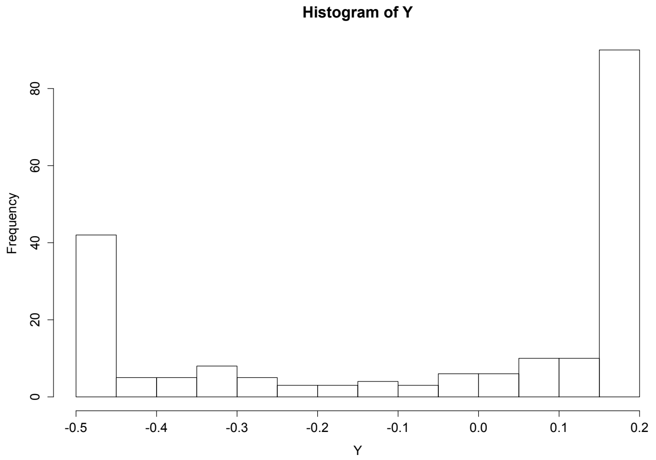


Figure 3.9: *Density of the simulated two-sided censoring response variable*

Note: The data generating process assumption is $Y = \max(-0.5, min(X^3 - e, 0.2)$

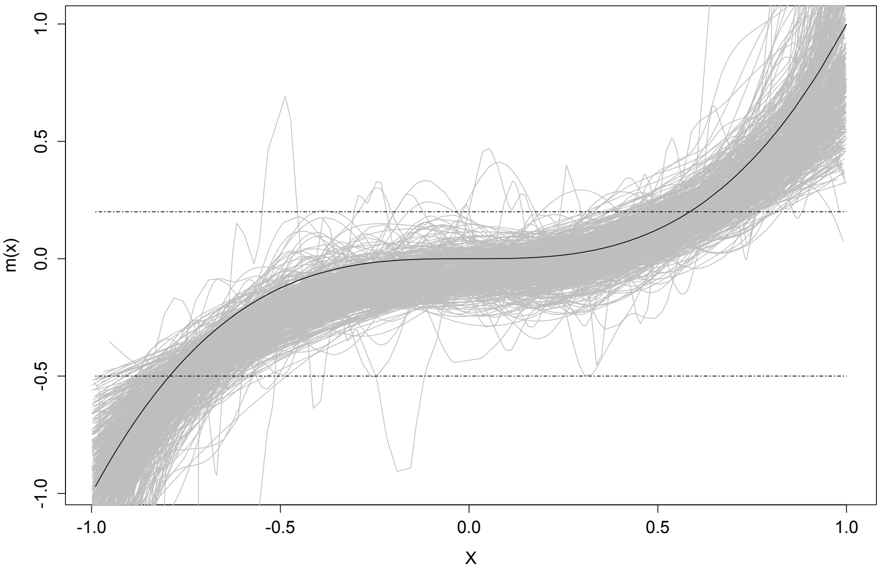


Figure 3.10: *Simulation result of the $w(x)$ estimate using two-sided censored nonparametric regression*

Note: The dark line solid is the true $w(x) = x^3$. The grey area is the $\hat{w}(x)$ estimated across 1,000 simulations. The dash lines at $m(x)$ = -0.5 and 0.2 indicate the two-sided fixed censoring points.

Chapter 4

Nonlinear quantile regressions for recovery rates

4.1 Introduction

So far, the modelling of RR has been limited to central tendency (see chapters 2 and 3 for details), it would be beneficial to banks and regulators to determine the influence of covariates on RR, and their heterogeneity across various parts of the conditional distribution of RR. These lead to the applications of the quantile regression (QR) in RR modelling in this chapter. Moreover, as the QR models' estimates vary across the conditional RR distribution, they would accommodate the heteroscedastic errors and the bimodality of the distribution. In comparison to the applications of the proposed conditional mean regressions in Chapter 3, although the distribution of error terms was not assumed, because of peculiar RR distribution, the quantile regression is more appropriate than the mean regression to address such property.

QR has been employed in several of areas. Fitzenberger, Koenker, and Machado (2013) have reviewed studies in the labour and public economic applications of quantile regressions, which are mostly related to income inequality. These studies highlight significant effects of, for example, education (Arias, Hallock, & Sosa-Escudero, 2002) and gender (Buchinsky, 2002) on income, which were found to vary across the quantile. In public finance, Okada and Samreth (2012), and Billger and Goel (2009) investigate the effects of macroeconomic and country-specific variables on the level of corruption using QR. They find that some current anti-corruption policies may be reconsidered for nations that are at the lower 0.1 and upper 0.9 quantiles of the corruption distribution. The QR is also applied in financial studies in various areas such as housing price analysis (Zietz, Zietz, & Sirmans, 2008), capital structure (Fattouh, Scaramozzino, & Harris, 2005), financial market (Ma & Pohlman, 2008; Baur, Dimpfl, & Jung, 2012; Meligkotsidou, Panopoulou, Vrontos, & Vrontos, 2014), among others. Moreover, the analysis using nonlinear QR can be found in Fenske, Kneib, and Hothorn (2011), who employ additive quantile regression using gradient boosting estimation method to identify risk factors of childhood malnutrition. They reveal that the effects of some factors such as the ages of children and their mothers have nonlinear effects on childhood malnutrition, and those effects vary at the different conditional quantiles.

In this chapter, motivated by the attractive properties of QR, we focus on the QR proposed by Koenker and Bassett (1978), which provides more complete information regarding the statistical analysis of the relationship between the response variable and the covariates. Recently, Siao et al. (2015); Krüger and Rösch (2017) applied the conventional parametric linear QR (L-QR) to analyse heterogeneous effects and predict RR. The aim of this chapter is to propose nonparametric (NP) and partially linear additive (PLA) quantile regression models and uncover the nonlinear and heterogeneous effects of the covariates on RR at various quantile of the RR distribution. Although the nonlinearity in RR modelling was addressed in the conditional mean regressions in Chapter 3, it has not been explored in the conditional quantile specifications. We will fill this gap in this chapter. In addition, the QR regression has an invariant property under data transformation and back-transformation, which does not introduce bias in the model's estimates (discussed in section 4.2.4). Thus, we also apply a partially linear additive model with logit transformation of RR (PLA-QR(tr)) to restrict the model's predictions to be within the unit interval. To evaluate the performances of the models included in this chapter, we employ several model selection criteria: goodness of fit and point prediction measurement for QRs, as well as distributional fit and the application in the Value at Risk framework which were proposed and implemented by Krüger and Rösch (2017).

This chapter makes several significant contributions to the literature on RR modelling. First, this is the first study to propose a NP-QR model for RR. Such a model, along with the local constant estimation method, generates RR predictions in the bounded [0,1] interval. As a consequence, there is no requirement for trimming and transforming the RR data as done in parametric regression models discussed in chapter 2. Second, there is no need to assume any shapes for the functional forms for the responses of the covariates. The data-driven method estimates the underlying responses as functions of the covariates themselves, which facilitates the estimation of the idiosyncratic marginal and interaction effects. For example, the response of debt cushion on RR for a specific loan characteristics of each individual can be nonparametrically estimated, as well as the way in which this response varies over the entire DC range during economic upturns and downturns. These idiosyncratic effects may differ among individuals conditioning on their own loan's characteristics. The model would provide tools that lenders can use to design optimal treatment rules for a particular borrower.

Third, the proposed PLA-QR contains two functions, as the model's specification allows both nonlinear effect in the additive functions, and linear effect in the parametric functional form. It provides an alternative marginal effect analysis, which offers more general and straight forward estimates of the effect, due to its estimates of linear coefficients and the nonlinear additive functions. This might be useful for regulators, who are more interested in the overall effect of the covariate on RR, rather than the idiosyncratic effects analysis as banks do.

It is apparent that accurate RR predictions would aid banks and regulators in adequately quantifying credit risk. However, an imperative question is how banks can mitigate the expected credit risk exposure. Providing banks with information on the heterogeneous marginal and interaction effects of borrower characteristics on RR under certain economic conditions would constitute an answer. For example, if a defaulted borrower has an expected low level of RR, a treatment program could be initiated to reduce the potential loss to the bank. The treatment could be designed using the outcomes of the marginal effect analysis. However, the impact of a characteristic or an economic condition could depend on the values of other characteristics themselves as well as on the point of the quantile of the conditional distribution. Our analysis could be beneficial to developing appropriate policies to mitigate the underlying credit risk exposure, which in turn would improve risk management, risk monitoring, and credit risk pricing.

As we apply the proposed models to the realised RR from the Moody's Ultimate Recovery Database[1], we find evidence of nonlinearity and heterogeneity in the effects of the covariates on RR. A noteworthy point is that the empirical outcomes of the impact of economic downturns on RR provide detailed and distinct associations between the RR and economic downturns across the various quantiles of the distribution. This is much more informative than the average effect provided by the mean regression. As Basel requires banks to estimate the downturn RR and

[1] The full data description, summary statistics as well as the preliminary analysis are discussed and provided in section 3.4 of this thesis

thus downturn credit risk, the findings based on QR models are very beneficial to banks.

The remainder of chapter 4 is organised as follows: the next section explains the model specifications and the estimation methodologies. The results of the empirical study are reported and analysed in section 4.3. Concluding remarks are made in section 4.4.

4.2 Methodology

In what follows, we consider three different conditional quantile regressions: (i) linear quantile regression (L-QR); (ii) nonparametric quantile regression (NP-QR); and (iii) partial linear additive quantile regression (PLA-QR). One of the main differences between these three models is the degree of flexibility in estimating the nonlinear relationship. The fully nonparametric model does not require a presumed functional form. Also, we introduce the PLA-QR, which contains linear and nonlinear components, and we also propose an improved estimation method for bandwidth selection in PLA-QR. Lastly, we also consider the linear and partial linear additive models with back-transformation (L-QR(tr) and PLA-QR(tr), respectively) to address the boundary problem of RR data.

4.2.1 Linear quantile regression

Let the conditional distribution function be $F(y|x) = P(Y \leq y|x)$, and the conditional τ^{th} quantile of the conditional distribution function be defined as:

$$q_\tau(x) \equiv \inf\{q : F(q|x) \geq \tau\} = F^{-1}(\tau|x), \tag{4.2.1}$$

where $\tau \in (0,1)$, x is $k \times 1$ realization of the vector of covariates. One can estimate q in (4.2.1) by minimizing the following loss function:

$$\rho_\tau(u) = u(\tau - I(u < 0)), \tag{4.2.2}$$

where $I(\cdot)$ is the indicator function, $u = Y - q$, and then we find $\hat{q}$ minimizes expected loss. We seek to minimize:

$$E[\rho_\tau(Y-q)] = (\tau - 1)\int_{-\infty}^{q} (y-q)dF(y|x) + \tau \int_{q}^{\infty} (y-q)dF(y|x).$$

The first derivative of the expected loss function with respect to q is defined as:

$$0 = (1-\tau)\int_{-\infty}^{q} dF(y|x) - \tau \int_{q}^{\infty} dF(y|x) = F(q|x) - \tau,$$

where $F(q|x) = \tau$, and q is the estimator of the τ^{th} conditional quantile defined in (4.2.1). In the linear quantile regression, let

$$y = x'\theta_\tau + \varepsilon, \qquad \tau \in (0,1), \tag{4.2.3}$$

where x is a $k \times 1$ vector of independent variables, θ_τ is a vector of unknown parameters, and $P(\varepsilon \leq 0|x) = \tau$. Therefore, the conditional quantile is defined as:

$$q_\tau(x) = x'\theta_\tau, \tag{4.2.4}$$

where $q_\tau(x)$ is the τ^{th} conditional quantile of y given x. The parameter θ_τ is estimated by minimizing the following check function:

$$\hat{\theta}_\tau = \min_{\theta_\tau \in R} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\theta_\tau), \tag{4.2.5}$$

where the estimate of the τ^{th} conditional quantile in (4.2.1) is $x_i'\hat{\theta}_\tau$. This optimization problem can be solved efficiently using a linear programming method. The asymptotic covariance matrix of $\sqrt{n}(\hat{\theta}_\tau - \theta_\tau)$ using the asymptotic distribution of $\hat{\theta}_\tau$:

$$Q_n^{-1/2}\sqrt{n}(\hat{\theta}_\tau - \theta_\tau) \to N(0,(1-\tau)), \tag{4.2.6}$$

where $Q_n = H_n^{-1}J_nH_n^{-1}$, $J_n(\tau) = n^{-1}\sum_{i=1}^{n} x_ix_i'$, $H_n(\tau) = n^{-1}\sum_{i=1}^{n} x_ix_i'f_i(q_\tau(x_i))$, and $f_i(q_\tau(x_i))$ is the conditional density of y_i evaluated at the τ^{th} conditional quantile.

4.2.2 Nonparametric quantile regression

To nonparametrically estimate the quantile function, we apply two-stage estimation as follows:

(i) estimate the conditional distribution $\hat{F}(y|x)$ nonparametrically using the weighted Nadaraya-Watson method introduced by Q. Li and Racine (2008); and

(ii) estimate the conditional quantile $\hat{q}_\tau(x)$ through $\hat{F}(y|x)$ given the definition of the quantile function in (4.2.1).

In (i), the conditional distribution is estimated by smoothing both dependent and independent variables:

$$\hat{F}(y|x) = \frac{\sum_{i=1}^{n} G_{h_0}(Y_i,y)\mathbb{K}_H(X_i,x)}{\sum_{i=1}^{n} \mathbb{K}_H(X_i,x)}, \tag{4.2.7}$$

where $G(v) = \int_{-\infty}^{v} w(u)du$ is the distribution function associated with the density function $w(\cdot)$; h_0 is the bandwidth for Y_i; $X_i = (X_i^c, X_i^d)$ is a mixture of continuous and discrete covariates defined as $X_i^c \in \mathbb{R}^p$, and X_i^d is a vector of $r \times 1$ discrete variables; $H \in \{h, l\}$, which h and l are bandwidths associated with X_i^c and X_i^d respectively; and $\mathbb{K}_H(\cdot) = \kappa_h(X_i^c, x^c)\cdot\lambda_l(X_i^d, x^d)$ is the product of all kernel functions of continuous and discrete variables discussed in Chapter 3.

To select the bandwidths, we apply the data-driven least-square cross validation method (Q. Li, Lin, & Racine, 2013). In this method, the bandwidths are chosen by minimizing the following objective function:

$$CV(h_0, H) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left[I(y_i \leq y_j) - \hat{F}_{-i}(y_j | x_i) \right]^2 M_i, \tag{4.2.8}$$

where $I(\cdot)$ is an indicator function, M_i is a trimming parameter to ensure that the objective function CV is finite, and $\hat{F}_{-i}(\cdot)$ is the leave-one-out estimator of $F(y|x_i)$ defined as:

$$\hat{F}_{-i}(y_i|x_i) = \frac{\sum\limits_{j=1,\, j\neq i}^{n} G_{h_0}(Y_j, y_i) \mathbb{K}_H(X_j, x_i)}{\sum\limits_{j=1,\, j\neq i}^{n} \mathbb{K}_H(X_j, x_i)},$$

In the step (ii), we use the conditional distribution estimate ($\hat{F}(y|x)$) in (4.2.7) to define the conditional quantile in (4.2.1), which can be written as:

$$\hat{q}_\tau(x) = \hat{F}^{-1}(\tau|x) = \inf\{y : \hat{F}(y|x) \geq \tau\}. \tag{4.2.9}$$

As we directly estimate the conditional distribution, it allows us to estimate the conditional quantile by minimizing the following loss function:

$$\hat{q}_\tau(x) = \min_q |\tau - \hat{F}(q|x)|. \tag{4.2.10}$$

Advantages of NP-QR include the following aspects: (i) the model is fully flexible, and there is no need to impose assumptions; (ii) the model can capture the underlying nonlinearity in the relationship of RR-covariates for each individual; and (iii) the predicted values of y are bounded within the range of observed y.

The application of this nonparametric methodology is suitable for RR modelling in terms of capturing nonlinearity and addressing boundary problems. However, one of the main difficulties is obtaining the marginal effects of the independent variables, especially in the high dimensional covariates (De Gooijer & Zerom, 2003), as the marginal effect estimates are not directly available.

One of the aims of this chapter is to estimate the partial marginal effects. Therefore, we turn to an additive model, the specification of which facilitates the estimation of the partial marginal effects of the continuous variables. However, the traditional additive model normally does not accommodate categorical and dummy variables. As the set of variables that we use in the empirical study includes continuous and discrete variables, we adopt the partially linear additive quantile regression model proposed by Hoshino (2014) for our purposes.

4.2.3 Partially linear additive quantile regression

In this section, we specify the model and explain in detail the estimation method, which involves a two-stage process. We also discuss the ways in which we integrate the methods proposed by Hoshino (2014), Wang and Yang (2009), Horowitz and Lee (2005) and Q. Li and Racine (2008), and present methods to improve the two-stage estimation process of the PLA-QR model using least-squares cross-validation method for bandwidth selection.

Model specification and assumptions

Consider the PLA-QR model:

$$y = \mu_\tau + \sum_{j=1}^{t} m_{j,\tau}(x_j^m) + \sum_{j=1}^{s} z_j \beta_{j,\tau} + \varepsilon, \qquad \tau \in (0,1) \tag{4.2.11}$$

where $m_{1,\tau}(\cdot),\ldots,m_{t,\tau}(\cdot)$ are unknown continuous univariate functions[2], x is partitioned as $\{x^m, z\}$, $x_j^m (j = 1,\ldots,t)$ is a continuous variable assumed in the additive

[2] $m_{j,\tau}(\cdot)$ can accommodate only a continuous variable.

functions[3], $z_j (j = 1,...,s)$ denotes the j^{th} component of s remaining covariates which are all discrete variables and the remaining continuous variables, $t + s = k$, μ_τ is an unknown constant, $\beta_{j,\tau}(j = 1,...,s)$ is the j^{th} component of an unknown $(s \times 1)$ parameters vector β_τ, and $\varepsilon \equiv \varepsilon_\tau$ is the quantile error such that

$$P(\varepsilon \leq 0|x^m, z) = \tau, \text{ for } \tau \in (0,1).$$

Estimation method

This section describes a two-stage procedure for estimating $m_{j,\tau}(\cdot)$. Let us assume $t = p$, hence we include all continuous variables in the additive component. We assume that the support of x^m is $\mathcal{X}^m \equiv [-1,1]^t$, and normalize $m_1,...,m_t$ so that $\int_{-1}^{1} m_j(v)dv = 0$ for $j = 1,...,t$. As in Horowitz (2009, 2012), the assumption that x^m takes value in the compact set can be made without loss of generality, because it can always be satisfied by monotonically increasing function of the components of x^m. For notational simplicity, we write $m_\tau(x^m) \equiv \mu_\tau + \sum_{j=1}^{t} m_{j,\tau}(x_j^m)$, where $x^m = (x_1^m,...,x_t^m)'$ is a generic element in $\mathcal{X}$. The PLA-QR model in (4.2.11) can be written in a vector form as:

$$y = m_\tau(x^m) + z'\beta_\tau + \varepsilon, \qquad \tau \in (0,1).$$

We describe the two-stage estimation of $m_\tau(x^m)$:

Stage I:

Let $\{\Pi_d : d \in \mathbf{Z}^+\}$ denotes a complete orthogonal basis for smooth functions on [-1,1], where $\mathbf{Z}^+$ is a positive integer; see Horowitz and Lee (2005) for conditions that the basis functions must satisfy. For any positive integer d, we define:

$$\Pi_d(x^m) = [1, \pi_1(x_1^m), ..., \pi_d(x_1^m), \pi_1(x_2^m), ..., \pi_d(x_2^m), ..., \pi_1(x_t^m), ..., \pi_d(x_t^m)].$$

[3] Given that $x^c \in \mathbf{R}^p$ in (4.2.7), we define $t \leq p$, as it is not necessary to include all continuous variables in the additive component. This allows the flexibility in accommodating both linear and nonlinear effects of x^c

Then, for $\alpha_{d,\tau} \in \mathbb{R}^{dt+1}$, $\Pi_d(x^m)\alpha_{d,\tau}$ is a series approximation to $\mu_\tau + m_\tau(x^m)$. For sufficiently large d, α_τ is the set of $(dt+1)\times 1$ unknown parameter vector, and β_τ can be estimated by minimizing the following check function:

$$\{\hat{\alpha}_\tau, \hat{\beta}_\tau\} = \min_{\alpha_\tau,\beta_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \Pi_d(x_i^m)'\alpha_\tau - z_i\beta_\tau). \tag{4.2.12}$$

Note that we use B-spline function denoted as $\Pi_d(\cdot)$. $\Pi_d(x^m)'\hat{\alpha}_\tau$ is the stage I series approximation of $\mu_\tau + \sum\limits_{j=1}^{t} m_{j,\tau}(x_j^m)$ which is defined as $\tilde{\mu}_\tau + \tilde{m}_\tau(x^m)$. In other words, the series estimate $\tilde{m}_{j,\tau}(x_j^m)$ is the product of $\pi_1(x_j^m),\ldots,\pi_d(x_j^m)$ with the appropriate components of $\hat{\alpha}_\tau$ (Doksum & Koo, 2000). Finally, we choose an optimum value for integer d^* by selecting d that minimizes:

$$SIC(d) = \log\left(\sum_{i=1}^{n} \rho_\tau(\tilde{\varepsilon}_i)\right) + \frac{td\log n}{2n}, \tag{4.2.13}$$

where $\tilde{\varepsilon}_i = y_i - \Pi_d(x_i^m)'\hat{\alpha}_\tau - z_i\hat{\beta}_\tau$.

Stage II:

We describe the second stage estimate of $m_{j,\tau}(x_j^m)$ for $j = 1,\ldots,t$. To estimate the unknown additive function of $x_{j=\delta}^m$ which is $m_{\delta,\tau}(x_\delta^m)$, we define:

$$m_{-\delta,\tau}(\bar{x}^m) = \mu_\tau + m_{1,\tau}(x_1^m) + \ldots + m_{t,\tau}(x_t^m) \text{ with } m_{\delta,\tau}(x_\delta^c) \text{ being excluded,}$$

and its series approximation from the estimation in Stage I is:

$$\tilde{m}_{-\delta,\tau}(\bar{x}^m) = \tilde{\mu_\tau} + \tilde{m}_{1,\tau}(x_1^m) + \ldots + \tilde{m}_{t,\tau}(x_t^m) \text{ with } \tilde{m}_{\delta,\tau}(x_\delta^m) \text{ being excluded.}$$

Then, let us define, for $i = 1,\ldots,n$,

$$\begin{aligned} y_\delta &\equiv y - m_{-\delta,\tau}(\bar{x}^m) - z'\beta_\tau \\ &= m_{\delta,\tau}(x_\delta^m) + \varepsilon, \end{aligned} \tag{4.2.14}$$

and

$$\tilde{y}_\delta \equiv y - \tilde{m}_{-\delta,\tau}(\bar{x}^m) - z'\hat{\beta}_\tau.$$

From (4.2.14), it is clear that one can obtain a consistent estimate for $m_{\delta,\tau}(x_\delta^m)$ by a one-dimensional nonparametric quantile regression of y_δ on x_δ^m. However, as $y_{i,\delta}$ in (4.2.14) is unknown for $i = 1,...,n$, it can be replaced with $\tilde{y}_{i,\delta}, i = 1,...,n$ defined in (4.2.14).

Finally, we can estimate the unknown additive component $m_{\delta,\tau}(\cdot)$ by appling nonparametric quantile regression of $\tilde{y}_{i,\delta}, i = 1,...,n$. We employ the nonparametric quantile regression using weighted Nadaraya-Watson estimator proposed by Q. Li and Racine (2008). Furthermore, we introduce the least-square cross validation for bandwidth selection, rather than the plug-in bandwidths as suggested by Hoshino (2014). Thus, we estimate $m_{j,\tau}(x_{i,j}^m); j = 1,...,t$ using (4.2.7) and (4.2.8), which completes stage II estimation.

To describe the estimation of $m_{j,\tau}(x_{i,j}^m)$ given $\{\tilde{y}_{i,j}, x_{i,j}\}_{i=1}^n$ for $j = \delta$, applying (4.2.7) yields:

$$\hat{F}(\tilde{y}_\delta | x_\delta^m) = \frac{\sum_{i=1}^n G_{h_0}(\tilde{y}_{i,\delta}, \tilde{y}_\delta)\kappa_h(x_{i,\delta}^m, x_\delta^m)}{\sum_{i=1}^n \kappa_h(x_{i,\delta}^m, x_\delta^m)}, \tag{4.2.15}$$

where $\kappa_h(\cdot)$ is a univariate kernel function with bandwidth h. The bandwidth is selected via the least-square cross validation in (4.2.8). Then, we estimate the unknown function $m_{\delta,\tau}(\cdot)$ by:

$$\hat{m}_{\delta,\tau}(x_\delta^m) = \arg\min_{\hat{y}_\delta} |\tau - \hat{F}(\hat{y}_\delta | x_\delta^m)|. \tag{4.2.16}$$

One of the main advantages of the two-stage approach is that the asymptotic properties of the estimate $\hat{m}_{\delta,\tau}(x_\delta^m)$ can be established (Hoshino, 2014). It can be shown that:

$$n^{\frac{2}{5}}(\hat{m}_{\delta,\tau}(x_\delta^m) - m_{\delta,\tau}(x_\delta^m)) \rightarrow^d N(b_\tau(x_\delta^m), V_\tau(x_\delta^m)), \quad \text{for } \delta = 1,...,t, \tag{4.2.17}$$

$$b_\tau(x_\delta^m) = -\frac{1}{2}C_h^2\kappa_2 F_{\varepsilon,\delta}^{(2)}(0|x_\delta^m)/f_{\varepsilon,\delta}(0|x_\delta^m),$$

and

$$V_\tau(x_\delta^m) = \frac{v_0\tau(1-\tau)}{C_h f_{x_\delta^m}(x_\delta^m) f_{\varepsilon,\delta}^2(0|x_\delta^m)},$$

where $C_h = hn^{\frac{1}{5}}$, $\kappa_2 = \int_{-1}^{1} u^2 K(u)du$, $F_{\varepsilon,q}^{(2)}(\cdot|x_\delta^m)$ is the second derivative[4] of the CDF of ε conditional to x_δ^m, $f_{\varepsilon,q}(\cdot|x_\delta^m)$ is the PDF of ε conditional to x_δ^m, $f_{x_\delta^m}(x_\delta^m)$ is the PDF of x_δ^m, and $v_0 = \int_{-1}^{1} K(u)^2 du$. Asymtotic results are useful to conduct statistical inference.

Implementation of the PLA-QR method to recovery rate modelling

Let us assume the specification of the partially linear model in (4.2.11) as:

$$RR = \mu_\tau + m_{1,\tau}(DC) + m_{2,\tau}(SI) + Z\beta_\tau + \varepsilon, \tag{4.2.18}$$

where RR is the observed recovery rate, $m_{j,\tau}(\cdot); j = 1,2$, are the unknown additive components of the debt cushion (DC) and the stress index (SI), and $Z =$ (loan type, instrumental rank, collateral) is a matrix of three discrete variables. The above two-stage estimation process is implemented for model (4.2.18) in the following steps:

1. Estimate the unknown parameters of the stage I:

$$\{\hat{\alpha}_\tau, \hat{\beta}_\tau\} = \min_{\alpha_\tau,\beta_\tau} \sum_{i=1}^{n} \rho_\tau(RR_i - \Pi_d(X_i^m)'\alpha_\tau - Z_i\beta_\tau),$$

   where $\Pi_d(X_i^m) = (1, \pi_1(DC_i), ..., \pi_d(DC_i), \pi_1(SI_i), ..., \pi_d(SI_i))$, integer d is a degree of B-spline function that minimizes SIC in (4.2.13), and $\alpha_\tau = (\alpha_0, \alpha_{1,DC}, ..., \alpha_{d,DC}, \alpha_{1,SI}, ..., \alpha_{d,SI})'$ is $(2d+1)\times 1$ vector of unknown parameters associated with each component in $\Pi_d(X_i^m)$.

[4] Hoshino (2014) suggest linear power series regressions with the higher order than 3

2. Obtain the unknown oracle responses in (4.2.14) of the stage II using the estimates $\{\hat{\alpha}_\tau, \hat{\beta}_\tau\}$ in step 1:

   - $\tilde{RR}_{DC} = RR - \tilde{m}_{SI,\tau}(SI) - Z\hat{\beta}_\tau$,
     where $\tilde{m}_{SI,\tau}(SI) = (1, \pi_1(SI), ..., \pi_d(SI))\hat{\alpha}_{-DC,\tau}$,
     and $\hat{\alpha}_{-DC,\tau} = (\hat{\alpha}_0, \hat{\alpha}_{1,SI}, ..., \hat{\alpha}_{d,SI})$.
   - $\tilde{RR}_{SI} = RR - \tilde{m}_{DC,\tau}(DC) - Z\hat{\beta}_\tau$,
     where $\tilde{m}_{DC,\tau}(DC) = (1, \pi_1(DC), ..., \pi_d(DC))\hat{\alpha}_{-DC,\tau}$,
     and $\hat{\alpha}_{-SI,\tau} = (\hat{\alpha}_0, \hat{\alpha}_{1,DC}, ..., \hat{\alpha}_{d,DC})$.

3. Estimate the unknown additive functions of DC and SI by regressing DC on $\tilde{RR}_{DC}$, and SI on $\tilde{RR}_{SI}$, respectively, using one dimensional nonparametric quantile regression in (4.2.15) and (4.2.16). Also, the bandwidths are chosen by least-square cross validation in (4.2.8). The additive function estimates are denoted as $\hat{m}_{1,\tau}(DC)$ and $\hat{m}_{2,\tau}(SI)$.

4.2.4 Quantile regressions with logit back-transformation

To overcome the problems associated with the boundaries zero and one, a transformation can be applied to both linear models and partially linear additive models. In fact, the additive component in PLA-QR may mitigate the boundary issue of the estimates, as the model allows for some degree of nonlinearity in the effects of the continuous covariates. However, there is no guarantee that the conditional quantile function estimates will lie within the unit interval [0,1].

As discussed in the earlier chapters, the inequality resulting in the bias of back-transformation in mean regression has been criticised in chapter 2, section 2.3.1. Such a bias does not arise in the present study, as it focuses on QR modelling (Bottai, Cai, & McKeown, 2010). Thus, the following equality holds:

$$\tau = P(Y < y|x) = P(\Phi(Y) < \Phi(y)|x), \tag{4.2.19}$$

where $\Phi(\cdot)$ is the transformation function.

In this chapter, the transformation function is defined as:

$$\Phi(y) = \ln\left(\frac{y+\nu}{1+\nu-y}\right) = y^*, \tag{4.2.20}$$

where $-\infty < y^* < \infty$ is the transformed y, and ν is an arbitrary positive value. The transformation function in (4.2.20) is in fact the logit function with an adjustment (ν), which ensures the validity of transformation since $y \in [0,1]$. With this logit transformation function, we can derive its inverse function:

$$\Phi^{-1}(y^*) = \frac{(1+\nu)exp(y^*)-\nu}{1+exp(y)} = y, \tag{4.2.21}$$

where $\Phi^{-1}(\cdot)$ is a logistic function that takes into account of the adjustment value ν, rather than trimming the boundaries of one and zero discussed in chapter 2, section 2.3.1. In what follows, the linear model with the logit transformation is explained. This is followed by the partially linear additive model with the transformation.

Linear quantile regression model with logit transformation (L-QR(tr))

The linear quantile regression with the logit transformation of y can be specified as:

$$\Phi(y) = x'\varphi_{\Phi,\tau} + \epsilon, \quad \text{or} \quad q_{\Phi,\tau}(x) = x'\varphi_{\Phi,\tau}, \tag{4.2.22}$$

where $P(\epsilon < 0|x) = \tau$, and $q_{\Phi,\tau}(x)$ is a conditional quantile of the $\Phi(y)$ given x. To estimate $\hat{\varphi}_\tau$, we regress the $\Phi(y)$ on the set of covariates x using the linear quantile regression discussed in section 4.2.1.

Due to the invariant property of the probability distribution in (4.2.19), it can be shown that:

$$\tau = Pr(Y \leq q_\tau(x)) = Pr(\Phi(Y) \leq q_{\Phi,\tau}(x)),$$

then we immediately have:

$$Pr(Y \leq q_\tau(x)) = Pr(Y \leq \Phi^{-1}(q_{\Phi,\tau}(x)),$$

where $\Phi^{-1}(\cdot)$ is an inverse of the transformation. Hence, we can applied the logistic transformation function in (4.2.21) to $\hat{q}_\tau(x)$:

$$\Phi^{-1}(\hat{q}_{\Phi,\tau}(x)) = \frac{(1+\nu)exp(x'\hat{\varphi}_{\Phi,\tau}) - \nu}{1 + exp(x'\hat{\varphi}_{\Phi,\tau})} = \hat{q}_\tau(x), \tag{4.2.23}$$

where $\hat{q}_{\Phi,\tau}(x) = x'\hat{\varphi}_{\Phi,\tau}$. Therefore, this allows us to estimate the conditional quantile of the bounded [0,1] RR.

Partially linear additive model with logit transformation (PLA-QR(tr))

We can also apply PLA-QR in (4.2.11) to $\Phi(y)$, defined as:

$$\Phi(y) = m_{\tau,\Phi}(x^m) + z\beta_{\tau,\Phi} + \epsilon, \quad \text{or} \quad q_{\Phi,\tau}(x) = m_{\tau,\Phi}(x^m) + z\beta_{\tau,\Phi}, \tag{4.2.24}$$

where $m_{\tau,\Phi}(x^m) = \mu_{\tau,\Phi} + \sum_{j=1}^{t} m_{j,\tau,\Phi}(x_j^m)$. Hence, the two-stage estimation of the additive model can be employed to $\Phi(y)$ in (4.2.20). Then, the same transformation procedure as the linear model is applied where the estimation of the PLA-QR with the transformation is $\hat{q}_\tau(x) = \Phi^{-1}(\hat{m}_{\tau,\Phi}(x^m) + z\hat{\beta}_{\tau,\Phi})$.

4.2.5 Model selection criteria for quantile regressions

In this section, we explain several criteria used to evaluate the performances of our QRs. First, we compare the conditional quantile goodness of fit of all included QRs at the given quantiles. Second, each model is evaluated based on the difference between the sample and predicted distributions. Finally, we evaluate the models with the Value at Risk framework.

Quantile regression goodness of fit

In order to evaluate the goodness of fit for the quantile regression, the pseudo R^2 is used. The pseudo R^2 is defined as:

$$\tilde{R}^2 = 1 - \frac{\hat{S}}{\bar{S}}, \tag{4.2.25}$$

where $\hat{S} = \sum_{i=1}^{n} \rho_\tau(y_i - \hat{\beta}_\tau x_i)$ is the sum of the residuals using the loss function in (4.2.2), and $\bar{S} = \sum_{i=1}^{n} \rho_\tau(y_i - q_\tau(y))$ is the sum of the loss function evaluated at $q_\tau(y)$, which is the τ^{th} sample quantile of y. This is similar to the standard R^2, which is one minus the sum of squares of residuals over the total sum squares at the mean. However, the pseudo R^2 constitutes a local measure of goodness of fit for a particular quantile (Koenker & Machado, 1999) instead of an average.

Point prediction of quantile regression

It is possible to apply the QR at a particular quantile for prediction purposes. The median regression[5], for example, provides the conditional median estimate, which is commonly expected to be more robust to outliers than a conditional mean regression. The standard criteria to evaluate the point prediction, such as mean squared error (MSE) and mean absolute error (MAE), can be employed to compare the predictability of the conditional mean and median regressions (Siao et al., 2015).

In our study, the QR's predictions are generated by allowing $\tau = \tau_i$, where τ_i is a sample quantile of y_i associated with realizations $y_1, ..., y_n$. Let us define the sample quantile as:

$$\tau_i = \frac{1}{n} \sum_{j=1}^{n} I(y_j \leq y_i),$$

[5] The conditional median regression is estimated by specified $\tau = 0.5$ in the QR's specification

where $I(\cdot)$ is indicator function. Then, we denote:

$$\begin{aligned} MSE^{(q)} =& \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{q}_{\tau=\tau_i}(x_i))^2, \text{ and} \\ MAE^{(q)} =& \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{q}_{\tau=\tau_i}(x_i)|, \end{aligned} \tag{4.2.26}$$

where $\hat{q}_{\tau=\tau_i}(x_i) = \hat{y}_i$ is the predicted y_i from a given quantile regression given $\tau = \tau_i$ and x_i. For example, if we observe that y_i is a sample median ($\tau_i = 0.5$), then the prediction of y_i will be generated by a median regression, $\hat{q}_{\tau=0.5}(x_i)$. These allow us to compare the predictive errors of the given quantile regressions, namely L-QR, NP-QR, and PLA-QR, for the various quantiles. Importantly, we note that this is only for model selection purposes among the quantile regressions, as the additional information of τ_i remains unknown for prediction purposes in practice.

Distributional fit of the quantile regression

To evaluate the sample distribution of QR predictions defined in (4.2.26) in comparison to the observed sample distribution of y, let $u = (u_1, ..., u_{2n})'$ be a $2n\times 1$ vector of ordered $\{y_{(1)}, ..y_{(n)}, \hat{y}_{(1)}, ..., \hat{y}_{(n)}\}$, and define:

$$\begin{aligned} F(u_j) =& \frac{1}{n}\sum_{i=1}^{n} I(y_i \leq u_j), \text{ and} \\ \hat{F}(u_j) =& \frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \leq u_j), \end{aligned} \tag{4.2.27}$$

where $j = 1, ..., 2n$, $F(\cdot)$ and $\hat{F}(\cdot)$ are the sample quantiles associated with the realisations of $y_1, ..., y_n$ and $\hat{y}_1, ..., \hat{y}_n$, respectively. Then, the distributional fit of QR can be seen as the difference between $F(u)$ and $\hat{F}(u)$ which is measured by:

$$HWMI = \frac{1}{2n}\sum_{j=1}^{2n}|F(u_j) - \hat{F}(u_j)|, \tag{4.2.28}$$

where HWMI stands for Harmonic Weighted Mass Index for two finite sample distributions (Hinloopen, Wagenvoort, & van Marrewijk, 2012). High values of HWMI indicates that the underlying samples are less likely to be drawn from the same distribution.

Furthermore, given the definition in (4.2.27), the Kolmogorov-Smirnov (KS) test can be performed to evaluate whether the sample and estimated distributions have a common distribution (Krüger & Rösch, 2017). The KS test statistic is:

$$D = \max_{j=1,\dots,2n} |F(u_j) - \hat{F}(u_j)|, \tag{4.2.29}$$

where the critical value is $D_\alpha = c(\alpha)\sqrt{\frac{2}{T}}$, $c(\alpha) = \sqrt{-0.5\ln(\frac{\alpha}{2})}$, $T = 2n$, and α is a level of significance. The null hypothesis that $F(\cdot)$ and $\hat{F}(\cdot)$ have a common distribution is rejected, if the maximum difference between both of them measured by D is larger than the critical value D_α.

Value at Risk evaluation

For the credit risk management purposes of RR modelling, the lower quantile of the RR distribution would be of most interest to lenders and regulators, as it shows the extreme losses that these lenders would experience. This leads to the application of QR at the lower-specified quantiles, also known as the Value at Risk (VaR) of defaulted loan recoveries. To estimate VaR of RR, this chapter considers QRs at the lower tail quantiles of $\bar{\tau} = \{0.05, 0.10, 0.15, 0.20\}$.

The VaR for a given $\bar{\tau}$ quantile represents that there is $(1-\bar{\tau})\cdot 100$ percent chance that we will underestimate RR. As far as the risk exposure is concerned especially during the economic downturns, we would prefer the model that generates the low percentage of the overestimations. Therefore, we can evaluate the VaR prediction by:

$$HR_{\bar{\tau}} = \frac{1}{n}\sum_{i=1}^{n} I(\hat{q}_{\bar{\tau}}(x_i) < y_i) \cdot 100\%, \tag{4.2.30}$$

where $HR_{\bar{\tau}}$ denotes as a hit rate of predicted VaR for a given $\bar{\tau}$ which is the percentage of underestimated RR, $\hat{q}_{\bar{\tau}}(x_i)$ is a predicted VaR using a QR with $\tau = \bar{\tau}$. The model with a hit rate which has the smallest distance between the predicted hit rate and the expected rate of $(1 - \bar{\tau}) * 100\%$ is the most preferable (Krüger & Rösch, 2017).

4.3 Empirical results

The dataset used in this empirical application is described in chapter 3, section 3.4. In this section, the NP-QR and PLA-QR models are applied to RR data at the three quantiles $\tau = \{0.25, 0.5, 0.75\}$. In what follows, we discuss these results.

4.3.1 Empirical results of nonparametric quantile regression

The conditional distribution and the quantile functions of the RR are nonparametrically estimated given the five determinants as specified in section 3.4. We then study the marginal and interaction effects among RR covariates and the way these effects change over the various quantiles of the conditional distribution. As NP-QR does not directly provide such information, we infer the effects by the following steps. First, we hold the categorical variables constant at three levels of risk characteristics[6]: high-, medium-, and low-risk. Second, the RR is estimated at various values of DC and SI. In doing so, it allows us to analyse the idiosyncratic marginal and interaction effects of these three given risk characteristics loans. Finally, we plot the estimated RR as a function of DC and SI, in order to illustrate the conditional effect of a particular variable on the given loans recoveries.

Table 4.1 reports the selected bandwidths by the least-square cross validation method in (4.2.8). Then, the conditional quantile functions are estimated at $\tau = \{0.25, 0.5, 0.75\}$, which represents lower, median and upper quantiles, by

[6] These variables are collateral status, instrumental rank and the types of loan. The definition of each risk characteristic will be provided later in this section

minimizing the objective function (4.2.10). The following results of the NP-QR provide the more complete picture of the relationships of RR-covariates on the RR of the loans with low-, medium-, and high-risk characteristic, respectively.

The effects of debt cushion and stress index on low-risk characteristic loan

To explain the impact, we consider the collateralised revolving loan with rank 1, which is specified by three categorical variables: Col = 1, Type = 2 and Rank = 1; and define it as a *low-risk characteristic loan*[7]. Figure 4.1 depicts the contours for all three quantiles, $\tau = \{0.25, 0.5, 0.75\}$. It clearly shows that the effects of DC and SI are nonlinear and their nonlinear shapes have changed across three different quantiles.

[————— Insert [Figure 4.1] here —————]

(i) The effect of debt cushion

To illustrate the effect of DC in Figure 4.1 , we let the SI be fixed as SI = {-1.0, -0.5, 0.0, 0.5, 1.0}, reflecting various economic conditions[8]. Figures 4.2a to 4.2c illustrate the effects of DC on the low-risk characteristics loan recoveries at the 0.25th, 0.5th, and 0.75th quantiles, respectively.

[————— Insert [Figure 4.2] here —————]

At the 0.25th quantile, as presented in 4.2a, a nonlinear positive effect of DC is observed and is similar in all given economic conditions. Debt cushion has the weakest impact on the RR for DC < 0.2, while a positive but increasing effect is found for the other range of DC. An increase in DC from 0 to 0.2 in Figure 4.2a leads to an increase in RR of the low-risk loans during neutral conditions (SI = 0) by only 0.1, while the RR increases by 0.4 with an increase in the DC from 0.4 to 0.6. The similar nonlinear effects of DC are observed for the median regression in

[7] In chapter 3, section 3.4 on data description, we found that the collateralised revolving loan with rank 1 has the highest RR on average

[8] The negative and positive SI are the proxies of economic upturn (SI <0), and downturn (SI > 0), respectively. See section 3.4 for the detailed explanation.

Figure 4.2b. We also find that the given defaulted loan is more likely to achieve full RR if its level of DC is greater than 0.6 at the 0.25th quantile, and greater than 0.4 at the median, except during the downturn economic (SI > 0).

At the upper 0.75 quantile, however, Figure 4.2c shows that DC mostly has no effect on RR of the low-risk loans, and the full RR is expected at any level of DC. Although the RRs of these loans are slightly less than one during the economic downturn, they are highly responsive to changes in DC. Our finding in Figure 4.2 suggests that lenders would be advised to pay more attention to loans with DC > 0.2 at the 0.25 quantile and the median, as a further increase in level of DC is more effective to improve the loan recoveries.

(ii)The effect of stress index

[———— Insert [Figure 4.3] here ————]

Figure 4.3 represents the effects of SI on the low-risk characteristic loan conditional on five fixed levels of DC = {0.00,0.25,0.50,0.75,1.00}. Figures 4.3a to 4.3c illustrate the effects of SI at the 0.25, 0.50 and 0.75 quantiles, respectively. Stress index mostly has a negative impact on RR, which is in line with expectations. On the other hand, only when the loan has a very high level of DC, especially when DC = 1, there is no effect of SI on RR. The low-risk characteristics loans with a DC of 1 are most likely to have a full RR at all conditional quantiles and are not sensitive to changes in economic conditions.

For the negative effect of SI, we find that the negative SI typically has a stronger effect than positive SI. At the lower 0.25 quantile, with DC fixed at 0.00, 0.25, and 0.50 in Figure 4.3a, an increase in SI from -1 to 0 causes a drop in RR by 0.3, while the RR decreases by only 0.1 as SI increases from 0 to 1. This behaviour is also consistent with the results at the median when the given DCs are 0 and 0.25 in Figure 4.3b. This implies that the RR is more sensitive to changes in economic stress during an upturn than changes during a downturn, especially when the loan has relatively low DC. On the other hand, the change in economic condition does not affect the RR at the upper 0.75 quantile as observed in Figure 4.3c.

The nonlinear effect of SI can be explained, as banks may impose more conservative policies during the downturn to mitigate the systematic risk. This then leads to the reduced impact of economic conditions on the given loans recoveries. We also find that the effect of SI seems to be weaker at the upper 0.75 quantile than at the median and lower quantiles, as well as the loan with higher level of DC.

Effect of debt cushion and stress index on the recovery rate for medium- and high-risk characteristic loans

We now consider the conditional effects of DC and SI on loans with other risk characteristics. *Medium*- and *high-risk characteristic loans* are specified as (i) uncollateralised revolving loans with Rank 1 (Col = 0, Type = 2, Rank = 1); and (ii) uncollateralised senior bonds with Rank 4 (Col = 0, Type = 5, Rank = 4), respectively[9]. To analyse the effects on RR of both specified loans, we employ the same procedure as the low-risk characteristic loans.

(i)The effect of debt cushion

In terms of the effects of DC at the 0.25, 0.5, and 0.75 quantiles, the findings and explanations of each quantile of the medium-risk characteristic loans (figures 4.4a to 4.4c) are mainly similar to those of the low-risk characteristic loans in the previous analysis (figures 4.2a to 4.2c). An increase in the low range of DC < 0.2 typically has less effect on RR than an increase in the higher ranges. Only when we consider such effect at 0.25 quantile during the economic upturn conditions (SI = {-1,-0.5}), the effect of the low range DC is as high as the other ranges in Figure 4.4a.

[9] This is based on the data description and preliminary analysis in chapter 3 which suggest that: (i) the uncollateralized loan is riskier than collateralized loan; (ii) revolving loan is less risky than the bond; and (iii) the loan with rank 1 is less risky than that with rank 4.

We find that it is more difficult for medium-risk characteristic loans to reach the full RR by increasing DC than it is for the low-risk characteristic loans. For the medium risk loans at the 0.25 quantile during a high-stress economic condition where SI = {0.5,1.0} in Figure 4.4a, an increase in DC does not appear to lead to the full RR as found in low-risk loans. The maximum RR is approximately 0.70-0.75 during those economic conditions.

[———— Insert [Figure 4.4] here ————]

The effect of DC for high-risk characteristic loans, shown in Figures 4.4d to 4.4f, differs substantially from its effects on low- and medium-risk characteristic loans. This would imply that the effect of DC on RR might depend on type of loan and/or instrumental rank [10]. At the 0.25 quantile in Figure 4.4d, defaulted loans with DC < 0.5, their recoveries are nearly zero (RR = 0), and an increase in DC does not improve the RR of the given loan in all economic conditions. Then, positive effects are observed when DC is greater than approximately 0.6, with a strong effect during an economic upturn (SI = {-1,-0.5}). However, in practice, it is less likely for loans with the given high-risk characteristics to have a DC greater than 0.5. Our finding implies that DC would not be an effective tool to improve RR for high-risk characteristic loans at the lower quantile.

For the median quantile in Figure 4.4e, although a change in DC between 0 and 0.4 tends to have no effect on the RR, the conditional effect of DC is positive with an increasing rate as DC increases from 0.4. Additionally, we find that the effect of DC on high-risk characteristic loans RR at the upper 0.75 quantile (Figure 4.4f) is similar to that on medium-risk characteristic loans at the median (Figure 4.4b).

(ii) The effect of stress index

The effects of SI, conditional on fixed levels of DC for loans with medium- and high-risk characteristics, are illustrated in Figure 4.5. Overall, the result is still

[10] The low- and medium-risk characteristics are revolving loans with Rank 1, while the high-risk characteristic is a senior bond with rank 4

consistent with the previous findings, where the effect of negative SI is mostly stronger than that of positive SI, especially when the effect is conditional on fixed levels of DC less than 0.2 in Figures 4.5a to 4.5c, 4.5b and 4.5e.

There are some exceptions for the nonlinear effect of SI on the high-risk characteristic loans recoveries. At the 0.25 quantile in Figure 4.5d, there is no effect of SI on RR of the high-risk characteristics loan with the given DC less than 0.5. Although, there is a strong negative impacts of SI on RR of the loan with DC = {0.75,1} are observed in Figure 4.5d, the high-risk characteristic loans with DC higher than 0.5 is not observed in practice.

Moreover, for the effect of SI at the 0.75 quantile in Figure 4.5f, the negative effect of the positive SI tends to be stronger than the negative SI, which is different from most findings in low- and medium-risk characteristics. This behaviour is also found in the effect of SI given high level of DC in Figure 4.5a and 4.5b for medium-risk characteristics at 0.25 and 0.50 quantiles respectively. These results imply that these loans are highly sensitive to the economic downturn, while the smaller effects are expected for the other loans.

[————— Insert [Figure 4.5] here —————]

Notably, the marginal effect analysis using NP-QR can be extended to reveal the idiosyncratic effect of the RR covariates on other given risk characteristics. This provides the lender with a comprehensive analysis of the specific defaulted loan which reveals various complex forms of the RR covariates' effects: nonlinear marginal effects, interaction effects, and heterogeneous effects. Such an analysis would be useful for lenders to understand each defaulted loan in their portfolio. On the other hand, we also provide, in the following section, an alternative analysis of the marginal effects using PLA-QR.

4.3.2 Empirical results of the partially linear additive quantile regression

In this section, we discuss the estimation results of the two-stage estimation method of PLA-QR model, which contains linear and additive components, followed by the marginal effect analysis using those components. The first-stage estimation of PLA-QR begins with the application of the B-spline function to approximate the nonlinear relationships of DC and SI in (4.2.13). The result is reported in the first row of Table 4.2, Panel (A). It shows that SICs are minimized as the degrees of the spline functions for DC are d = 9, d = 5, and d = 6 for the 0.25, 0.5, and 0.75 quantiles, respectively; and d = 1 is selected for SI in all quantiles. This suggests that the effect of SI is approximately linear, which is similar to our result in chapter 3.

Comparing this first-stage estimation of the PLA-QR in Table 4.2, Panel (A), with the linear model in Panel (B), the SICs of the proposed models at all given quantiles are smaller than those of the linear quantile regression model. These results indicate that some improvements can be made when the nonlinear relationships are estimated via the spline functions. Therefore, we specify our PLA-QR model specification as follows:

$$q_{RR,\tau}(x) = \mu_\tau + m_\tau(DC) + Z\beta_\tau, \tag{4.3.1}$$

where $q_{RR,\tau}(x)$ is the τ^{th} conditional quantile of RR, Z is a vector of covariates {SI, types of loan, instrumental ranks, collateral status}, which are assigned in the parametric component, β_τ is a vector of unknown parameters for the linear component, and DC is assigned to the additive component $m_\tau(\cdot)$. Due to the model structure, the marginal effect analysis is more transparent and general than the previous analysis using NP-QR. The insight idiosyncratic effects are generalized to the estimates of parametric coefficients and additive functions.

[————— Insert [Table 4.2] here —————]

Marginal effect estimates of linear components

Given the selected degrees of the spline functions, consistent parameter estimates are provided for the variables in the linear component, which are the marginal effects of categorical variables and SI. Negative relationships are consistently found for SI and the instrumental ranks, while a positive relationship is found for the collateral status in Table 4.2, panel (A). These relationships are mostly in line with expectations. However, we observe some heterogenous effects: first, there are substantial differences in the impacts of SI at the 0.25th and 0.75th conditional quantiles. The negative effect of SI is relatively small and insignificant at the 0.75 quantile compared to the median and lower quantiles. These findings reveal that loans at the lower quantile of RR tend to be more sensitive to the economic environment compared to those in the upper parts of the distribution. This could imply that the decrease in overall RR during an economic downturn is caused mainly by loans at the lower quantiles. The effect of the macroeconomic conditions impacts the level of RR differently, depending on the location of the distribution. These findings are also consistent with the standard linear quantile regression model in Table 4.2, panel (B), although the estimated parameters of SI are insignificant at all three quantiles.

Second, we expect that the higher the instrumental rank, the more risky the borrower is expected to be, which lowers RR. This intuition is reflected in our empirical analysis in Table 4.2, especially at the median and 0.75 quantiles. The RR of loans at the lower 0.25 quantile with Rank 2-4 is expected to be lower than the RR of those with Rank 1 by approximately 0.13. This observation implies that the effects of instrumental rank tend to be binary (either rank 1 or not) at the lower quantile. On the other hand, any increase in rank causes a lower RR only in the median and upper quantiles. For example, borrowers at the median and upper quantiles with Rank 4 have the lowest RR, followed by those with Ranks 3, 2, and 1 respectively, given that other variables are held constant.

Third, loans with collateral are intuitively expected to have higher RRs than those without, and our results in Table 4.2 indeed show what is expected. The parameter estimates of collateral are negative at all three conditional quantiles. However, the marginal effect estimates of collateral on RR are relatively high at the lower quantile, and become smaller magnitudes in the median before insignificant at the upper 0.75 quantile. Specifically, loans with collateral are expected to have an RR that is 0.17 higher than those of loans without collateral for lower 0.25 quantiles, while there is no significant difference between these loans at the 0.75 quantile. This implies that collateral status has an insignificant effect on RR level for loans in the higher quantiles.

Fourth, loan types impact RRs differently depending on the level of quantile in Table 4.2. Compared to term loans (baseline), revolving loans have significantly higher RRs than term loans at the 0.25 and 0.5 quantile regressions, while significantly lower RRs are expected for bonds at the median and upper quantiles.

Marginal effect estimates of additive components

[———— Insert [Table 4.3] here ————]

The second-stage estimation focuses solely on DC in the additive component. Table 4.3 shows the optimal bandwidth based on least-squares cross-validation in (4.2.8), which is employed to estimate the additive component. The selected bandwidths are reported in Table 4.3. Then, we proceed to the second-stage estimation based on the optimal bandwidths to estimate the additive function using the nonparametric quantile regression estimate given in (4.2.15).

[———— Insert [Figure 4.6] here ————]

Figure 4.6 represents the additive component estimates across the different quantiles, along with the 95% bootstrapping confidence intervals. The figures clearly show a nonlinear relationship between DC and RR with different shapes depending on the level of the conditional quantiles, which diverge notably from the linear relationship. The results show that an additional increase in DC has no impact on RR when DC ranges between [0.0,0.2] and [0.6,1.0] for the 0.25 quantile; between [0.0,0.1] and [0.5,1.0] for the median quantile; and between [0.3,1.0] for the 0.75 quantile. These findings are somewhat similar to the analysis in NP-QR in terms of the functional shapes: a change in the low levels of DC ineffectively increases the RR in the lower quantiles. Specifically, we find that an increase in DC of [0.0,0.2] has the most effect on loans at the upper quantiles.

Our finding suggests that for the observed heterogeneous and nonlinear effects of DC, the strength of the effects depends on the conditional quantile as well as on the level of DC itself. We find that an increase in DC from 0.0 to 0.4 leads to an increase in RR at the median and 0.75 quantile by 0.2, compared to that of the lower 0.25 quantile by 0.1. This implies that the effect of the given DC is stronger for the upper quantiles. On the other hand, if we consider an increase in DC between 0.4 and 0.6, this leads to a 0.3 increase in RR for the 0.25 quantile, a 0.2 increase for the median quantile, and no change for the 0.75 quantile. Thus, the effect of the higher range of DC at the low quantile is strongest.

Boundaries issue of PLA-QR compared to the linear quantile regression

We now turn to problems associated with boundaries. As discussed earlier, RR is bounded in the [0,1] unit interval. We find that the nonlinear effect of DC, which is captured by the additive component, ensures that most estimates will fall between zero and one. Table 4.4 reveals the degree of the boundary problems in PLA-QR compared to L-QR. Overall, PLA-QR has consistently lower percentages of negative predicted RRs than L-QR, while comparable percentages of estimates exceeding full RR (1) are observed in both models. However, the upper and lower bounds for the additive models are substantially closer to the unit interval than those for the linear model. Specifically, the boundaries of the fitted 0.25, 0.5, and 0.75 additive quantile models are $[-0.0170_{(0.9\%)}, 1.0006_{(0.5\%)}]$; $[0.0100_{(0.6\%)}, 1.0349_{(4.10\%)}]$; and $[0.2627_{(0.25\%)}, 1.0138_{(22.22\%)}]$, respectively, where the percentages in brackets are the number of estimates exceeding the boundaries. The lower and upper bounds of the additive models are approximately five times and two times smaller than those of the linear model. Although we find that 22% of the PLA-QR estimates at 0.75 quantile are greater than one, the maximum of value of the estimates is 1.01, compared to 1.03 for L-QR.

[————— Insert [Table 4.4] here —————]

4.3.3 Empirical results of the partially linear additive regression on transformed recovery rate

In order to overcome the boundary problems, we introduce a model with a logit transformation of RR to ensure that the predicted conditional quantile RR would lie in the [0,1] interval. As the set of RR determinants are regressed on the transformed RR, the parameter estimates reflect the relationships between the covariates and the transformed RR.

[————— Insert [Table 4.5] here —————]

Table 4.5 shows the result of the first-step estimation. Firstly, the SIC suggests the degrees of B-spline as d=2 for the 0.25 quantile and d=3 for both the 0.5 and 0.75 quantiles for DC, and d=1 for SI in all cases. These degrees are lower than those of the models with no transformation. Secondly, the signs of the estimators in the linear component of PLA-QR(tr) in Table 4.5 are consistent with the results of PLA-QR (Table 4.2). Specifically, in Table 4.5, negative relationships are consistently observed for SI and rank, while collateral shows negative signs across all quantile regressions.

[———————— Insert [Figure 4.7] here ————————]

Given the output of the first-step estimation, Figure 4.7 illustrates the additive estimates of DC[11]. The nonlinear shapes of the additive component estimates across the three quantiles are comparable to the findings provided by PLA-QR in Figure 4.6.

4.3.4 Evaluation of the quantile regressions performance

In this section, the performances of NP-QR, PLA-QR, L-QR, PLA-QR(tr), and L-QR(tr) are compared according to various aspects, including their goodness of fit, predictability, distributional fit, and predicted VaR evaluation.

Goodness of fit

The goodness of fit of each model is measured by an average residuals using the loss function in (4.2.2) and $\tilde{R}^2$ in (4.2.25). Overall, Table 4.6 suggests that the nonlinear quantile regressions outperforms the linear model. NP-QR provides the best goodness of fit in both criteria in all three quantiles, as it yields the lowest average errors and the highest $\tilde{R}^2$, followed by PLA-QR. Although the PLA-QR(tr) can overcome the PLA-QR's boundary issue, the goodness of fit of both models is comparable. PLA-QR(tr) slightly outperforms PLA-QR in $\tilde{R}^2$.

[———————— Insert [Table 4.6] here ————————]

Point prediction performance

As discussed in (4.2.26), we generate the predicted RR for the entire family of empirical RR quantiles given $\tau = \tau_i$. The predicted RR of each model is then denoted as $\hat{y}_i = \hat{q}_{\tau=\tau_i}(x_i)$. Table 4.7 shows the MSE and MAE of the in-sample and out-of-sample predictions. To define the in- and out-of-sample data, the split was made by randomly partitioning 30% of the full data and defining it as the

[11] This is the effect of DC on the transformed RR

out-of-sample data, with the remaining observations forming the in-sample data. The result shows that NP-QR and PLA-QR provide the most precise predictions for the in-sample data. On the other hand, NP-QR provides the least accurate out-of-prediction, while PLA-QR's predictive error remains small. We also find that L-QR offers a lower out-of-sample MSE and similar MAE when compared to PLA-QR. Moreover, we find that PLA-QR(tr) and L-QR(tr) provide the least accurate predictions.

[———— Insert [Table 4.7] here ————]

Distributional fit

Table 4.8 reports the HWMI of all the included models to measure the distributional fit of the RR predictions. It shows that our proposed PLA-QR has the best distributional fit for the in-sample data, followed by L-QR and NP-QR, respectively. On the other hand, L-QR yields the highest out-of-sample HWMI, followed by PLA-QR. Regarding the models with back-transformation, the performances of both PLA-QR(tr) and L-QR(tr) are consistently poor. In terms of the hypothesis test, for which the KS test was used, our results suggest that the null hypothesis of all models is rejected, due to the high KS statistics. The sample distribution differs significantly from the predicted distributions generated by all models. However, Krüger and Rösch (2017) caution against the application of the KS test, as the test statistic only considers the maximum deviance and fails to take into account the entire distribution.

Value at Risk evaluation

[———— Insert [Table 4.9] here ————]

To estimate the VaR, we consider the various lower conditional quantile regressions, where $\tau = \{0.05, 0.1, 0.15, 0.2, 0.25\}$, using all five included models. The hit rate of each model is then calculated using (4.2.30), which is compared with the expected hit rates are 95%, 90%, 85%, 80%, and 75%, respectively. The results are shown in Table 4.9. They indicate that NP-QR offers the best hit rates in generating the VaR at all five given quantiles, while PLA-QR(tr) provides relatively poor hit rates. Among the given five conditional quantiles, the VaR estimates at the 0.2 and 0.25 quantiles of NP-QR have hit rates of approximately 79% and 76%, which are almost the same as the expected hit rates (80% and 75%, respectively); followed by PLA-QR and L-QR. Furthermore, we find that the VaR estimates generated by L-QR(tr) at $\tau = 0.05, 0.1, 0.15$ have noticeably high hit rates, which are better than those of all models except for NP-QR.

4.4 Conclusion

This study proposes conditional nonparametric regression and partially linear additive quantile regression models, denoted as NP-QR and PLA-QR, respectively, and investigates which model(s) provide(s) a complete picture of the relationships between the response variable, RR, and the determinants over the entire conditional probability distribution. The conditional QR models are flexible and constitute improvements of the conventional conditional mean regression studied in chapter 3, which focuses on modelling the central tendency. Moreover, we assess their performance in four aspects including goodness of fit, in- and out-of-sample predictions, distributional fit, and the value at risk. Our findings deepen the understanding of the effects of covariates on RR which are found to be nonlinear, dependent on other variables (interaction), and heterogeneous across different quantiles.

The effect of debt cushion (DC) is nonlinear, where RR is less responsive to an increase in DC when DC < 0.2, especially at the lower quantiles. Although a positive effect is observed, the strength of the effect varies idiosyncratically and is strongly dependent on the other risk characteristic variables. The analysis using

NP-QR indicates that the effect of DC is much weaker for defaulted loans with high-risk characteristics (unsecured senior bonds with Rank 4) compared to its effect on low- and medium-risk characteristic loans (collateralised and uncollateralised revolving loans with Rank 1, respectively). The RR of the high-risk characteristic loans do not response to the change in DC when DC < 0.4, at the 0.25, median, and 0.75 quantiles.

The results of PLA-QR suggests that the effect of SI on RR is strong only at the lower quantile. Then, the analysis using NP-QR further reveals that an increase in SI during the economic upturn (SI ≤ 0) is likely to affect the RR more than during the economic downturn (SI > 0). However, the effect of the downturn economy is prominent for the RR of high-risk characteristics loans at the upper 0.75 quantile and that of medium-risk characteristics loans at the 0.25 and 0.5 quantiles, especially those loans with DC > 0.5.

Furthermore, based on four criteria to compare the performances of the models studied in this chapter, we find that the proposed NP-QR performs the best in terms of goodness of fit, in-sample prediction, and VaR. On the other hand, the proposed PLA-QR outperforms most alternative models in goodness of fit, in- and out-of-sample prediction, in-sample distributional fit, and VaR at the 0.20 and 0.25 quantiles. Moreover, as the boundary problem in PLA-QR is vastly mitigated but not completely resolved, we apply the model with a back-transformation technique (PLA-QR(tr)) to eliminate the problem. However, the models with data transformation are the least preferable according to our model selection criteria.

Despite the proposed QR models performing well at the various quantiles of conditional RR distribution and demonstrating the heterogeneity in the RR-covariate relationship, the problems associated with boundary are not resolved in the PLA-QR. Also, an analysis of NP-QR reveals an evidence of interaction effects of some covariates on RR, and by taking such effects into account might lead to outstanding performances of the model proposed in this chapter. In the next chapter, we will propose a model which by construction will not have boundary problems and will be able to address the interaction effects of the covariates on defaulted loan recoveries.

4.5 Appendix C: Tables and Figures

Variables	NP-QR Bandwidth
Recovery rate	0.0133
Debt cushion	0.2327
Stress index	0.5021
Type of loan	0.6819
Instrumental rank	0.1180
Collateral status	0.2440

Table 4.1: *Selected bandwidths of the nonparametric quantile regression*

Note: These bandwidths of NP-QR are selected based on least square cross validation method in (4.2.8). Then, they are applied to estimate the conditional distribution in (4.2.7), and the quantile function in (4.2.10)

	(A) Partially linear additive quantile regression			(B) Linear quantile regression		
	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
Additive component						
Degree of B-spline for DC	9	5	6	1	1	1
SIC	5.5275	6.0069	5.7700	5.6185	6.0384	6.0148
Linear component						
Debt cushion	NA	NA	NA	0.6308 ***	0.4444 ***	0.0549 *
				(0.0143)	(0.0203)	(1.9276)
Stress Index	-0.0378 **	-0.0266 ***	-0.0049	-0.0168	-0.0406	-0.0072
	(0.0164)	(0.0027)	(0.0303)	(0.0033)	(0.0046)	(0.0097)
Revolving loan	0.0450 ***	0.0170 ***	0.0005	0.1163 ***	0.0084	0.0015
	(0.0131)	(0.0060)	(0.0014)	(0.0238)	(0.0134)	(0.0094)
Senior Secured Bond	-0.0098	0.0102	-0.0006	0.0480 **	-0.0683 ***	-0.0118
	(0.0128)	(0.0120)	(0.0043)	(0.0238)	(0.0184)	(0.0150)
Senior Subordinated Bond	-0.0079	-0.2283 ***	-0.3698 ***	-0.0052	-0.2147 ***	-0.5194 ***
	(0.0129)	(0.0377)	(0.0601)	(0.0273)	(0.0401)	(0.0515)
Senior Unsecured Bond	0.0086	-0.0134	-0.0667 **	0.0133	-0.0076	-0.2045 ***
	(0.0161)	(0.0338)	(0.0313)	(0.0297)	(0.0384)	(0.0506)
Subordinated Bond	-0.0088	-0.2420 ***	-0.3849 ***	-0.0072	-0.2409 ***	-0.5463 ***
	(0.0129)	(0.0398)	(0.0719)	(0.0270)	(0.0386)	(0.0724)
Rank 2	-0.1242 ***	-0.1526 ***	-0.0186	-0.1230 ***	-0.1657 ***	-0.0244
	(0.0108)	(0.0178)	(0.0245)	(0.0088)	(0.0188)	(0.0231)
Rank 3	-0.1373 ***	-0.2063 ***	-0.1098 ***	-0.1358 ***	-0.2154 ***	-0.1387 ***
	(0.0111)	(0.0332)	(0.0388)	(0.0092)	(0.0312)	(0.0369)
Rank 4	-0.1352 ***	-0.2443 ***	-0.1622 ***	-0.1302 ***	-0.2635 ***	-0.1995 ***
	(0.0111)	(0.0217)	(0.0372)	(0.0097)	(0.0187)	(0.0582)
Collateral	0.1982 ***	0.0676 **	0.0009	0.1462 ***	0.1307 ***	-0.0078
	(0.0117)	(0.0304)	(0.0188)	(0.0387)	(0.0341)	(0.0452)

Table 4.2: *Linear component estimates of the partially linear additive model*

Note: Panel (A) reports the estimates of PLA-QR in which DC is included as an additive component. Panel (B) reports the estimates of the linear quantile regression as specified in (4.2.3). *** and ** represent significances at the 1% and 5%, respectively. The value in bracket is a standard error.

	Bandwidth		
Variables	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
$\tilde{y}_{dc,\tau}$	0.0001	0.0018	0.0172
$m_{DC,\tau}(DC)$	0.1073	0.0500	0.1878

Table 4.3: *Selected bandwidths of the partially linear additive model for the second-step estimation*

	$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
	PLA-QR	L-QR	PLA-QR	L-QR	PLA-QR	L-QR
The predicted RR exceeding upper boundary(1)	0.50%	1.17%	4.10%	6.88%	22.22%	15.08%
The predicted RR exceeding lower boundary(0)	0.97%	4.75%	0.62%	1.04%	0.25%	0.31%
Maximum estimate (upper bound)	1.0006	1.0537	1.0349	1.1458	1.0138	1.0331
Minimum estimate (lower bound)	-0.0170	-0.0802	0.0100	0.0104	0.2627	0.2091

Table 4.4: *The percentage of predicted RR exceeding zero and one boundaries*

	(A)Partially linear additive quantile regression			(B) Linear quantile regression		
	τ = 0.25	τ = 0.5	τ = 0.75	τ = 0.25	τ = 0.5	τ = 0.75
Additive component						
Degree of B-spline for DC	2	3	3	1	1	1
SIC	5.5706	5.9741	5.9468	5.5949	5.9907	5.9666
Linear component						
Debt cushion	NA	NA	NA	6.3014 ***	7.3820 ***	0.6351
				(0.4098)	(0.0286)	(0.5630)
Stress Index	-0.1720 ***	-2.3796 ***	-0.6231	-0.3317 ***	-0.5385 ***	-0.0778
	(0.0265)	(0.2415)	(0.3616)	(0.0380)	(0.0633)	(0.0854)
Revolving loan	0.5216 ***	0.3820 ***	0.0620 **	1.4683 ***	0.6442 ***	0.0123
	(0.1221)	(0.0802)	(0.0303)	(0.2644)	(0.2208)	(0.1619)
Senior Secured Bond	0.2741 ***	0.0927	-0.1894	1.2988 ***	-0.5153 **	-0.2172
	(0.0962)	(0.1169)	(0.0996)	(0.1398)	(0.2185)	(1.0177)
Senior Subordinated Bond	-0.9004 ***	-1.0205 ***	-2.5564 ***	-0.1445	-0.8272 **	-6.5882 ***
	(0.2551)	(0.2914)	(0.7351)	(0.4185)	(0.3659)	(0.5669)
Senior Unsecured Bond	1.1402 ***	0.4946 **	-1.2236	1.6824 ***	0.6103	-5.3511 ***
	(0.2016)	(0.2057)	(0.6747)	(0.3978)	(0.3418)	(0.4813)
Subordinated Bond	-1.5508 ***	-1.2565 ***	-2.5152 ***	-0.7021	-0.9649 **	-6.6625 ***
	(0.1854)	(0.3016)	(0.7132)	(0.4110)	(0.4099)	(0.4974)
Rank 2	-0.6984 ***	-0.7675 ***	-0.2586	-0.7501 ***	-0.7449 ***	-0.3088
	(0.0848)	(0.0872)	(0.2094)	(0.1423)	(0.1293)	(0.2264)
Rank 3	-2.5000 ***	-1.2440 ***	-0.7619 ***	-2.8155 ***	-1.3058 ***	-0.8095 ***
	(0.0985)	(0.2492)	(0.1967)	(0.1491)	(0.2139)	(0.2138)
Rank 4	-2.5289 ***	-1.8577 ***	-1.2304 ***	-2.7392 ***	-1.6205 ***	-1.2717 ***
	(0.2578)	(0.3309)	(0.4620)	(0.4209)	(0.3690)	(0.3826)
Collateral	1.9699 ***	0.7697 ***	-0.4339	1.5915 ****	1.0227 ***	-0.1070
	(0.1733)	(0.1935)	(0.2254)	(0.3946)	(0.2866)	(0.3247)
	(0.2577)	(0.1935)	(0.2004)	(0.1645)	(0.2866)	(0.4310)

Table 4.5: *Linear component estimates of the partially linear additive model with back transformation*

Note: The models in Panel (A) and (B) are estimated by regressing the set of the determinants on the transformation of the recovery rate. Hence, the parameter estimates reflect only the relationships between the determinants and the transformed recovery rate.

	Models without transformation			Models with transformation		
Models	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
(i) Average residuals using the loss function						
Fully nonparametric quantile regression	0.0844	0.1100	0.0846	-	-	-
Partially linear additive quantile regression	0.0865	0.1149	0.0910	0.0870	0.1122	0.0895
Linear quantile regression	0.0896	0.1190	0.0920	0.0917	0.1141	0.0913
(ii) Pseudo R^2						
Fully nonparametric quantile regression	0.3397	0.3943	0.2563	-	-	-
Partially linear additive quantile regression	0.3128	0.3509	0.1803	0.3219	0.3606	0.1910
Linear quantile regression	0.2903	0.3303	0.1698	0.2907	0.3614	0.1736

Table 4.6: *Goodness of fit of the quantile regressions*

Models	In-sample		Out-of-sample	
	$MSE^{(q)}$	$MAE^{(q)}$	$MSE^{(q)}$	$MAE^{(q)}$
NP-QR	0.0374	0.1124	0.0954	0.2075
PLA-QR	0.0386	0.1151	0.0404	0.1202
L-QR	0.0400	0.1228	0.0389	0.1205
PLA-QR(tr)	0.0452	0.1235	0.0512	0.1342
L-QR(tr)	0.0539	0.1363	0.0527	0.1358

Table 4.7: *Point prediction of the quantile regressions*

Note: Out-of-sample data is formed by randomly selected 30% of the full sample data. Each model is estimated by the remaining 70% to predict the out-of-sample RR.

Models	In-sample		Out-of-sample	
	HWMI	KS-statistic	HWMI	KS-statistic
NP-QR	0.0314	0.0877	0.0389	0.1003
PLA-QR	0.0296	0.0841	0.0324	0.0823
L-QR	0.0302	0.0904	0.0266	0.0720
PLA-QR(tr)	0.0348	0.0990	0.0530	0.1073
L-QR(tr)	0.0391	0.1097	0.0550	0.1118

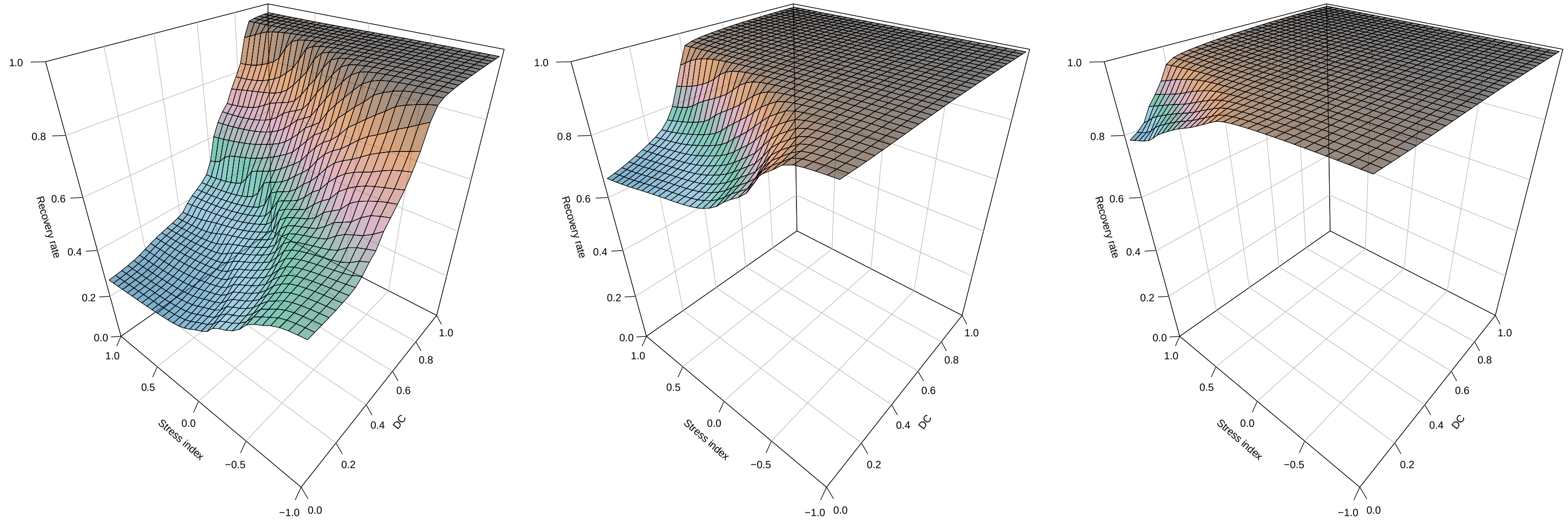
Table 4.8: *Distributional fit of the quantile regressions*

Note: HWMI is the Harmonic Weighted Mass index which is defined in (4.2.28). The KS-statistic is calculated by the maximum difference in the sample quantiles of the RR sample and predicted RR sample distributions defined in (4.2.29). The critical values at 1% and 5% levels of significance are 0.004 and 0.003.

Model	$\bar{\tau} = 0.05$	$\bar{\tau} = 0.1$	$\bar{\tau} = 0.15$	$\bar{\tau} = 0.2$	$\bar{\tau} = 0.25$
NP-QR	89.92%	87.04%	82.99%	79.39%	75.65%
(difference)	(5.08%)	(2.96%)	(2.01%)	(0.61%)	(0.65%)
PLA-QR	87.17%	83.44%	79.30%	75.34%	71.24%
(difference)	(7.83%)	(6.56%)	(5.70%)	(4.66%)	(3.76%)
L-QR	86.72%	82.99%	78.89%	75.02%	71.11%
(difference)	(8.28%)	(7.01%)	(6.11%)	(4.98%)	(3.89%)
PLA-QR(tr)	87.58%	82.68%	76.51%	72.41%	67.46%
(difference)	(7.42%)	(7.32%)	(8.49%)	(7.59%)	(7.54%)
L-QR(tr)	88.24%	85.24%	80.51%	74.39%	64.76%
(difference)	(6.76%)	(4.76%)	(4.49%)	(5.61%)	(10.24%)
Expected hit rate	95.00%	90.00%	85.00%	80.00%	75.00%

Table 4.9: *Hit rates of the predicted Value at Risks using quantile regressions*

Note: The hit rate is defined as a percentage of underestimated predicted VaR for each model. The percentage in bracket is a difference between the expected hit rate and the hit rate from predicted VaR of each model. The model that provides the lowest difference is the most preferable.



(a) $\tau = 0.25$

(b) $\tau = 0.5$

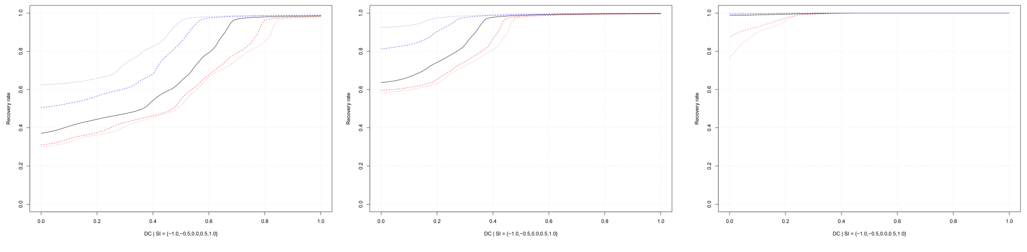
(c) $\tau = 0.75$

Figure 4.1: *Effects of debt cushion and stress index on the RR of the low risk characteristic loan using NP-QR*

Note: The low risk characteristics are defined as revolving loan (type = 2), with instrumental rank 1 (Rank = 1), and with collateral (Col = 1).

The figure show how RR of the given loan responds to the change in the stress index and debt cushion at three conditional quantiles

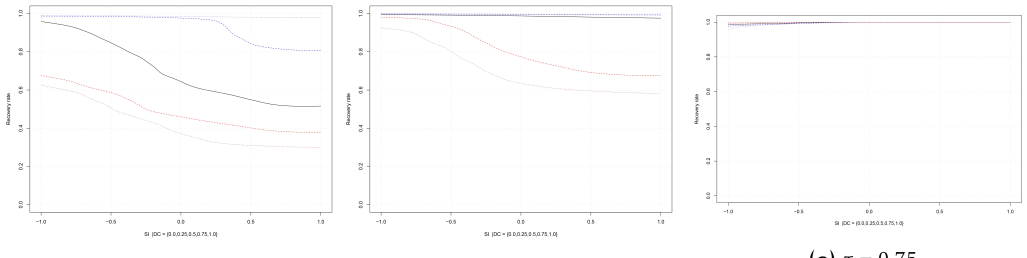
: $\tau = 0.25, 0.5 and 0.75$



(a) $\tau = 0.25$ (b) $\tau = 0.5$ (c) $\tau = 0.75$

Figure 4.2: *Conditional effects of the debt cushion on RR of the low risk characteristic at various economic conditions*

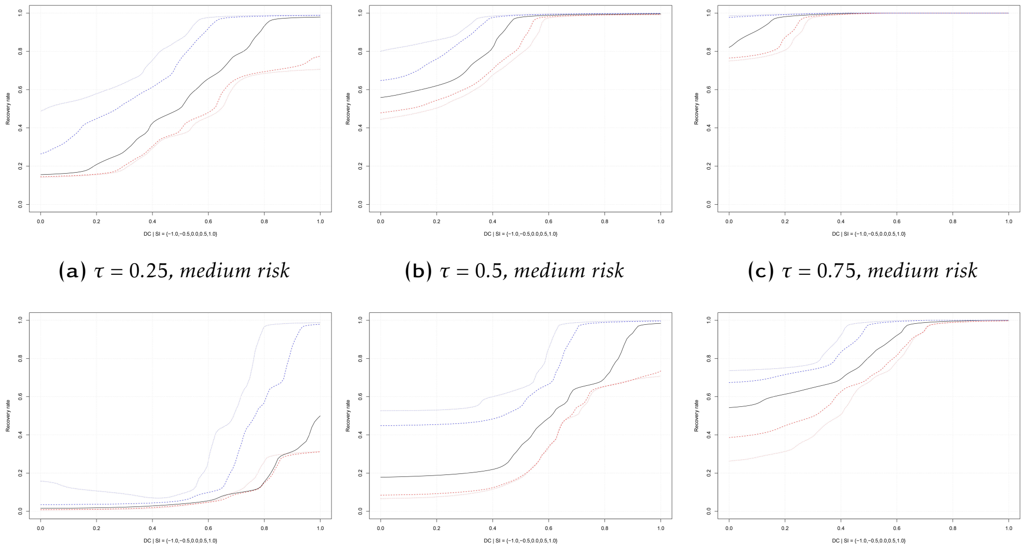
Note: Sub-figures (a)-(c) illustrate the NP-QR estimates of the conditional effects of debt cushion on RR of low risk characteristics loan for 0.25, 0.5, and 0.75 quantiles, respectively. The effects are conditional on five economic scenarios as SI = $\{-1, -0.5, 0, 0.5, 1\}$. In each sub-figure, the blue dotted line and the blue dashed line are the conditional effects of DC on RR during the negative SI (economic upturn) at -1 and -0.5, respectively. On the other hand, the red dotted and dashed lines are the conditional effects of DC during the positive SI (economic downturn) at 0.5 and 1, respectively. The dark solid line is the conditional effect of DC given SI = 0 (neutral condition).



(a) $\tau = 0.25$ (b) $\tau = 0.5$ (c) $\tau = 0.75$

Figure 4.3: *Conditional effects of the stress index on RR of the low risk characteristic at various levels of debt cushion*

Note: Sub-figures (a)-(c) illustrate the NP-QR estimates of the conditional effects of stress index on RR of low risk characteristics loan for 0.25, 0.5, and 0.75 quantiles, respectively. The effects are conditional on five levels of DC as $DC = \{0, 0.25, 0.5, 0.75, 1\}$. In each figure, the blue dotted line and the blue dashed line are the conditional effects of SI on RR with DC at 0 and 0.25, respectively. On the other hand, the red dotted and dashed lines are the conditional effects of SI given DC at 0.75 and 1, respectively. The dark solid line is the conditional effect of SI given DC = 0.5.



(a) $\tau = 0.25$, *medium risk*

(b) $\tau = 0.5$, *medium risk*

(c) $\tau = 0.75$, *medium risk*

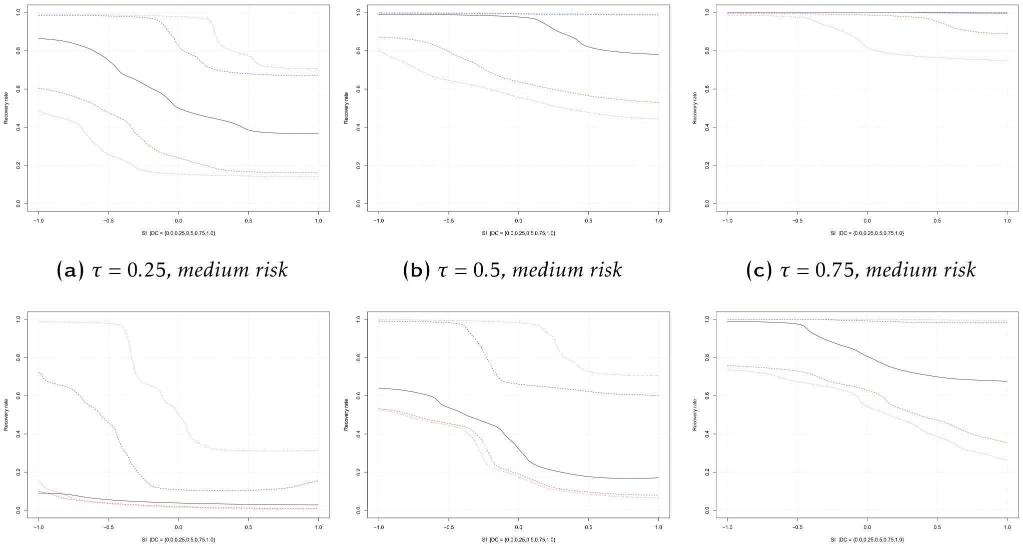
(d) $\tau = 0.25$, *high risk*

(e) $\tau = 0.5$, *high risk*

(f) $\tau = 0.75$, *high risk*

Figure 4.4: *Conditional effects of debt cushion on the recovery rate for the loans with medium and high risk characteristics*

Note: The figures illustrate the NP-QR estimates of the conditional effects of debt cushion on RR of medium- and high-risk characteristics loan. The effects are conditional on five economic scenarios as SI = $\{-1, -0.5, 0, 0.5, 1\}$. Sub-figures (a)-(c) represent the effect on the medium-risk characteristic loan: Type = 2, Rank = 1, and Col = 0, at 0.25, 0.5, 0.75 quantiles, respectively. Sub-figures (d)-(e) represent the effect on RR of high risk characteristic: Type = 5, Rank = 4, and Col = 0, at 0.25, 0.5, 0.75 quantiles, respectively. In each sub-figure, see the descriptions in Figure 4.2.



(a) $\tau = 0.25$, *medium risk*

(b) $\tau = 0.5$, *medium risk*

(c) $\tau = 0.75$, *medium risk*

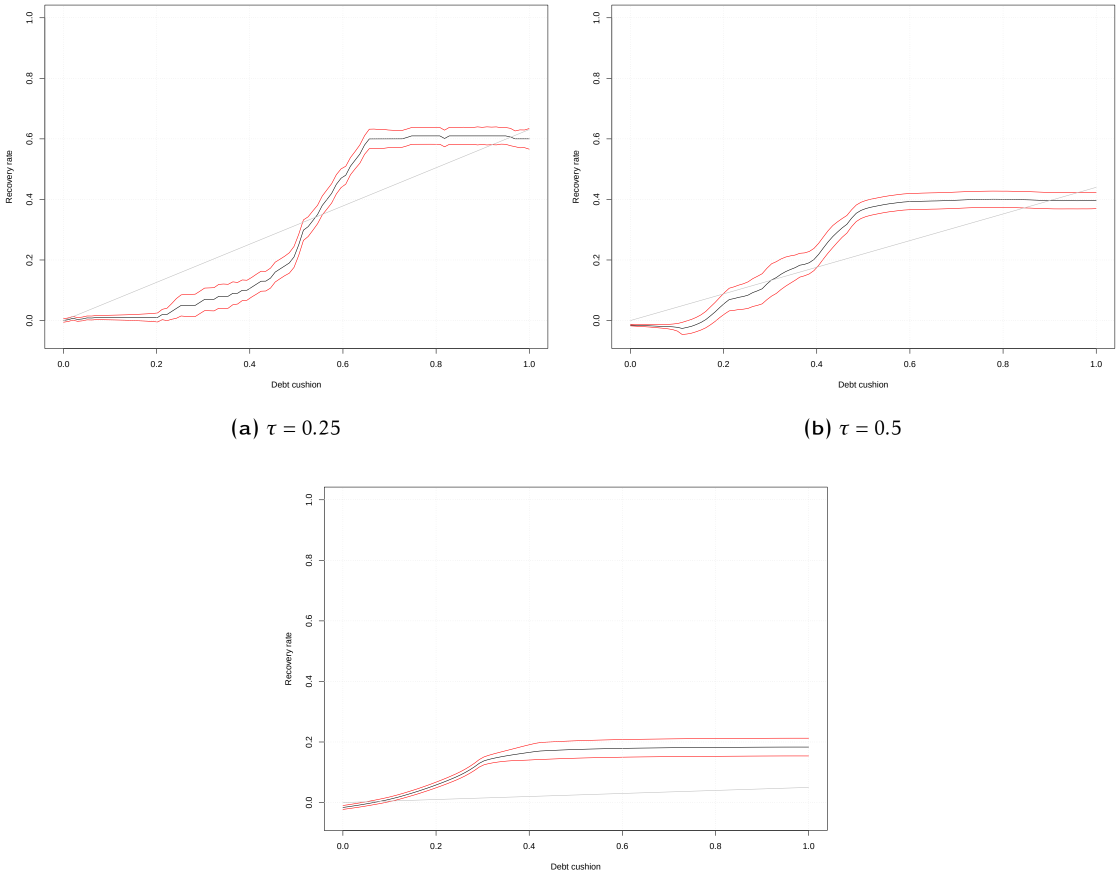
(d) $\tau = 0.25$, *high risk*

(e) $\tau = 0.5$, *high risk*

(f) $\tau = 0.75$, *high risk*

Figure 4.5: *Conditional effects of the stress index on the recovery rate of the revolving loan with medium and high risky characteristics*

Note: The figures illustrate the NP-QR estimates of the conditional effects of stress index on RR of medium- and high-risk characteristics loan. The effects are conditional on five levels of debt cushion as $DC = \{0, 0.25, 0.5, 0.75, 1\}$. Sub-figures (a)-(c) represent the effects on the medium-risk characteristic loan: Type = 2, Rank = 1, and Col = 0, at 0.25, 0.5, 0.75 quantiles, respectively. Sub-figures (d)-(e) represent the effects on RR of high risk characteristic: Type = 5, Rank = 4, and Col = 0, at 0.25, 0.5, 0.75 quantiles, respectively. In each sub-figure, see the descriptions in Figure 4.3



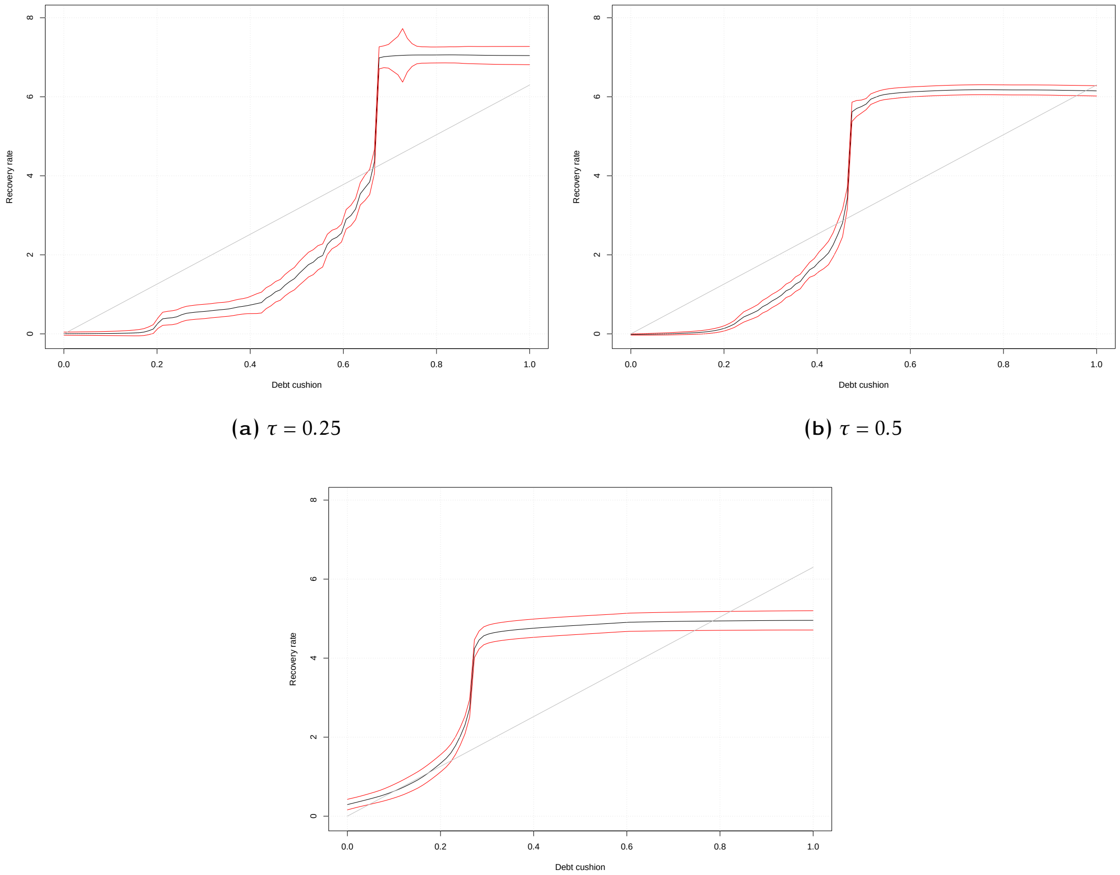
(a) $\tau = 0.25$

(b) $\tau = 0.5$

(c) $\tau = 0.75$

Figure 4.6: *Nonlinear effect estimates of debt cushion using the partially linear quantile regression*

Note: As we assume DC in the additive component $m(DC)$ of PLA-QR, the dark line illustrates additive estimates representing the effect of DC. The red lines are the bootstrapping confident interval at 5% level of significant. The grey solid line represents the marginal effect of DC using L-QR.



(a) $\tau = 0.25$

(b) $\tau = 0.5$

(c) $\tau = 0.75$

Figure 4.7: *Nonlinear effect estimates of debt cushion using the partially linear quantile regression with logit transformation*

Note: The figures show the additive component estimates of DC using PLA-QR(tr). Note importantly, the figures show the effects of DC on the transformed RR, as we employ the logit transformation technique. In each figure, see the detailed descriptions in Figure 4.6

Chapter 5

Local logit regression for recovery rate modelling

5.1 Introduction

The topic investigated in this chapter is motivated by the findings in chapters 3 and 4. Although the partially linear conditional mean regression (PL) in chapter 3 and the partially linear additive quantile regression (PLA-QR) in chapter 4, both extensively studied in the respective chapters, offer relatively high out-of-sample predictive accuracies, the boundary problem has not been completely resolved. Despite the fact that the PL and PLA-QR models greatly mitigate the [0,1] boundary issue of the RR by allowing nonlinearity in the relationship between the RR and its covariates, we find that a very small percentage (say, 0.5%) of the out-of-sample RR predictions exceed the boundaries zero and one. On the other hand, the nonparametric regressions with local constant method were able to ensure the boundary condition, but the out-of-sample predictive power was rather low due to the problem of high dimensional covariates.

Furthermore, the popular and simple *transformation* and *back-transformation* technique introduces bias to the conditional mean model estimates, as previously discussed in chapters 2 and 3. The studies that have used back-transformation appear to have overlooked the presence of bias in the model estimates. An exception is the QMLE[1] regression developed by Papke and Wooldridge (1996) specifically for fractional data (QMLE-RFRV). The aforementioned bias does not arise in this model by construction. Several studies have applied linear QMLE-RFRV regression to RR modelling and have found that this model provides better RR prediction than the other regression models (Dermine & De Carvalho, 2006; Khieu et al., 2012; Qi & Yang, 2009). In chapter 3, we show that the QMLE-RFRV has better out-of-sample predictive accuracy than the alternative parametric regressions that have been popular in the RR modelling literature.

In this chapter, we build on insights from the findings of large-scale empirical research in the literature as well as our studies in chapters 3 and 4 documenting the merits of non-parametric and semiparametric approaches and regression for fractional data for the purpose of recovery predictions, marginal effect and interaction effect analysis. The primary aim of this chapter is to propose a flexible and robust nonparametric local logit model for the RRs of defaulted loans. The proposed model specification facilitates marginal and interaction effects, as well as generating RR predictions that would lie within[0,1]. This chapter makes several principal contributions as follows.

First, our proposed local logit model has a flexible model specification, in that the unknown coefficients are assumed to be functions of all covariates and can be locally estimated. The data-driven kernel estimation method uncovers the underlying nonlinear recovery covariate relationships. This facilitates the analysis of the marginal and interaction effects of the conditioning variables on RR, as will be demonstrated in our empirical application presented in this chapter.

[1] Quasi maximum likelihood estimation

Second, the local logit model estimates are robust to the various shapes and features of recovery distribution discussed in the previous paragraphs, thus providing reliable statistical inference. Third, our model is developed specifically for fractional data. In proposing the local logit model for fractional data, we integrate the studies of Papke and Wooldridge (1996) who preoposed QMLE-RFRV, and Frölich (2006) who developed the local logit model for binary discrete variables, and demonstrated its superiority to its parametric counterparts. Thus, there is no need for trimming and transforming recoveries for regression modelling. As a result, the aforementioned bias will not arise in the local logit model estimates, improving further the reliability of statistical inference and recovery prediction.

Fourth, we apply the local logit regression to the widely studied Moody's RR dataset, which spans 18 years. We demonstrate the ways in which loan/borrower characteristics and the economic condition at the time of default and their interactions influence the recoveries of defaulted loans and their predictions. We provide a comprehensive analysis of nonlinear marginal and interaction effects on recoveries, whereas the main focus of previous studies has largely been on the prediction of recoveries and linear marginal effects. Our model does not only capture the nonlinearity in the marginal effects of DC and SI, it also accommodates nonlinear interactions between continuous and discrete variables, and their effects on RR.

In addition, we illustrate the framework that integrates the applications of the nonparametric regression to improve the QMLE-RFRV model specification. Specifically, we improve the parametric functional form by means of a "calibration" method using the information gained from the comprehensive empirical analysis of the local logit estimates. This aims to mitigate the misspecification problem in the parametric regression.

These contributions highlight the novelty of our proposed local logit model, in particular its flexibility in accommodating nonlinear recovery covariate relationships and thus enriching the model specification, which supports improved recovery prediction.

The remainder of this chapter is organised as follows: in the next section, we propose the nonparametric local logit regression for [0,1] bounded response data along with the estimation method. This is followed by a brief discussion of the parametric QMLE-regression for fractional data and the estimation method. Section 5.3 conducts a simulation study to assess various properties and the robustness of the proposed model and analyses the results. Section 5.4 provides a specification test. Section 5.5 conducts the empirical analysis and assesses the out-of-sample recovery predictability of the models. Section 5.6 concludes this paper. Some additional results of the simulation study in Section 5.3 are further discussed in Appendix D.

5.2 Methodology

In this section, we discuss the parametric QMLE regression for fractional response variable (QMLE-RFRV) and propose a nonparametric local logit model and the estimation methods which include the choice of kernel functions and bandwidth selection criterion. Furthermore, we briefly discuss several criteria in order to evaluate the predictive performance of the proposed model relative to the parametric counterpart.

5.2.1 Parametric regression for [0,1] bounded data

The parametric QMLE-RFRV is the theoretically valid model for the fractional response variable, such as the recovery rate (RR). The conditional mean is given as:

$$E(Y|X=x) = \Lambda(x'\gamma) \qquad , \tag{5.2.1}$$

where Y is the continuous [0,1] bounded variable (i.e. $0 \leq Y \leq 1$), X is the vector of k covariates (which is individual loan characteristics - a mixture of continuous and discrete variables in the empirical example), $\Lambda(\cdot)$ is the logistic function, $0 < \Lambda(\cdot) < 1$, and γ is a vector of unknown parameters. Papke and Wooldridge (1996) proposed a quasi-maximum likelihood estimation (QMLE) method. The unknown vector of parameters are estimated as:

$$\hat{\gamma} = \arg\max_{\gamma} \sum_{i=1}^{n} Y_i \log(\Lambda(X_i'\gamma)) + (1-Y_i)\log(1-\Lambda(X_i'\gamma)). \tag{5.2.2}$$

The estimator in (5.2.2) is consistent and asymptotically normal, these properties being robust to various conditional distributional assumptions.

The main assumption of the QMLE-RFRV is the correctly specified functional form for the conditional mean. However, the conditional mean of this model can be misspecified in practice because the underlying correct functional form is largely unknown. We want to improve the specification of the conditional mean of QMLE-RFRV, which might include sufficient number of interaction terms, polynomials and discretized continuous variables and so on, by exploiting information provided by the estimates of local logit model. The calibrated QMLE-RFRV is presented in Section 5.5.3.

5.2.2 Local logit regression

This study proposes a local logit regression for fractional response variable and a data driven nonparametric method to estimate the model. As will be seen, the local logit model is flexible to accommodate the underlying any complex nonlinear relationship between RR and covariates. The conditional mean is defined as:

$$E(Y|X = x) = \Lambda(x'\beta(x)) \tag{5.2.3}$$

where $x = (x_1,..,x_k)'$ is $k \times 1$ vector, $\beta(x)$ is a vector of unknown local logit estimator is the function of x.

We obtain the estimators of local logit model by maximizing the local likelihood function (Tibshirani & Hastie, 1987) as:

$$\hat{\beta}(x) = \arg\max_{\beta(x)} \sum_{i=1}^{n} Y_i \log(\Lambda(X_i'\beta(x))) + (1 - Y_i)\log(1 - \Lambda(X_i'\beta(x)))\mathbb{K}_H(X_i, x), \tag{5.2.4}$$

where $\mathbb{K}_H(X_i, x)$ is a product of k kernel functions associated with $(x_1,...,x_k)$ for a given a vector of bandwidths $H = (h_1,...,h_k)'$. The local logit model parameter $\beta(x)$ - which is a function of covariates x - is locally estimated based on a kernel weights $\mathbb{K}_H(X_i, x)$, which determine the local distance between X_i and a specified value of vector x for a given set of bandwidths H.

Our study employs two different kernel functions: a Gaussian kernel function for continuous variables and a kernel function which is constructed specifically for categorical variables. Let us define:

$$X_i = (X_i^c, X_i^d),$$

where the continuous regressors with p dimensions is $X_i^c \in \mathbf{R}^p$, the remaining regressors X_i^d is a $q \times 1$ vector of categorical variables, and $p + q = k$. For any t^{th} component in X_i^d, where $t \in \{1,...,q\}$, each component can take a discrete value such as $X_{t,i}^d \in \{0, 1, ..., c_t - 1\}$, where $c_t \geq 2$ is the number of categories of $X_{t,i}^d$. Clearly, $c_t = 2$ for the dummy variable. In what follows, the two kernel functions that we use in the estimation of local logit are defined, which also have been fully discussed in Chapter 3.

A kernel function for continuous variable

The standard Gaussian kernel function is employed for any continuous variable (X_i^c) which is defined as:

$$\kappa_s\left(X_{s,i}^c, x_s^c, h_s\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{X_{s,i}^c - x_s^c}{h_s}\right)^2\right), \tag{5.2.5}$$

where $s = 1,..,p$, $\kappa(\cdot)$ is the Gaussian kernel function, and h_s is a bandwidth associated with s^{th} continuous variable.

A kernel function for discrete variable

For the discrete variable, we apply the kernel function proposed by Racine and Li (2004) which is defined as:

$$\lambda_t(X_{t,i}^d, x_t^d, l_t) = \begin{cases} 1, & \text{if } X_{t,i}^d = x_t^d, \\ l_t, & \text{otherwise,} \end{cases} \tag{5.2.6}$$

where we assume that the t^{th} categorical variable, l_t is the bandwidth associated with $\lambda_t(\cdot)$, and $0 < l_t \leq 1$.

The product of the kernel functions[2] in (5.2.5) and (5.2.6) are functions of bandwidths and the optimum selection of which are crucial in the estimation of the local logit model. There are several bandwidth selection methods available for the non-parametric estimation. Although the plug-in method is popular, its application is limited as it does not work well in the small sample setting and the high dimensional independent variables x. This study use, on the other hand, the least-squares cross-validation which is commonly applied in practice to select the bandwidth that minimize a certain loss function.

[2] which is defined as $\mathbb{K}(X_i, x) = \kappa_1(X_{i,1}^c, x_1^c, h_1) \cdots \kappa_p(X_{i,p}^c, x_p^c, h_p) \cdot \lambda_1(X_{i,1}^d, x_1^d, l_1) \cdots \lambda_q(X_{i,q}^d, x_q^d, l_q)$

In this study, we select the set of bandwidths $H = (h_1,..,h_p,l_1,..,l_q)$ that minimizes an objective function, which is the sum of prediction error squares, defined as:

$$CV = \sum_{i=1}^{n} \Big(Y_i - \Lambda(X_i'\hat{\beta}(x_{-i}|H)) \Big)^2, \tag{5.2.7}$$

where $\hat{\beta}(x_{-i}|H)$ is a $k \times 1$ vector of leave-one-out estimates of local logit estimators associated with x_i which is a solution to:

$$\arg\max_{\beta(X_i)} \sum_{j=1, i\neq j}^{n} Y_j \log(\Lambda(X_j'\beta(X_i))) + (1-Y_j)\log(1-\Lambda(X_j'\beta(X_i)))\mathbb{K}_H(X_j, X_i). \tag{5.2.8}$$

It is worth noting that the optimal bandwidth in (5.2.7) would be very large if the unknown underlying functional form is indeed the standard linear function, $\Lambda(X'\gamma)$. When the sizes of all bandwidths increase as n goes to infinity, the approximately equal kernel weights are assigned for all i. Specifically, the large bandwidths would cause the product of kernel functions in (5.2.4) to be the same regardless of the local distance between X_i and x. Thus, the local estimators $\beta(x)$ in (5.2.4) converge to the global estimator γ in (5.2.2) as the bandwidths become larger. It shows that the local logit model encompasses the global parametric QMLE-RFRV.

Moreover, when the dimension p of continuous variables is large, in general, the nonparametric estimation method has curse of dimensionality problem (Greene, 2003). This is a common criticism in nonparametric method due to the sparsity of data in hight dimensional space which leads to a decrease in convergence rate for regression function estimators as number of regressors increase. As mentioned in Frölich (2006), the variance of the error term is bounded in the model with fractional response variable, and therefore, the curse of dimensionality problem does not arise in the local logit model we study in this paper. In addition, in the estimation of local logit model with binary response variable, Frölich (2006) assigns one common bandwidth for all discrete and another common bandwidth for all continuous regressors. In our study, we apply different bandwidths for the continuous variables and only one bandwidth for all discrete variables.

5.3 Simulation study

In this section, we conduct an extensive simulation study in order to assess the finite sample properties of the proposed local logit model estimators and their robustness to various nonlinear functional forms for the conditional mean and to various symmetric and asymmetric error distributions. For the comparison purpose, we consider the QMLE-RFRV as the benchmark model with correct linear and nonlinear functional forms. On the other hand, there is no assumption on the conditional mean specification of local logit model. Additionally, the shape of the error distribution is assumed to be unknown for both models. We also ensure that the response variable generated is bounded in [0,1] with high intensity at the boundaries zero and one, which reflect the typical features of the RR data to be modelled in the empirical application - one of the main objectives of this paper. We generate the data for two sample sizes, n=200 and n=500.

5.3.1 Experimental design

We generate seven sets of univariate and multivariate X variables with different degrees of nonlinearity in the conditional mean specifications. Furthermore, the data-generation processes include various distributional assumptions.

A1 Univariate data-generation process

We generate the data as follows: $X_1 \sim N(1,1)$, $U \sim N(0,1)$, and the response variable with two-sided censoring as $Y = \max(0, \min(1, Y^*))$. $Y^* = f(X_1)$ is the conditional mean specification with three different functional forms:

(U1) $Y^* = 0.5X_1 + U$

(U2) $Y^* = X_1^2 + U$

(U3) $Y^* = \sin(X_1) + U$

In other words, the univariate functional forms include linear, quadratic, and sine functions. Figures 5.1a to 5.1c demonstrate the densities of the simulated response variable under U1 to U3, respectively, which indicates the boundaries of [0,1] with intensities at both ends. Although, we assume two-sided censoring, this information is assumed to be unavailable. Therefore, both parametric fractional regression and local logit regression incorrectly specify the likelihood assumption and the link function.

[————— Insert [Figure 5.1] here —————]

A2 Bivariate data-generation process Bivariate data (X_1, X_2) is generated. Then, similar to A1, $Y = \max(0, \min(1, Y^*))$, where $Y^* = f(X_1, X_2)$. For a given data-generation process U where $U \sim N(0,1)$, (X_1, X_2) and Y^* are generated as follows:

(B1) $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, and $Y^* = 0.2X_1 + 0.5X_2 + U$

(B2) $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, and $Y^* = 0.2X_1 + 0.5X_2^2 + U$

(B3) $X_1 \sim \chi^2_{(3)}$, $X_2 \sim N(0,1)$, and $Y^* = 0.5\sin(X_1) + 0.5X_2 + 0.2X_2^2 + U$.

These assumptions assume the linear functional form and its combination with quadratic function in B1 and B2, respectively. On the other hand, a highly nonlinear functional form, which is a mixture of sine and quadratic functions, is applied in B3.

[————— Insert [Figure 5.2] here —————]

Figures 5.2a, 5.2b, and 5.2c show the densities of the generated Y for B1, B2, and B3, respectively, which are bounded with different proportions of the clustering at zero and one.

A3 Multivariate data-generation process

We generate a multiple data set which is a mixture of continuous and discrete independent variables. This is common in many practical applications arising in economics, finance, and other disciplines.

$$\text{(M1)} \quad Y = \Phi(-0.02X_1 + \sin(X_2) + D_1 + 0.5D_2 + D_3 + 0.5X_1D_2 + U),$$

where $\Phi(\cdot)$ is a probit link function, $X_1 \sim \chi^2_{(3)}$, $X_2 \sim N(1,1)$, $D_1 \sim Ber(1,0.75)$, $D_2 \sim Ber(1,0.4)$, $D_3 \sim Ber(1,0.2)$, and U is generated from an equally weighted mixture of $N(-2,1)$ and $N(2,1)$. Given the complexity of the functional form[3], we consider only n = 500. The sample density is presented in Figure 5.3.

[————— Insert [Figure 5.3] here —————]

5.3.2 Results of the simulation study: A summary

We assess the finite sample properties of the proposed local logit model in comparison with those of the parametric QMLE-RFRV, which is the benchmark model in terms of in-sample and out-of-sample predictabilities and the interpretability of the model estimates. We do this in the following four steps: (i) partition the full sample into in-sample and out-of-sample data; (ii) evaluate the predictability of the models using MSE and MAE criteria; (iii) repeat steps (i) and (ii) 1,000 times, then compute the average MSE and MAE; and (iv) compare the local logit model estimators with those of the benchmark model with correct model specifications. We assess these properties

[3] As there are three dummy variables, we consider only the moderate sample size to avoid the possibility of causing discontinuity in the conditional mean

for the three data-generation processes given in A1 to A3, and n = 200 and n = 500.

Predictive performance

Tables 5.1 and 5.2 report the in- and out-of-sample predictive measures MSE and MAE of the local logit and the benchmark model for n = 200 and 500, respectively. The results show that the proposed model consistently outperforms the benchmark model in in-sample prediction, while the out-of-sample performance of the local logit model is comparable to that of the benchmark model with correctly specified functional form. Full detailed results and discussion of the simulation study in terms of the predictive performance is provided in Appendix D1.

A noteworthy result is that the performances of the proposed model in U1 and B1 are identical to the benchmark model. We observe that the selected bandwidths are large when the true conditional mean is linear. This might indicate that the local logit estimates identify the model specification correctly as discussed in section 5.2.

[————— Insert [Tables 5.1 and 5.2] here —————]

Local logit analysis

In this section, we examine how close the local logit estimates are to those of the benchmark model with correct functional form when the data-generation process is multivariate (M1) a mixture of continuous and discrete variables. The remaining results of assumptions A1 and A2 are also provided in Appendix D2.

Let us denote the estimate of the benchmark model as:

$$\hat{y} = \Lambda(\hat{\gamma} + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 \sin(x_2) + \hat{\gamma}_3 d_1 + \hat{\gamma}_4 d_2 + \hat{\gamma}_5 d_3 + \hat{\gamma}_6 x_1 d_2). \qquad (5.3.1)$$

and the estimate of the local logit regression as:

$$\hat{y} = \Lambda(\hat{\beta}_0(x) + \hat{\beta}_1(x)x_1 + \hat{\beta}_2(x)x_2 + \hat{\beta}_3(x)d_1 + \hat{\beta}_4(x)d_2 + \hat{\beta}_5(x)d_3) \tag{5.3.2}$$

First, we analyse the interaction of x_1 and d_2. That is, the effect of x_1 on y depends on d_2. We expect the local estimate $\hat{\beta}_1(x)$ conditional on d_2 to be the same as the benchmark model estimate $\hat{\gamma}_1 + \hat{\gamma}_6$ when $d_2 = 1$, and $\hat{\gamma}_1$ when $d_2 = 0$. A plot of the local estimate $\hat{\beta}_1(x)$ given $d_2 = 0$ appears in Figure 5.4, which is clearly comparable to $\hat{\gamma}_1$ in Figure 5.4c. The results suggest that both models generate similar conditional marginal effect estimates of x_1. On the other hand, the local estimate $\hat{\beta}_1(x)$ given $d_2 = 1$ is shown in Figure 5.4b, which we compare with $\hat{\gamma}_1 + \hat{\gamma}_6$ in Figure 5.4d. We find that the average estimates over the iterations in both models are approximately 0.5. These findings indicate that on average the local logit estimate adequately captures the interaction effect between x_1 and d_2, although its variation is higher than that of the benchmark model estimate.

[———— Insert [Figure 5.4] here ————]

Second, consider the nonlinear component $\sin(x_2)$. The local marginal effect estimate $\hat{\beta}_2(x)$ in Figure 5.5a is compared with $\hat{\gamma}_2 \cos(x_2)$ in Figure 5.5b. These figures show that the local estimate approximates some of the nonlinear behavior of the benchmark model. The local logit estimate shows a positive effect with a diminishing rate when $x_2 > 0$. The effect is then negative, which is similar to the estimate of the correctly specified benchmark model.

Third, the marginal effect estimates of the discrete variables d_1 and d_3 are plotted in Figures 5.6a and 5.6b for the local logit and the parametric QMLE-RFRV, respectively. It is approximately 0.6 for both the local logit and benchmark models. However, the local estimates have slightly higher variations than those of the QMLE-RFRV model.

[———— Insert [Figures 5.5, 5.6] here ————]

Fourth, Figure 5.7 shows the estimate of the interaction of d_2 and x_1. Given the correct specification of the benchmark model in (5.3.1), the marginal effect of d_2 is $\hat{\gamma}_4 + \hat{\gamma}_6 x_1$, which is shown in Figure 5.7b. Clearly, the effect of d_2 on the response variable is a linear function of x_1. The local logit estimate $\hat{\beta}_4(x)$ indicates a positive relationship between d_2 and x_1, as shown in Figure 5.7a, which is approximately linear.

[———— Insert [Figure 5.7] here ————]

The overall results of the simulation study show that the local logit estimators can detect the nonlinear relationship between the response variable and covariates, including various forms of nonlinearity and interactions between continuous and discrete variables. In the empirical study of RR modelling, we will exploit this information from the local logit estimation to "calibrate" the QMLE-RFRV model (see section 5.5 for details).

5.3.3 Robustness of the local logit model

In this section, we evaluate the robustness of the proposed model under various assumptions regarding the error distribution, including bimodality and asymmetry. We consider two model specifications, which include (M1), defined in section 5.3, and (M2) defined as:

$$(M2) \qquad Y = \Phi(-1.5\sqrt{X_1} + \sin(X_2) + D_1 + 0.5D_2 + 0.5X_3D_2 + 0.2X_3 + d_3 + U)$$

where $X_1 \sim \chi^2_{(3)}$, $X_2 \sim \chi^2_{(1)}$, $X_3 \sim N(0,2)$, $D_1 \sim Ber(1,0.75)$, $D_2 \sim Ber(1,0.4)$, and $D_3 \sim Ber(1,0.2)$. We consider two assumptions for the error distribution: an asymmetric $U^{(1)} \sim \chi^2_{(1)}$, and a bimodal $U^{(2)}$ which is generated as the equally weighted mixture of $N(-2,1)$ and $N(2,1)$.

This study estimates the MSE and MAE measures of the local logit model and the benchmark model relative to those of the correctly specified parametric QMLE-RFRV model for the purpose of performance assessment. Note that the QMLE-RFRV with a standard linear functional form is the benchmark model used here for comparison purposes. Specifically, if the relative MSE and MAE are equal to or less than one, then the model performance is the same or better than that of the correctly specified QMLE-RFRV. We set three sample sizes, $n = 200$, 500 and 1,000, where evaluations are made in both the in-sample and out-of-sample data.

The in-sample performance measures (the relative MSE and MAE) of the proposed local logit model are reported in Panel (a) of Tables 5.3 and 5.4, respectively. These relative measures are consistently lower than those of the parametric regression. The results also show that both relative MSE and MAE are mostly less than or equal to 1.00 for both asymmetric and bimodal error distributions. On the other hand, the QMLE-RFRV, as the benchmark model, performs poorly for asymmetric error distribution, with both the MSE and MAE being greater than 1.00 and close to 2.00 in many cases.

Panel (b) of Tables 5.3 and 5.4 reports the models' out-of-sample performance measures. The local logit model continues to outperform the parametric regression in most cases. Additionally, the local logit model tends to have substantially lower MSE and MAE for the Chi-squared error assumption compared with the bimodal error distribution. Moreover, we notice that the local logit model has relatively large MSE and MAE for bimodal distributions for a small sample size n = 200, while vast improvements are observed for the larger sample sizes.

[———— Insert [Tables 5.3 and 5.4] here ————]

5.4 Specification testing

In this section, we briefly discuss a specification test for the null hypothesis that the parametric QMLE-RFRV model with a given specification fits the RR data well against the alternative hypothesis that the local logit model fits the data well. The testing procedure employs the generalised maximum likelihood ratio (Fan, Zhang, & Zhang, 2001) and is augmented with a bootstrap method for calculating the p-value of the test statistic. The test statistic is defined as:

$$TS = \frac{RSS_0 - RSS_1}{RSS_1} \tag{5.4.1}$$

where RSS_0 is the residual sum square under the null hypothesis which is $\frac{\sum_{i=1}^{n}(Y_i - \Lambda(X_i'\hat{\gamma}))^2}{n}$, and RSS_1 is under the alternative which is $\frac{\sum_{i=1}^{n}(Y_i - \Lambda(X_i'\hat{\beta}(x)))^2}{n}$. The null hypothesis is rejected if the p-value of the TS is less than the nominal level. To compute the p-value, we apply the wild bootstrap procedure as follows:

1. Under the null hypothesis, generate $Y_i^* = \Lambda(X_i'\hat{\gamma} + e_i^*)$ for each $i = 1, \ldots, n$, where e_i^* is generated as follows:

   - Estimate the residual $\hat{e}_i = \Lambda^{-1}(Y_i^{(\nu)}) - X_i\hat{\gamma}$ where $Y_i^{(\nu)} = \frac{Y_i + \nu}{1+2\nu}$, and ν is a small arbitrary value[4].

   - Obtain $e_i^* = (\hat{e}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i) \cdot \eta_i$ where $\{\eta_i\}$ is a sequence of independent and identically distributed random variables drawn from N(0,1).

2. Use the dataset $\{(Y_i^*, X_i) : i = 1, \ldots, n\}$ to estimate the models under both null and alternative hypotheses. Then, the test statistic is calculated as $TS^* = \frac{RSS_0^* - RSS_1^*}{RSS_1^*}$.

[4] This allows $\Lambda^{-1}(Y_i^{(\nu)})$ defines for all $Y \in [0,1]$.

3. Repeat Steps 1 and 2 B times to draw the empirical distribution for TS^*. Then, the p-value is computed by $\frac{1}{B}\sum_{b=1}^{B} I(TS_b^* \geq TS)$, where $I(\cdot)$ is an indicator function and TS_b^* is calculated based on the b-th bootstrap sample.

5.5 Empirical results

The local logit model is applied to the RR dataset[5] to uncover the nature of the underlying unknown nonlinear RR covariate relationships, to conduct marginal and interaction effects analysis, and to generate RR predictions. The results of this empirical investigation will be utilised to "calibrate" the functional form of the parametric QMLE-RFRV in order to mitigate the misspecification problem of the linear model.

5.5.1 Bandwidth selection

We estimate the proposed model with the full dataset of $3,573$ defaulted loans. The local logit regression is specified as:

$$\begin{aligned} y = E(Y|X=x) &= \Lambda\Big(x'\beta(x)\Big) \\ &= \Lambda\Big(\beta_0(x) + \beta_{DC}(x)\cdot DC + \beta_{SI}(x)\cdot SI \\ &\quad + \sum_{d=2}^{6}\beta_{Type^{(d)}}(x)\cdot Type^{(d)} + \sum_{d=2}^{4}\beta_{Rank^{(d)}}(x)\cdot Rank^{(d)} \\ &\quad + \beta_{Col}(x)\cdot Col\Big) \end{aligned} \tag{5.5.1}$$

where $\beta(x) = (\beta_0(x), \beta_{DC}(x), ..., \beta_{Col}(x))'$ is a vector of the unknown parameters associated with x, and $Type^{(d)}$ and $Rank^{(d)}$ are dummy variables representing each category of $Type$ and $Rank$, respectively (see Table 3.1 in chapter 3 for more details). For the comparison purpose, we estimate the benchmark model, which is the QMLE-RFRV, denoted as $\Lambda(x'\gamma)$. The local logit parameters are estimated

[5] The data description and preliminary analysis are provided in Chapter 3

with the kernel function (5.2.5) for the continuous variables DC and SI, and kernel function (5.2.6) for the categorical variables Type, Rank, and Col. The bandwidth is selected by the leave-one-out least-squares cross-validation method (5.2.7). Let us define,

$$H = (h_1, h_2, \ell_1),$$

where H is the 3×1 vector of the bandwidths, h_1 and h_2 are associated with DC and SI, respectively, and ℓ_1 is a single bandwidth for all three categorical variables: Rank, Type and Col. $H = (0.1121, 1.2734, 1.0000)$ is the set of selected optimal bandwidths. We estimate the local logit model with the selected bandwidths for the full dataset. We find that the MSE of the local logit model is 0.076, compared 0.089 for the benchmark model, indicating the better fit of local logit model for the data than the benchmark parametric model.

5.5.2 Local logit analysis

The marginal effects of the continuous and discrete variables are analysed in the local logit and the parametric QMLE-RFRV models. In this section, we provide the marginal effects analysis of continuous and discrete variables, followed by a discussion on the interaction effect between continuous and discrete variables

Local logit estimates of continuous variables

[———— Insert [Figure 5.8] here ————]

Figure 5.8a shows that the local logit estimate of DC is a nonlinear function of itself, denoted as $\hat{\beta}_{DC}(x)$, while the QMLE-RFRV estimate $\hat{\gamma}_{DC}$ is represented by the solid horizontal line. The local logit model estimates clearly show the nonlinear marginal effect of DC, whereas it is constant in the parametric model in figure 5.8a.

In the local logit model, the effect of DC on RR increases with somewhat constant rate for $0 < DC < 0.6$, reaching the highest impact when DC = 0.6. This is followed by a decreasing effect on the RR of the defaulted loans for $0.6 \leq DC < 0.8$, which reaches its minimum effect at $DC = 0.8$. Then, the positive effect with an increasing rate is reappeared for DC > 0.8. These results imply that defaulted loans with $0 < DC < 0.6$ tend to be more responsive to an increase in additional DC than loans with higher DC. Figure 5.8a also shows that the estimated parametric coefficient $\hat{\gamma}_{DC}$ is 2.4, which is similar to the average local logit estimates $\hat{\beta}_{DC}(x)$.

To analyse the effect of SI on RR, $\hat{\beta}_{SI}(x)$ and $\hat{\gamma}_{SI}$ are plotted (solid line) in Figure 5.8b. To explain the marginal effect of SI on RR, we consider three ranges of SI: the low SI as SI<0, the high SI as $0 \leq SI < 1.5$, and the crisis SI[6] as SI ≥ 1.5, indicating good, poor, and (global financial) crisis economic conditions. The effect of SI on RR is negative and increasing with SI, and it then becomes positive and increasing for SI > 1.5. The variation of the effect of SI on RR is very high for low values of SI. This result indicates that RR is more sensitive to changes in economic condition during a low-SI period, compared to high-SI and the GFC crisis periods. The variation of the local logit estimate shows that although SI has a nonlinear negative effect on RR, the magnitude of the effects are different conditional on the characteristics of each loan. A relatively small variation of the local estimate observed for high SI implies that the effect of high SI is less dependent on the loan characteristics than for the low SI and the crisis SI. In practice, these findings indicate changes in the behaviour and expectations of both banks and borrowers during an economic downturn ($0 < SI \leq 1.5$). Lenders would have similarly adopted more conservative financial strategies in preparation for a negative scenario. This leads to the smaller negative effect of high SI on RR with lower variation.

[———— Insert [Figure 5.9] here ————]

[6] The data description and summary statistics in chapter 3, section 3.4.1, show that the stress index is greater than 1.5 is observed only during the recent Global Financial crisis

Furthermore, we consider SI = $\{-1, 0, 1\}$ and examine the effect of DC on RR for collateralised revolving loans with rank 1 ($Type = 2, Rank = 1, Col = 1$) across various economic conditions measured by SI. The plot in Figure 5.9 indicates that the RR is a nonlinear function of DC. The marginal effect of DC on RR is zero for DC < 0.3, and positive & increasing until DC = 0.6, and then nearly zero for DC > 0.6. On the other hand, SI has a negative impact on RR. For example, if we consider a loan with DC = 0, then the RR is 0.63, 0.75, and 0.90 for high-stress periods (SI = 1), neutral-stress periods (SI = 0), and low-stress period (SI = -1), respectively.

Local estimates of discrete variables

We now turn to an analysis of the local estimates of the discrete variables, including type of loan, instrumental rank, and collateral status. The estimates indicate the levels of riskiness of each category in comparison to the reference category[7]. Specifically, a negative estimate means that the category of interest has a lower RR (higher risk) than the reference category, when other variables are held constant.

Table 5.5 compares the median of the local logit estimates of all discrete variables with the coefficients estimates of QMLE-RFRV. The results show that the medians of the local estimates and the QMLE-RFRV estimates are more or less the same. These results imply that the local logit and the parametric estimators for the discrete variables contain somewhat similar information.

[———— Insert [Table 5.5] here ————]

The local logit estimates of all discrete variables are presented in Figure 5.10. Their signs are mostly in line with expectations. However, there are some unexpected positive estimates for the local estimates of senior secured bonds (Type 3) and senior unsecured bond (Type 5) in Figure 5.10a, and unexpected negative estimates for Col in Figure 5.10c. In general, both types of senior bonds are expected to have lower RRs than term loans due to their priority in

[7] The reference categories of Type, Rank and Col are given in Table 3.1

the credit capital structure. Hence, only a negative sign is expected. On the other hand, collateralised loans are commonly expected to have a higher RR than loans without collateral, in which case a positive effect is expected. As shown earlier in the simulation study, the unexpected signs of the estimates may be due to the presence of interaction effects. In what follows, we analyse the potential interaction effect between the continuous and discrete covariates.

[———— Insert [Figure 5.10] here ————]

Interaction effects between continuous and discrete variables

We find that there are three significant relationships between DC and the local estimates of both Type = {3,5} and Col = 1, which indicates interaction effects among these variables resulting in the unexpected signs in the previous analysis. Figure 5.11 shows that the effects of both senior bonds are highly dependent on the levels of DC, which might indicate interactions between DC and both senior bonds. First, for senior secured bonds, Figure 5.11a shows the unexpected positive estimates for defaulted loans with $0.2 < DC < 0.6$. Second, for the senior unsecured bond, the expected negative signs are observed only for the loan with $0.1 < DC < 0.5$ in Figure 5.11b.

[———— Insert [Figure 5.11] here ————]

An analysis of the remaining marginal effects of $Type = \{2,4,5\}$ and DC is also provided, although the effects of these types are in line with expectations. We find an interaction effect between revolving loans and DC, Figure 5.12a illustrates how the marginal effect of revolving loan depends on DC. Although, a positive marginal effect is found as expected, the figure shows that the strength of the effect depends on the level of DC. The effect is stronger as the level of DC increases from DC = 0.5, while it is relatively small effect for DC < 0.5. This implies that revolving loans are expected to have substantially higher RR than the term loan, if the level of DC is relatively high. The remaining Figures 5.12b and 5.12c represent the dependency between the marginal effects of $Type = \{4,6\}$ and the level of

DC, respectively. The interaction effects between these variables are not found.

[————— Insert [Figure 5.12] here —————]

To explain the unexpected negative estimates of Col, Figure 5.13 shows the relationship between the local estimates of Col and the levels of DC. A clear pattern is observed in Figure 5.13a, the estimates are negative, when the defaulted loan has a DC between 0.2 and 0.5.

[————— Insert [Figure 5.13] here —————]

In addition, we find that the interaction effects of Type and SI as well as Col and SI cannot be identified. Figures 5.14a to 5.14e shows the marginal effect of each type of loan conditional on the level of SI, where the effect varies substantially across values of SI. It is only in Figures 5.14c and 5.14e that some relationships are observed, which show that the negative marginal effects of $Type = \{4, 6\}$ are roughly the same across all macro-economic conditions. Also, Figure 5.15 shows that the relationship between marginal effect of Col and the levels of SI cannot be detected. Therefore, interaction effects are found among loan characteristic variables, rather than between the loan characteristic and macro-economic systematic variables.

[————— Insert [Figures 5.14 and 5.15] here —————]

To verify the findings of the interaction effect analysis, in Table 5.6, we partition the empirical RR data based on the analysis. The results confirm that the local logit analysis can uncover the true interaction effects observed in the empirical data, which is discussed as follows. We compare the average RRs of the senior secured and unsecured bonds with those of the collateralised term loans (reference category) with various ranges of DC in Panels A and B, respectively. We find that the average RR of the bonds is lower than that of term loans, if we do not consider bonds' level of DC. On the other hand, as observed in the local logit estimates analysis, the unexpected positive effects are found for the secured senior bond with $0.2 \leq DC \leq 0.6$, and the unsecured bond with $DC \leq 0.25$ and $DC \geq 0.5$. We find that the results in table 5.6 is consistent to our previous analysis. For example, although the term loans' average RR are mostly higher than those of the senior secured bonds, only bonds with $0.2 \leq DC < 0.6$ have higher RRs than term loans.

In Panel C, the results show that collateralised loans mostly have substantially higher RRs than uncollateralised loans. Only when we consider loans with $0.2 \leq DC < 0.5$, the averages RR of both loans are somewhat similar, which is in line with the analysis of the interaction between Col and DC.

[————— Insert [Table 5.6] here —————]

5.5.3 Calibrated QMLE regression for fractional response variables

In Section 5.5.2, we found nonlinear marginal effects of DC and SI, and also determined how the level of DC interacted with the covariates Type and Col. In this section, we improve the linear specification of the parametric conditional mean regression, which may include sufficient numbers of interaction terms, polynomials and discretised continuous variables, by exploiting the information provided by the estimates of the local logit model.

This exercise has three main motivations: it will (i) simplify the finding of the local logit estimates; (ii) ease the interpretation of the estimate, as most practitioners are familiar with linear regression; and (iii) provide valid statistical inferences as well as the analytical form of the marginal effects. Moreover, the result in our simulation study shows that the local logit model estimate can illustrate the correct functional form of the parametric QMLE-RFRV.

Table 5.7 reports the improved specification of the QMLE-RFRV model, where the results in Panels A, B, C, and D are based on the local logit analysis. In Panel A, we introduce an indicator function to control the range of DC, as a positive marginal effect of DC with an increasing rate was observed when DC < 0.6 in Figure 5.8a. Therefore, a quadratic function is applied. On the other hand, a simple linear function is applied to the other ranges of DC. The results are consistent with the local logit analysis, as significant increasing positive marginal effects are observed for DC < 0.6. Also, the marginal effect of DC ≥ 0.6 is not as strong as in the former given range, as the estimated parameter γ_{DC_2} is significantly smaller than γ_{DC_1}.

In Panel B, the parameter estimates of γ_{SI_1}, γ_{SI_2}, and γ_{SI_3} represent the effects of SI for low SI, high SI, and crisis SI, respectively, as the negative effects of SI weaken as SI increases. The results show that the negative effects are stronger for low SI, followed by high SI and crisis SI, respectively, since $\gamma_{\text{SI}_1} < \gamma_{\text{SI}_2} < \gamma_{\text{SI}_3}$. In addition, the effect of the crisis SI, γ_{SI_3}, is insignificant, which is in line with the observed high variation of the local logit estimate of SI > 1.5 in Figure 5.8b.

Panels C and D take into account of the four interaction effects analysed in Section 5.5.2, including $Type = \{2, 3, 5\}$ and $Col = \{1\}$, which interact with DC. The results show that the signs of the parameter estimates are consistent with the local logit analysis. For example, the interaction effects between DC and the senior secured bonds are reported as $\gamma_{\text{T}_{31}}$, $\gamma_{\text{T}_{32}}$, and $\gamma_{\text{T}_{33}}$ in Panel C. The results show that only the $\gamma_{\text{T}_{32}}$ estimate is significantly positive, which represents the interaction effect between senior secured bonds and $DC \in [0.2, 0.6)$. This means that senior secured bonds with $0.2 \leq DC < 0.6$ are likely to have higher RRs than term loans, which is consistent with our previous findings in the local logit model. For Panel D, a negative marginal effect of Col is observed only when DC is between 0.2 and 0.5, γ_{C_2}, but the effect is insignificant. This result and together with the finding in Table 5.6 might imply that the negative interaction effect of Col and DC found in the local logit analysis would also be insignificant.

[———— Insert [Table 5.7] here ————]

To indicate the statistical validity of the improved functional form, we apply a wild bootstrap-based specification test with 1,000 iterations. The QMLE-RFRV with improved specification in Table 5.7 is the null hypothesis against the local logit model alternative. According to the p-value computed by the bootstrap method, we cannot reject the null hypothesis at the 5% level of significance with a p-value of 0.09. This offers statistical evidence that the calibrated model specification fits the RR data well[8].

5.5.4 Model selection based on in- and out-of-sample predictive performance criteria

In this chapter, we employ two predictive performance criteria to evaluate the predicted RR, which are point prediction evaluation, and quantile of simulated RR portfolio predictive evaluation.

Point prediction evaluation

We use three methods to partition the full samples into in- and out-of-samples and assess the sensitivity of the models' predictions to these methods. The three methods include:

(DF1) Partition the full sample randomly into a pre-specified 70:30 ratio of in-sample and out-of-sample, for 1,000 iterations. According to our empirical RR data, one borrower could have several defaulted loans. Randomly partitioning the full data allows overlapping information, as a borrower's information could be included in both the in- and out-of-sample data.

(DF2) Partition the full sample into, for example, an in-sample period 1994-2005, and an out-of-sample period 2006-2012. This way of partitioning ensures that there are no overlapping observations in the samples. This definition

[8] We also conduct the specification test on QMLE-RFRV with standard linear functional form. As a result and we reject the null hypothesis with a p-value of 0.72

also mimics the application of RR predictive models in practice, as banks would want to use the full set of observed data to predict the RR in the forthcoming years (see chapter 3, section 3.5.2 for the details and discussions).

(DF3) Select any particular year as the out-of-sample period, and the remaining years as the in-sample period. For example, the out-of-sample period is the start of the GFC, 2008, then the in-sample period is 1994-2007 and 2009-2012. This way of partitioning the in- and out-of-sample periods is very useful to predict RR at the various phases of the economic cycle.

Quantile of RR portfolio prediction evaluation

In this method, we evaluate the predictive performance of the models at various quantiles of the simulated RR portfolio distribution of the out-of-sample period (Altman & Kalotay, 2014). The following re-sampling procedure is employed to construct the RR portfolio distribution:

(Step 1) Define the in-sample data period as 1994-2004 and the out-of-sample period as 2004-2012

(Step 2) Draw a random sample of 100 RRs from the out-of-sample data with replacement. Assign each loan a $1.00 face value and construct an equally-weighted portfolio of the selected RRs. This RR portfolio represents the money that is recovered from a portfolio with a $100.00 face value.

(Step 3) Predict the selected out-of-sample RR using the benchmark QMLE-RFRV model, the local logit model, and the calibrated QMLE-RFRV model. The predicted RR portfolio is then constructed for each model.

(Step 4) Repeat (Step 2) and (Step 3) above 10,000 times and construct a simulated RR portfolio distributions for the three models under investigation.

The model performance is evaluated according to the predictive error of a simulated RR portfolio at various quantiles of the distribution.

5.5.5 Empirical results of predictive performance evaluations

In what follows, the predictive performance criteria in previous section are applied to compare the predictive power of the models included in this chapter.

Point prediction accuracy

We adopt the data partitioning method DF1. The out-of-sample MSE and MAE of the local logit model are 0.0824 and 0.2750, respectively. For the calibrated linear model, they are 0.0854 and 0.2880, respectively. On the other hand, the benchmark model has the highest predictive errors, 0.0964 and 0.3246. These results indicate that the local logit model outperforms the others.

The results of the out-of-sample evaluation of the models for DF2 are reported in Table 5.8, which includes the predictive performances of 11 different out-of-sample windows from 2001 to 2012. For the first window, we estimate the models for the in-sample period 1994 to 2000, and evaluate the predictions of the out-of-sample period 2001 to 2012. Then, the in-sample window is continually expanded by each calendar year until the eleventh window in-sample period is 1994 to 2010 and the out-of-sample period is only 2011 to 2012. The MSE and MAE of the predictions for each window are reported in Table 5.8.

[———— Insert [Table 5.8] here ————]

The results show that the proposed local logit model has the highest predictive accuracy, followed by the calibrated model. The benchmark model outperforms the proposed model only in the two out-of-sample windows of 2002-2012 and 2010-2012 under both the MSE and MAE criteria at the 5% level of significance

(Table 5.8). The table also provides the averages and variances of MSE and MAE over 11 windows. The MSE and MAE averages of the proposed local logit model as well as their variances are consistently lower than those of the benchmark model.

Noticeably, the differences in MSE among the three models are large for the out-of-sample predictions between 2004 and 2008 in Table 5.8. These years are crucial, since they partially cover the global financial crisis period of 2007 to 2010. The benchmark model is highly sensitive to the crisis year compared to the non-parametric and calibrated models. The MSE and MAE are very large during the crisis period for the benchmark model. The benchmark model's low accuracy during the GFC could be due to the unexpected shock with a substantially high level of SI. As a linear model, the constant negative effect of SI could lead to an underprediction of RR during the crisis.

[———— Insert [Table 5.9] here ————]

The results of the point prediction evaluations of the three models for DF3 are presented in Table 5.9, where we predict RR for every year from 2000 to 2011. Table 5.9 shows that the local logit regression consistently outperforms the benchmark regression. The MSE and MAE averages of the proposed model across the 11 year-period are 0.087 and 0.224, compared to 0.096 and 0.235 for the benchmark model. The benchmark model's predictions outperform the proposed model only in 2010 and 2011. The calibrated model mostly outperforms the benchmark model and its performance is comparable to that of the local logit model. As far as the economic cycle is concerned, the local logit model and the calibrated QMLE-RFRV model have comparable performance and outperform the benchmark model at all window sizes. In addition, the MSEs of the local logit and calibrated models are substantially lower than those of the benchmark model during the GFC period. On the other hand, we observe that the benchmark model yields a relatively high MSE during the recent GFC periods (2007-2009) when SI level is at its peak.

Quantile of RR portfolio prediction accuracy

We evaluate the performances of the models at various quantiles of the simulated portfolio distribution. The results in Table 5.10 compare RRs at the 0.05, 0.25, 0.5, 0.75, and 0.95 quantiles of the observed RR portfolio distribution with those of the predicted portfolio distributions. The local logit model and the calibrated model predict the RR portfolio at the five selected quantiles of the distribution more precisely than the benchmark model does. For example, at the 0.5 quantile of the portfolio distribution, the actual portfolio can recover $63.96 from a $100.00 face value. The predictions of both the proposed model and the calibrated model are approximately $61.30, compared to the benchmark model's prediction of $67.71. This implies that the benchmark model is more likely then the other two models to overestimate the RR portfolio value. In addition, we find that the local logit model outperforms the other models for the high-risk portfolios at the low quantiles followed by the calibrated model.

[———— Insert [Table 5.10] here ————]

In summary, the proposed local logit model outperforms the other two models as indicated by all predictive performance measure criteria in both the point and quantile predictions. We also find that the calibrated model has slightly lower predictive power than the proposed model, and outperforms the benchmark model.

5.6 Conclusion

In this chapter, we propose a nonparametric local logit model for [0,1] bounded response variables and assessed its finite sample properties relative to the QMLE regression for fractional response variables (QMLE-RFRV), which we use as the benchmark model. These two models were then applied to empirical RR data and covariates. The results of the marginal and interaction effect analyses of the local logit model were utilised to calibrate the QMLE-RFRV model. The in-sample and out-of sample predictive performances of the three models were assessed using the MSE and MAE measures. The main findings of this study can be summarized as follows.

First, an extensive simulation study establishes that the properties of local logit model estimates are as good as those of the correctly specified parametric model in moderate sample sizes, and they are robust to asymmetric and bimodal error distributions. Second, our application of local logit regression to model RR data uncovered the underlying nonlinear RR data and covariate relationships, including interaction effects among covariates. Third, the results of the local logit model were used to improve the parametric QMLE-RFRV model specification, producing what we call the ”calibrated model”. The calibrated model is nonlinear in variables, which includes some useful interaction terms. Fourth, we assessed the in-sample and the out-of-sample RR predictability of the local logit model and the calibrated model in comparison to the standard parametric model. The results show that the local logit model outperforms the others. In addition, the calibrated model is comparable to the local logit model in terms of predictive performance.

An attractive feature of the local logit and calibrated models is that they outperform the benchmark model in out-of-sample RR prediction, particularly during the crisis period. Our findings are useful to applied researchers and practitioners who are unfamiliar with the nonparametric machinery. They can also be used by banks to design treatment programs for their borrowers. More importantly, the effect of conditional marginal effects on loan characteristics can be locally estimated for each defaulted loan recoveries.

5.7 Appendix D: Additional results of the simulation study

D1. Local logit model's predictive performance

In this section, we compare the proposed model's accuracy rate in predicting the simulated response variables of A1 to A3 with that of the benchmark model (see section 5.3.1 for details of data generating process).

D1.1 Univariate data-generation process A1

In terms of bandwidth, the least-squares cross-validation using U1 data suggests that the selected bandwidth is a hundred[9] in both sample sizes n = 200 and 500. As the selected bandwidth is substantially large, both the in-sample and out-of-sample predictive performances are identical for the proposed model and the benchmark model. This result is in line with our expectation, as U1 has standard linear specification. Then, the local logit model is expected to converge to the benchmark model, causing the substantial large selected bandwidth.

The selected bandwidths of U2 and U3, on the other hand, are 0.2 and 0.3, respectively, for n = 200. As the sample size increases to 500, these bandwidths are reduced to 0.15 and 0.22, respectively. Figures5.16a - 5.16d and Figures 5.16e - 5.16f report the predictive performances of the models for U2 data using MSE and MAE, respectively. These figures indicate that both in-sample and out-of-sample accuracies are similar across the two models. However, the results indicate that the proposed model's in-sample predictions are slightly better than those of the benchmark model in Figures 5.16a, 5.16b, 5.16e, and 5.16f. This is common for a nonparametric regression, which can accurately fit in-sample data, especially with

[9] The size of 100 is restricted as the maximum size for our cross-validation algorithm. The further increase in the size of the bandwidth has almost no effect in the estimation. The substantial large bandwidth causes a converging in local parameters to a global parameter

a small bandwidth. However, substantial differences between the in-sample and out-of-sample predictions are observed. Nevertheless, our results show that the MSE and MAE between the in-sample and out-of-sample data are not markedly different. Moreover, not only the average MSE and MAE are similar, but also their variations across a number of simulations.

The models' predictabilities under the U3 assumption are shown in Figure 5.17, which indicates results similar to those of U2. However, the results show that the out-of-sample MSE and MAE are almost identical when a large sample size is considered in Figures 5.17d and 5.17h.

[————— Insert [Figure 5.16 and 5.17] here —————]

D1.2 Bivariate data-generation process A2

Similar to U1, the relatively large bandwidth of 100 is selected for B1, as both X_1 and X_2 have linear effects. Since the proposed model is identical to the benchmark model, the results of B1 are not discussed further. As B2 has a mixture of linear and quadratic functions, our results suggest the bandwidths of 55.56 for X_1 and 0.29 for X_2. The bandwidths become 99.00 and 0.25, respectively, for n = 500. Thus, relatively large bandwidths are selected for the linear effect of X_1, while smaller bandwidths are chosen for the nonlinear effect of X_2. Figure 5.18 evaluates the in-sample and out-of-sample predictabilities of the models in terms of MSE and MAE, respectively. The results suggest that the proposed model and the benchmark have largely similar behaviors in all cases. Figures 5.18a and 5.18b show that the proposed model has lower in-sample MSEs than the benchmark, while the MAEs are similar in Figures 5.18e and 5.18f. On the other hand, when the out-of-sample prediction is considered, the benchmark model slightly outperforms the proposed model based on both average MSE and MAE.

For B3, the bandwidths are 0.648 for X_1 and 0.646 for X_2, for n = 200. They then decrease to 0.583 and 0.528 as n = 500. The bandwidths are relatively small, since both underlying effects are highly non-linear. We find that the out-of-sample predictions of the proposed model have a lower accuracy rate than those of the benchmark, especially for the small sample size as shown in Figures 5.19c and 5.19g. However, the differences are reduced when n increases, although the benchmark model still consistently outperforms the proposed model. An increase in the number of in-sample data points leads to an improvement in the local estimation.

[———— Insert [Figures 5.18 and 5.19] here ————]

D1.3 Multivariate data-generation process A3

For the bandwidths, the kernel functions are assigned depending on the types of the variables, where (5.2.5) is employed for X_1 and X_2 and (5.2.6) for D_1 to D_3. To facilitate computational optimisation, we allow a single bandwidth for all three discrete variables. As a result, the selected bandwidths are 3.68 for x_1, 1.36 for x_2, and 0.79 for all discrete variables. The bandwidth of x_1 is relatively large, which is in line with expectations, as the true function is linear. However, the bandwidth is not substantially large as before, which indicates that x_1 also interacts with d_2, causing the change in slope, when $d_2 = 1$.

The predictive performances of all models are reported in Figure 5.20, given n = 500. They show that the in-sample predictabilities of the proposed and benchmark models are more or less the same, while the benchmark model's accuracy slightly outperforms that of the proposed model in the out-of-sample performance in Figures 5.20b and 5.20d.

[———— Insert [Figure 5.20] here ————]

D2. Local logit marginal effect estimates

To illustrate the proposed model's application for estimating marginal effects, a local logit analysis is conducted to recover the underlying effects of the simulated covariates through the model's estimates. The analysis is then evaluated by comparing its results with those of the benchmark model. This section provides the analysis of univariate and bivariate data-generation process[10] A1 and A2.

D2.1 Univariate data-generation process A1

As a two-sided censoring is assumed for U1, we plot the function $y_i = max(1, min(2(x_{i1}))$ as the dark solid line in Figure 5.21. It shows a sharp turning point at $x_1 = 0$, as a linear function is observed when $0 \leq x_1 < 2$. The results show that the models have identical graphical outcomes in Figures 5.21a and 5.21b for the proposed model with n = 200 and 500, respectively, compared to Figures 5.21c and 5.21d for the benchmark model. All figures indicate good approximations of the true underlying function. However, all results in Figure 5.21 show a smoother transition at x1 = 0 than the true function. This could be due to the application of the quasi-likelihood estimation, where the two-sided censoring assumption remains unknown for both models.

As a quadratic functional form is assumed for U2, we observe nonlinear relationships only when $-1 < x_1 < 1$ (see the solid line in Figure 5.22). The results clearly show that the local logit estimates can capture the nonlinear behaviors of the true shapes in Figures 5.22a and 5.22b. The estimates are also similar to the parametric estimates in Figures 5.22c and 5.22d.

Figure 5.23 illustrates the true functional plot under U3, where the sine function is employed as the true underlying relationship. A nonlinear shape is assumed only when $X \in [0, 3]$. Consistent with the previous results, the estimates from the proposed model tend to fit the true function. However, high variations are observed in the left tails of Figures 5.23a and 5.23b, where variation in the latter tends to reduce as n increases .

[————— Insert [Figures 5.21, 5.22, and 5.23] here —————]

[10] The results for multivariate data-generation process A3 have been discussed in section 5.3.2

D2.2 Bivariate data-generation process A2

Under assumptions B2 and B3[11], we denote the local logit regression estimate as:

$$\hat{y} = \Lambda(\hat{\beta}_0(x) + \hat{\beta}_1(x)x_1 + \hat{\beta}_2(x)x_2). \tag{5.7.1}$$

The local logit analysis of each assumption is discussed in what follows.

Bivariate assumption B2

The correctly specified benchmark model estimate under assumption B2 is denoted as:

$$\hat{y} = \Lambda(\hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 x_2^2). \tag{5.7.2}$$

Figure 5.24 compares the local estimators $\hat{\beta}_1(x)$ with the parametric estimator $\hat{\gamma}_1$. The results show that the local estimators, in Figures 5.24a and 5.24b, are constant for all values of X_1. This is in line with the underlying linear effect in assumption B2 as well as the parametric estimators in Figures 5.24c and 5.24d. Furthermore, although the variations of the local estimators in Figures 5.24a and 5.24b are higher than those of the parametric estimators in Figures 5.24c and 5.24d, the local estimators' variations decrease as the number of observation increases. However, we find that the averages of both the local and parametric estimators are similar at approximately 0.9.

[————— Insert [Figures 5.24 and 5.25] here —————]

To illustrate the estimators for a quadratic form on x_2, we firstly provide the local estimators $\hat{\beta}_2(x)$ in Figures 5.25a and 5.25b for n = 200 and 500, respectively. The estimators are negative for $x_2 \in [-2, 0)$ with an increasing rate until they become positive for $x_2 \in (0, 2]$. In addition, the local estimators are approximately zero for $x_2 = 0$. These results would imply a slope of the quadratic function with a

[11] See section 5.3.1 for details of data generating process. Also, we do not consider B1, as the result of the local logit is identical to the benchmark model in section 5.3.2

turning point at zero. Therefore, we then compare our local logit estimates with the quadratic slope estimates based on the benchmark model, $2\hat{\beta}_2 x_2$, in Figures 5.25c and 5.25d. The results indicate same effects of x_2.

In conclusion, our results in Figures 5.24 and 5.25 demonstrate that the local estimators $\hat{\beta}_1(x)$ and $\hat{\beta}_2(x)$ are good approximations of the benchmark model's estimators $\hat{\beta}_1$ and $2\hat{\gamma}_2 x_2$, respectively, with no prior knowledge about the functional form.

Bivariate assumption B3

The functional form in B3 is a combination of two nonlinear functions: sine and quadratic forms. The local logit regression estimates are the same as B2 in (5.7.1), while the correctly specified estimates for the benchmark model are:

$$\hat{y} = \Lambda(\hat{\gamma}_0 + \hat{\gamma}_1 \sin x_1 + \hat{\gamma}_2 x_2 + \hat{\gamma}_3 x_2^2). \tag{5.7.3}$$

First, Figures 5.26c and 5.26d reveal the slope estimates of x_1 based on the benchmark model in (5.7.3), which is defined as $\hat{\gamma}_1 \cos(x_1)$. Comparing the benchmark model with the proposed model, Figures 5.26a and 5.26b are the local estimators $\hat{\beta}_1(x)$, which show nonlinear behaviours similar to those of $\hat{\gamma}_1 \cos(x_1)$. However, the local estimators corresponding to the relatively high values of $x_1 > 8$ have noticeably high variations of the simulations (see Figure 5.26a). This could be due to the low information in this range, as the data are generated from the Chi-square distribution with three degrees of freedom, which has highly positive skewness. As the sample increases to $n = 500$ in Figure 5.26b, the local estimates show lower variation compared to Figure 5.26a, with similar shapes.

Second, a quadratic function is assumed for x_2, hence slopes with a constant increasing rate are expected. Figure 5.27c and 5.27d show the slope estimates of x_2 for the benchmark model defined as $\hat{\gamma}_2 + 2\hat{\gamma}_3 x_2$. These figures are compared with Figures 5.27a and 5.27b, which are the local estimators $\hat{\beta}_2(x)$. The results show that the local estimators have approximately linear shapes, reflecting the slopes of the quadratic function.

[————— Insert [Figures 5.26 and 5.27] here —————]

Overall, the simulation experiments for the bivariate settings show that the local estimators contain information about the slopes of the functions. The results indicate that the estimators provide informative approximations of the correctly specified parametric estimators. The approximations seem to improve in precision as sample size increases.

5.8 Appendix E: Tables and figures

n = 200	In-sample				Out-of-sample			
	Benchmark		LL		Benchmark		LL	
Specification	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
U1	0.0156	0.0720	0.0156	0.0720	0.0161	0.0722	0.0161	0.0722
U2	0.0195	0.0825	0.0182	0.0811	0.0199	0.0825	0.0200	0.0831
U3	0.0325	0.1362	0.0312	0.1361	0.0331	0.1382	0.0333	0.1399
B1	0.0211	0.1022	0.0211	0.1022	0.0215	0.1028	0.0215	0.1028
B2	0.0312	0.1299	0.0298	0.1299	0.0314	0.1300	0.0322	0.1312
B3	0.0260	0.1109	0.0234	0.1102	0.0256	0.1129	0.035	0.1271

Table 5.1: *In-sample and out-of-sample predictions of models in the simulation study with small sample size*

Note: The benchmark model is the standard QMLE-RFRV. The local logit regression is denoted as LL

n = 500	In-sample				Out-of-sample			
	Benchmark		LL		Benchmark		LL	
Specification	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
U1	0.0121	0.0682	0.0121	0.0682	0.0112	0.0691	0.0112	0.0691
U2	0.0191	0.0841	0.0187	0.0827	0.0181	0.0811	0.0182	0.0811
U3	0.0333	0.1355	0.0326	0.135	0.0325	0.1386	0.0326	0.1388
B1	0.2041	0.0958	0.2041	0.0958	0.0207	0.1005	0.0207	0.1005
B2	0.0309	0.1255	0.0301	0.1256	0.0313	0.1291	0.0314	0.1300
B3	0.0261	0.11	0.0241	0.1021	0.0252	0.1123	0.0259	0.1151
M1	0.1100	0.2731	0.109	0.2722	0.1154	0.2781	0.1162	0.2812

Table 5.2: *In-sample and out-of-sample predictions of models in the simulation study with moderate sample size*

Note: The benchmark model is the standard QMLE-RFRV. The local logit regression is denoted as LL

Relative mean squared error

	n = 200		n = 500		n= 1000	
Error (U) distribution	Benchmark	LL	Benchmark	LL	Benchmark	LL
Panel (a): In-sample prediction						
Specification: M1						
Chi-square	1.9618	0.7603	1.8828	0.9218	1.8645	0.9773
Bimodal	1.0893	1.1003	1.0844	0.9592	1.0839	0.9838
Specification: M2						
Chi-square	1.3811	0.9206	1.3685	0.8512	1.1466	0.9857
Bimodal	1.0615	1.0392	1.0600	1.0296	1.0569	0.9350
Panel (b): Out-of-sample prediction						
Specification: M1						
Chi-square	1.9178	1.4636	1.9816	1.2444	1.9005	1.1072
Bimodal	1.0854	1.1191	1.0800	1.0501	1.0750	1.0304
Specification: M2						
Chi-square	1.3727	1.4701	1.3553	1.2094	1.3713	1.1466
Bimodal	1.0491	1.0835	1.0572	1.0503	1.0527	1.0464

Table 5.3: *Mean square error of the local logit model relative to correctly specified QMLE-RFRV model*

Note: The benchmark model is the standard linear QMLE-RFRV model. LL is the proposed local logit model. The relative MSE is the MSE of the given model relative to the correctly specified QMLE-RFRV model. The error $U^{(1)} \sim \chi^2_{(1)}$ - asymmetric distribution. The error $U^{(2)}$ generated from the equally weighted mixture of $N(-2,1)$ and $N(2,1)$ - bimodal distribution.

Relative mean absolute error

	n = 200		n = 500		n= 1000	
Error (U) distribution	Benchmark	LL	Benchmark	LL	Benchmark	LL
Panel (a): In-sample prediction						
Specification: M1						
Chi-square	1.5536	0.8819	1.5366	0.9959	1.5341	1.0272
Bimodal	1.0664	1.0524	1.0622	0.9962	1.0627	1.0131
Specification: M2						
Chi-square	1.2464	0.9099	1.2410	0.9089	1.2370	1.0050
Bimodal	1.0408	1.0148	1.0413	1.0250	1.0380	0.9618
Panel (b): Out-of-sample prediction						
Specification: M1						
Chi-square	1.5434	1.1824	1.5612	1.1310	1.5219	1.0807
Bimodal	1.0627	1.0597	1.0594	1.0412	1.0563	1.0357
Specification: M2						
Chi-square	1.2376	1.1679	1.2301	1.0880	1.2440	1.0859
Bimodal	1.0364	1.0404	1.0397	1.0356	1.0360	1.0219

Table 5.4: *Mean absolute error of the local logit model relative to correctly specified QMLE-RFRV model*

Note: The benchmark model is the standard linear QMLE-RFRV model. LL is the proposed local logit model. The relative MAE is the MAE of the given model relative to the correctly specified QMLE-RFRV model. The error $U^{(1)} \sim \chi^2_{(1)}$ - asymmetric distribution. The error $U^{(2)}$ generated from the equally weighted mixture of $N(-2,1)$ and $N(2,1)$ - bimodal distribution.

Variables	Parametric coefficients (γ)	Median of local logit coefficients ($\beta(x)$)
Type of loan		
$Type^{(2)}$	0.4907***	0.4376
	(0.1358)	
$Type^{(3)}$	-0.0229	-0.0328
	(0.1433)	
$Type^{(4)}$	-0.3771	-0.3857
	(0.2331)	
$Type^{(5)}$	0.2535	0.2899
	(0.2010)	
$Type^{(6)}$	-0.4258	-0.4417
	(0.2476)	
Rank		
Rank 2	-0.4512***	-0.5076
	(0.1049)	
Rank 3	-0.7900***	-0.8919
	(0.1506)	
Rank 4	-0.9300***	-1.0193
	(0.1896)	
Collateral		
Collateralized loan	0.4420**	0.6127
	(0.1842)	

Table 5.5: *Estimates of the QMLE-RFRV model and the local logit regression*

Note: The median of the local estimates are calculated based on the results in Figure 5.10

Variables	LL estimates	Sample averages RR	
		Category of interest	Reference category
Panel A Senior secured bond			
$DC \in [0,1]$	N/A	0.59	0.71
$DC < 0.2$	Negative	0.44	0.53
$0.2 \leq DC < 0.6$	Positive	0.82	0.71
$DC \geq 0.6$	Negative	0.79	0.91
Panel B Senior unsecured bond			
$DC \in [0,1]$	N/A	0.43	0.71
$DC < 0.2$	Positive	0.41	0.34
$0.2 \leq DC < 0.5$	Negative	0.47	0.74
$DC \geq 0.5$	Positive	0.77	0.36
Panel C Loans with collateral			
$DC \in [0,1]$	N/A	0.73	0.37
$DC < 0.2$	Positive	0.53	0.31
$0.2 \leq DC < 0.5$	Negative	0.68	0.63
$DC \geq 0.5$	Positive	0.92	0.65

Table 5.6: *Average recovery rates of senior bonds and collateralized loans for various ranges of DC*

Note: The empirical RR is partitioned based on the findings in the interaction analysis of local logit model estimates of senior bonds and collateral status. Then RR of each sub-sample is compared with the reference group. For Panels A and B, given the ranges of DC in the first column, the categories of interest are the senior secured and unsecured bonds, respectively, and the reference categories are collateralized and uncollateralized term loans, respectively. For Panel C, the the category of interest is the collateralized loan, whereas the reference category is the uncollateralized loan.

Coefficients	Variables	Estimates	(SE)	
Panel A: Debt cushion				
γ_{DC_1}	$I(DC < 0.6)DC^2$	5.5071	(0.5890)	*
γ_{DC_2}	$I(DC \geq 0.6)DC$	2.4200	(0.2105)	*
Panel B: Stress index				
γ_{SI_1}	$I(SI < 0)SI$	-1.2283	(0.1436)	*
γ_{SI_2}	$I(0 \leq SI < 1.5)SI$	-0.4589	(0.0997)	*
γ_{SI_3}	$I(SI \geq 1.5)SI$	-0.0212	(0.0330)	
Panel C: Types of loan				
$\gamma_{T_{21}}$	$Type^{(2)}$	0.0817	(0.1124)	
$\gamma_{T_{22}}$	$I(DC > 0.8)Type^{(2)}$	0.2770	(0.5021)	
$\gamma_{T_{31}}$	$I(DC < 0.2)Type^{(3)}$	-0.4516	(0.1912)	*
$\gamma_{T_{32}}$	$I(0.2 \leq DC < 0.6)Type^{(3)}$	0.8757	(0.1883)	*
$\gamma_{T_{33}}$	$I(DC \geq 0.6)Type^{(3)}$	-1.5847	(0.4087)	*
$\gamma_{T_{41}}$	$Type^{(4)}$	-0.3635	(0.1265)	*
$\gamma_{T_{51}}$	$I(DC < 0.2)Type^{(5)}$	0.2713	(0.0574)	*
$\gamma_{T_{52}}$	$I(0.2 \leq DC < 0.5)Type^{(5)}$	-0.0456	(0.1806)	
$\gamma_{T_{53}}$	$I(DC \geq 0.5)Type^{(5)}$	0.8014	(0.4795)	
$\gamma_{T_{61}}$	$Type^{(6)}$	-0.3691	(0.1431)	*
Panel D: Collateral status				
γ_{C_1}	$I(DC < 0.2)Col$	0.5802	(0.1733)	*
γ_{C_2}	$I(0.2 \leq DC < 0.5)Col$	-0.1600	(0.1383)	
γ_{C_3}	$I(DC \geq 0.5)Col$	1.2458	(0.3536)	*
Panel E: Instrumental rank				
γ_{R_2}	$Rank2$	-1.0040	(0.0901)	*
γ_{R_3}	$Rank3$	-1.5447	(0.1204)	*
γ_{R_4}	$Rank4$	-1.6792	(0.1422)	*

Table 5.7: *Estimates of the calibrated parametric QMLE-RFRV model*

Note: The calibrated model (CM) parameters and the corresponding variables are presented in columns 1 & 2 respectively. Panels A, B, C, D and E list nonlinear, interactive, discretised DC and discrete variables in the model

Out-of-sample			MSE			MAE		
Year	Out-of-time obs.	%	Benchmark	LL	CM	Benchmark	LL	CM
2001-2012	2,738	77%	0.1071	0.1048	0.1203	0.2449	0.2425	0.2505
2002-2012	2,168	61%	0.1035	0.1052	0.1229	0.2420	0.2445	0.2508
2003-2012	1,478	41%	0.1021	0.1007	0.1411	0.2608	0.2557	0.2672
2004-2012	1,141	32%	0.1131	0.0900	0.1017	0.2536	0.2238	0.2284
2005-2012	965	27%	0.1279	0.1182	0.1053	0.2689	0.2563	0.2410
2006-2012	779	22%	0.1439	0.1045	0.1179	0.2843	0.2464	0.2556
2007-2012	708	20%	0.1517	0.1145	0.1219	0.2932	0.2578	0.2613
2008-2012	660	18%	0.1592	0.1144	0.1245	0.3016	0.2570	0.2647
2009-2012	496	14%	0.0841	0.0825	0.0831	0.2288	0.2170	0.2273
2010-2012	121	3%	0.0643	0.0825	0.0746	0.2024	0.2065	0.2004
2011-2012	35	1%	0.0709	0.0819	0.0954	0.2092	0.2055	0.2234
Average			0.1116	0.0999	0.1099	0.2536	0.2375	0.2428
Var.			0.0009	0.0002	0.0004	0.0010	0.0004	0.0004

Table 5.8: *Out-of-sample predictive performance of models*

Note: The table employs the data partitioning (DF2). Columns 1-3 indicate the out-of-sample period, the number of the observation in out-sample period, and the percentage of these observations relative to total number. Benchmark, LL and CM represent the standard QMLE-RFRV model, local logit model and the calibrated QMLE-RFRV model, respectively.

Out-of-sample	Sample size		MSE			MAE		
year	Out-of-sample	In-sample	Benchmark	LL	CM	Benchmark	LL	CM
2001	570	3003	0.1308	0.1235	0.1274	0.2878	0.2826	0.2860
2002	690	2883	0.0903	0.0861	0.0779	0.2088	0.2102	0.2175
2003	337	3236	0.1045	0.0987	0.1049	0.2814	0.2757	0.2533
2004	176	3397	0.0690	0.0555	0.0692	0.1828	0.1678	0.1786
2005	186	3387	0.1114	0.0815	0.0776	0.2245	0.2000	0.1959
2006	71	3502	0.0950	0.0708	0.0782	0.2067	0.1932	0.1947
2007	48	3525	0.0919	0.0739	0.0756	0.1959	0.1924	0.2078
2008	164	3409	0.1322	0.0937	0.1082	0.3454	0.2911	0.3127
2009	375	3198	0.0903	0.0860	0.0911	0.2373	0.2215	0.2365
2010	86	3487	0.0615	0.0728	0.0623	0.1996	0.2028	0.1901
2011	20	3553	0.0838	0.1183	0.1001	0.2210	0.2299	0.2510
	Average		0.0964	0.0873	0.0884	0.2356	0.2243	0.2295
	Var.		0.0018	0.0007	0.0011	0.0022	0.0015	0.0017

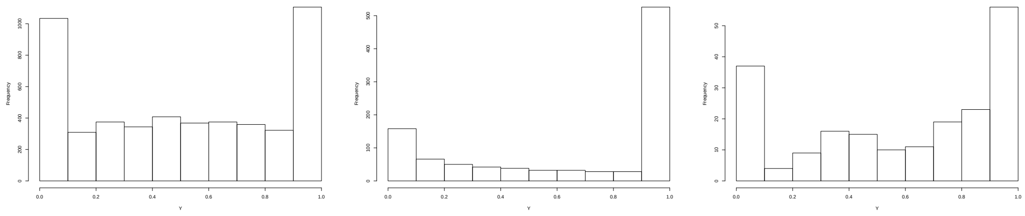
Table 5.9: *Out-of-sample predictive performance of the models over the course of economic cycle*

Note: The table employs the data partitioning (DF3). The model is estimated for the in-sample period, which excludes only one year - the out-of-sample year.

Quantiles	Actual	Benchmark	LL	CM
0.05	57.71	59.68	56.93	56.11
	(% different)	(3.4%)	(1.4%)	(2.9%)
0.25	61.40	64.54	59.80	59.25
	(% different)	(5.1%)	(2.7%)	(3.6%)
0.5	63.69	67.91	61.35	61.28
	(% different)	(6.6%)	(3.8%)	(3.9%)
0.75	65.91	70.55	62.99	62.97
	(% different)	(7.0%)	(4.6%)	(4.7%)
0.95	69.19	74.94	65.55	65.99
	(% different)	(8.3%)	(5.6%)	(4.8%)
MSE		22.09	16.52	15.72

Table 5.10: *Quantile predictive performance of the models*

Note: Portfolio distributions were generated from the out-of-sample predictions of RR by the three models.

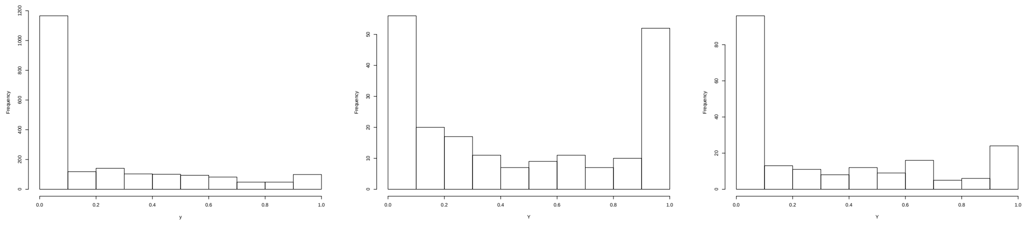


(a) *U1*

(b) *U2*

(c) *U2*

Figure 5.1: *Densities of the generated dependent variables of the assumption A1*



(a) *B1*

(b) *B2*

(c) *B3*

Figure 5.2: *Densities of the generated dependent variables of the assumption A2*

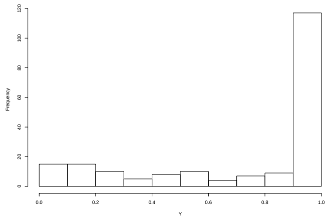
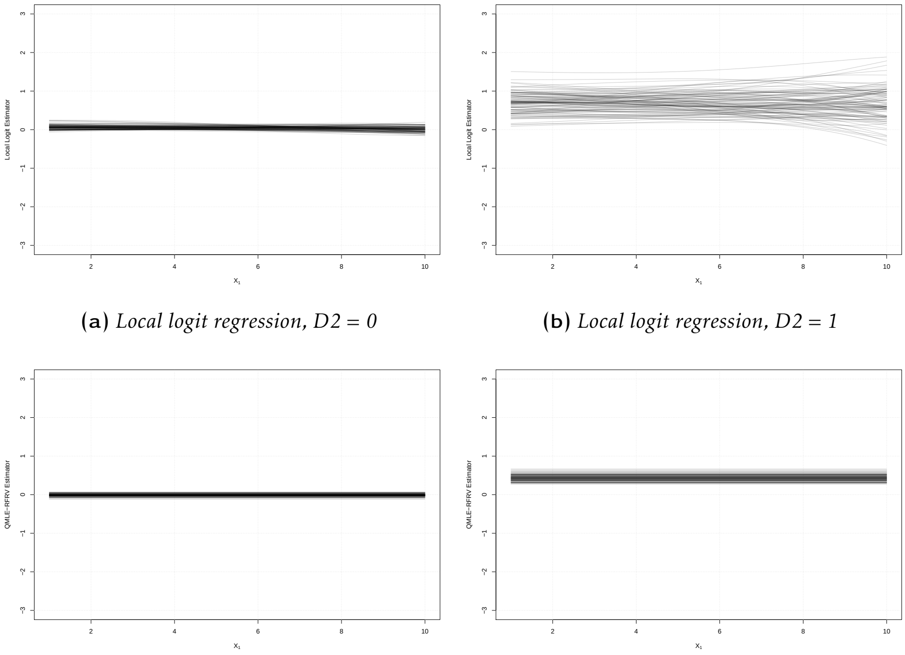


Figure 5.3: *Density of the generated dependent variable of the assumption A3*



(a) *Local logit regression, D2 = 0*

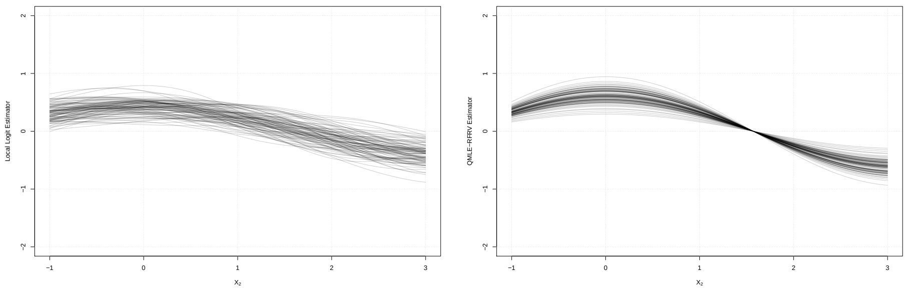
(b) *Local logit regression, D2 = 1*

(c) *QMLE-RFRV, D2 = 0*

(d) *QMLE-RFRV, D2 = 1*

Figure 5.4: *Interaction effects estimates of x_1 conditional on d_2 under simulation M1*

Note: These figures show the interaction effect estimates of x_1 and d_2 for the simulation assumption M1. (a) and (b) illustrate the local logit marginal effect estimates $\beta_1(x)$ as a function of x_1 conditional on $d_1 = 0$ and 1, respectively, as specified in (5.3.2). On the other hand, (c) and (d) represent the parametric QMLE-RFRV estimates γ_1 and $\gamma_1 + \gamma_6$, respectively, as specified in (5.3.1).

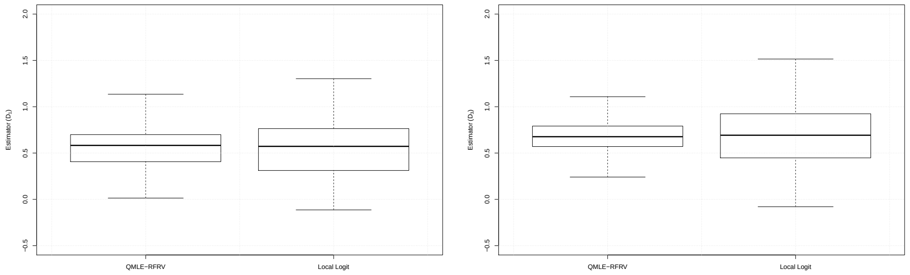


(a) *Local logit regression*

(b) *QMLE-RFRV*

Figure 5.5: *Nonlinear marginal effect estimates of x_2 under simulation M1*

Note: (a) is the local logit marginal effect estimate $\beta_2(x)$ in (5.3.2) as a function of x_2. (b) represents the parametric QMLE-RFRV estimate $\gamma_2 \cos(x_2)$ as the marginal effect estimate of x_2 in (5.3.1).

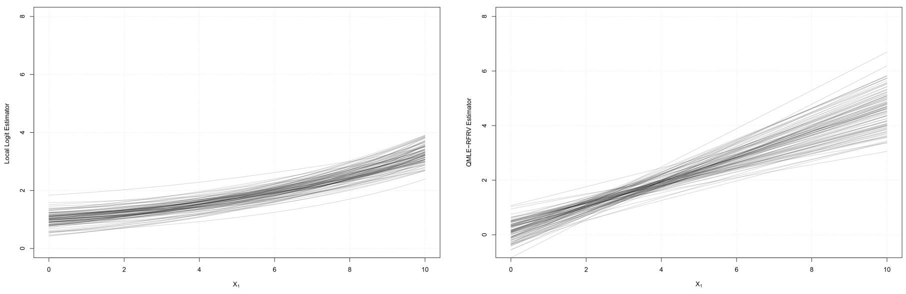


(a) D_1

(b) D_3

Figure 5.6: *Marginal effect estimates of D_1 and D_3 under simulation M1*

Note: (a) and (b) compare the marginal effect estimates of QMLE-RFRV and the local logit model for the discrete variables D_1 and D_3 under simulation assumption M1 in (5.3.1) and (5.3.2). (a) represents the marginal effect estimates of D_1, which compares $\hat{\gamma}_3$ and $\hat{\beta}_3(x)$, on the left and right hand sides of the figure, respectively. Similarly, (b) represents the comparison of the marginal effect of D_3 between $\hat{\gamma}_5$ and $\hat{\beta}_5(x)$.

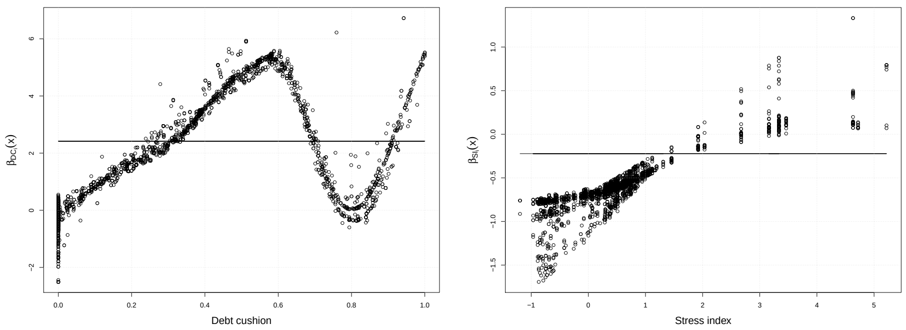


(a) *Local logit regression*

(b) *QMLE-RFRV*

Figure 5.7: *Interaction effect estimates of D_2 conditional on x_1 under simulation M1*

Note: These figures show the marginal effect estimate of d_2 as a function of x_1 as such interaction effect is specified in the simulation assumption M1. (a) illustrates the local logit marginal effect estimate $\beta_4(x)$ in (5.3.2) as a function of x_1. (b) is represents the marginal effect estimate $\gamma_4 + \gamma_6 x_1$ in (5.3.1).



(a) *Debt cushion*

(b) *Stress index*

Figure 5.8: *Local logit marginal effect estimates of the debt cushion and the stress index*

Note: (a) is the nonlinear local logit marginal effect estimate of DC, $\hat{\beta}_{DC}(x)$, as a function of DC. (b) illustrates the marginal effect estimate of SI, $\hat{\beta}_{SI}(x)$, as a function of SI. In addition, the dark solid lines in both figures are the parametric coefficients of QMLE-RFRV for DC and SI.

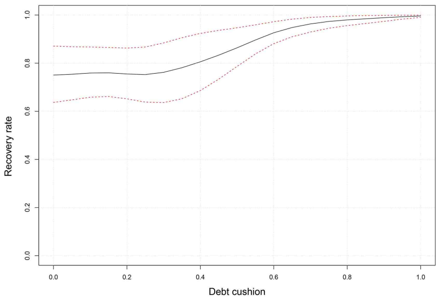
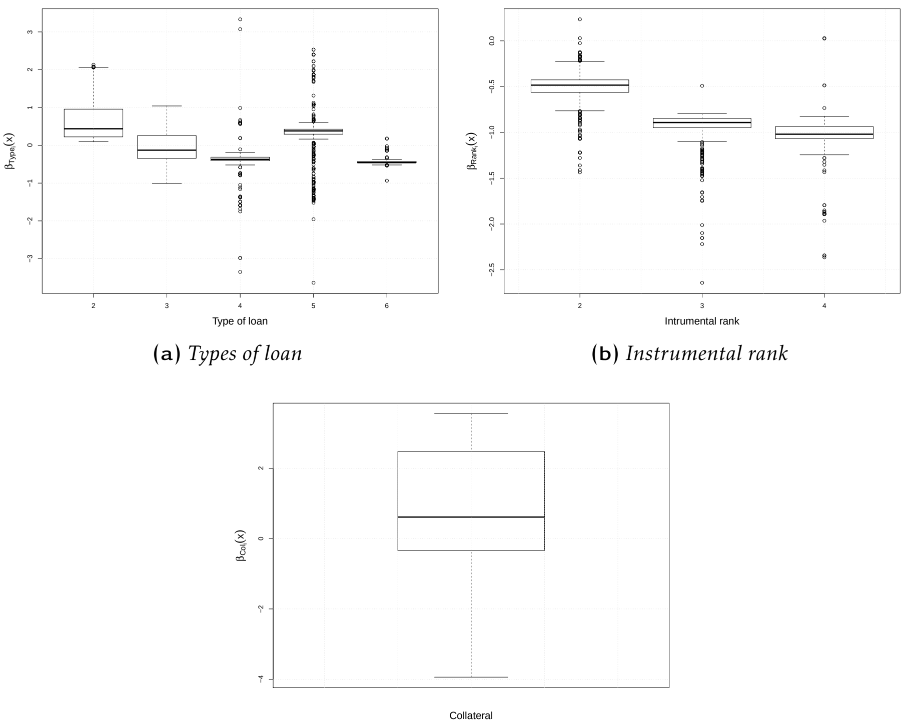


Figure 5.9: *Effect of debt cushion on the recovery rate of a specific defaulted loan*

Note: The figure illustrates the effect of debt cushion on the recovery rate of a collateralised revolving loan with rank 1 (Type = 2, Rank = 1, Col = 1) as a defined specific characteristics. We also consider the effect in three different economic scenarios by specifying the levels of the stress index. The dark solid line represents the effect of debt cushion on the recovery rate given SI = 0, while the red dashed lines represent the effects during SI = -1 for the lower bound, and SI = 1 for the upper bound.



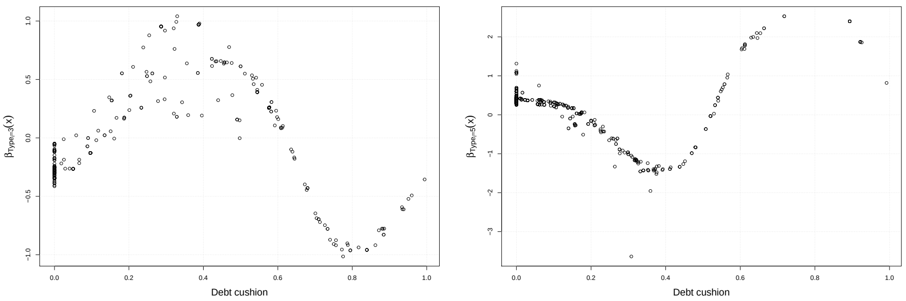
(a) *Types of loan*

(b) *Instrumental rank*

(c) *Collateral status*

Figure 5.10: *Local logit marginal effect estimates of Type, Rank, and Col categorical variables*

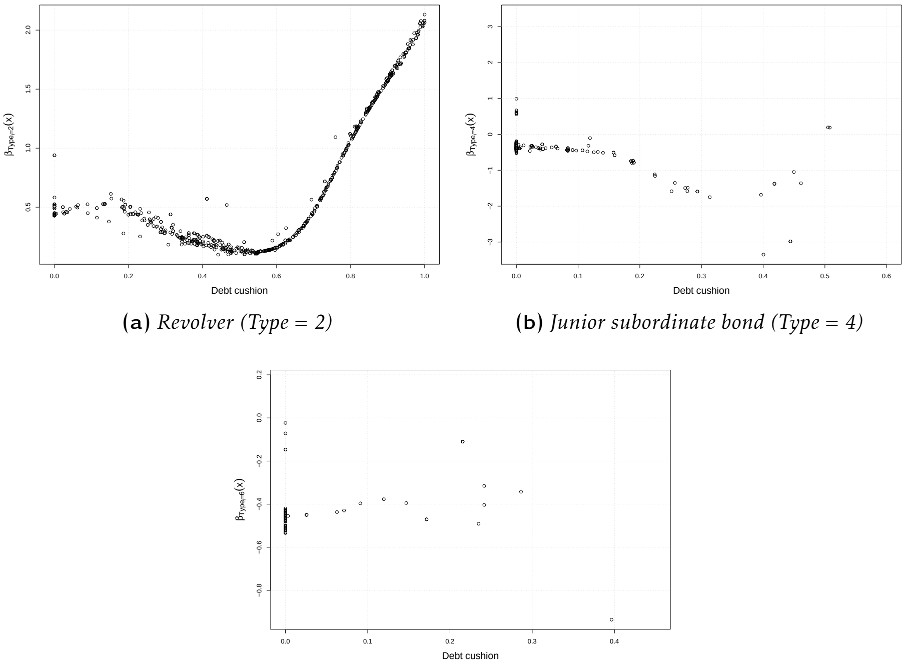
Note: The figures represent the marginal effect of each categorical covariate: (a) shows the marginal effects of 5 types of loan, Type = {2,..,6}, where the term loan (Type = 1) is the reference category; (b) shows the marginal effects of three instrumental ranks, Rank = {2,3,4}, where Rank 1 is the reference category; and (c) is the marginal effect of the collateral status, Col = 1, where uncollateralized loan is the reference category.



(a) *Senior secured bond (Type = 3)* (b) *Senior unsecured bond (Type = 5)*

Figure 5.11: *Local logit interaction effect estimates of the senior bonds conditional on level of the debt cushion*

Note: The figures illustrate the local logit marginal effect estimates of Type = {3,5}, which are $\hat{\beta}_{Type^{(3)}}(x)$ and $\hat{\beta}_{Type^{(5)}}(x)$, respectively, as a function of the debt cushion



(a) *Revolver (Type = 2)* (b) *Junior subordinate bond (Type = 4)*

(c) *Junior and subordinated bond (Type = 6)*

Figure 5.12: *Interaction effect estimates of Type = {2,4,6} conditional on the level of debt cushion*

Note: The figures illustrate local logit marginal effect estimates for given three types of loan as a function of the level of debt cushion as specified in (5.5.1): (a) represents the marginal effect estimate of revolving loan $\beta_{Type=2}(x)$ interacted with DC; (b) and (c) represent the marginal effect estimates of junior bonds $\beta_{Type=4}(x)$ and $\beta_{Type=6}(x)$, respectively, interacted with DC.

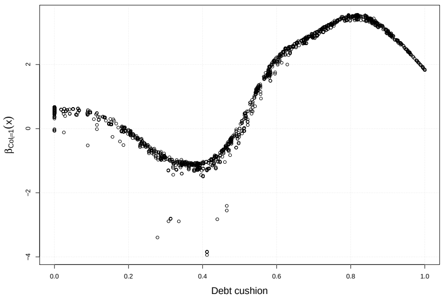
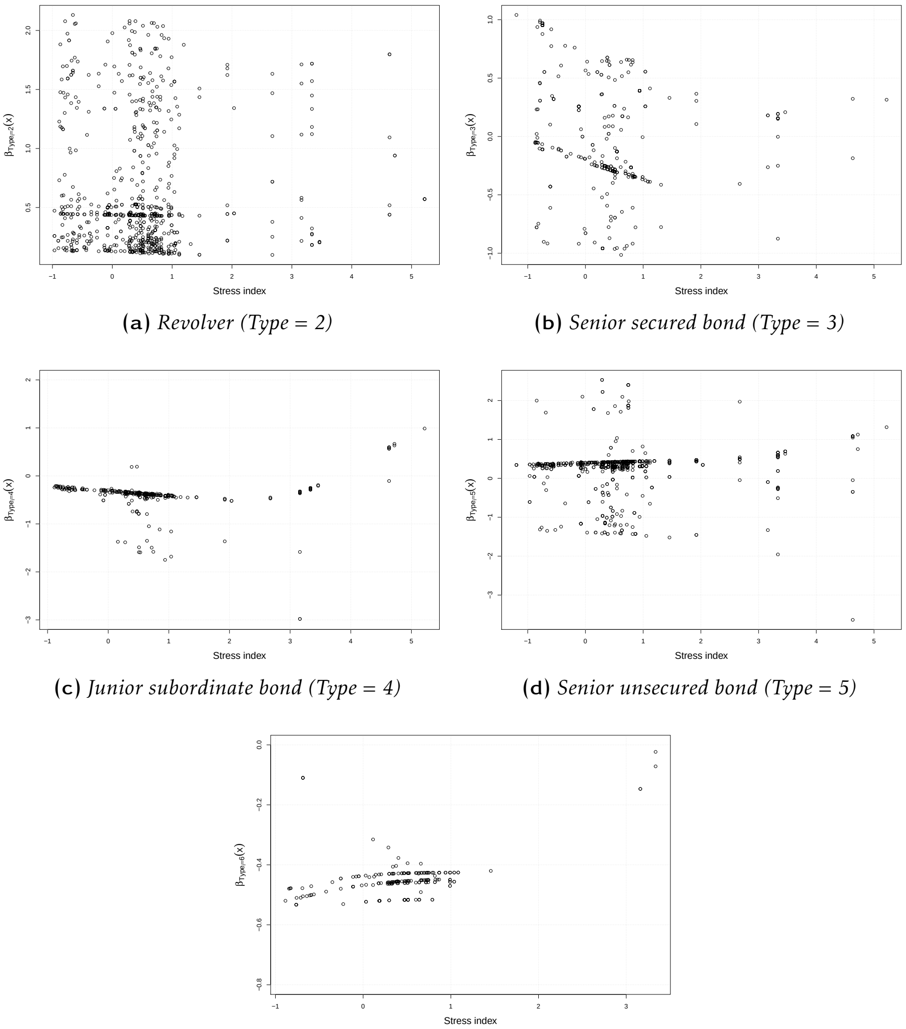


Figure 5.13: *Local logit interaction effect estimates of the collateral status conditional on level of the debt cushion*

Note: The figure illustrates the local logit marginal effect estimates of the collateralized loan, which is $\hat{\beta}_{Col}(x)$, as a function of the debt cushion



(a) *Revolver (Type = 2)*

(b) *Senior secured bond (Type = 3)*

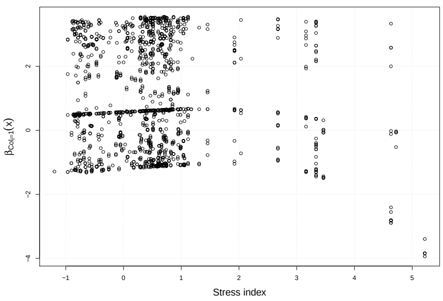
(c) *Junior subordinate bond (Type = 4)*

(d) *Senior unsecured bond (Type = 5)*

(e) *Junior and subordinated bond (Type = 6)*

Figure 5.14: *Interaction effect estimates of Type conditional on the level of the stress index*

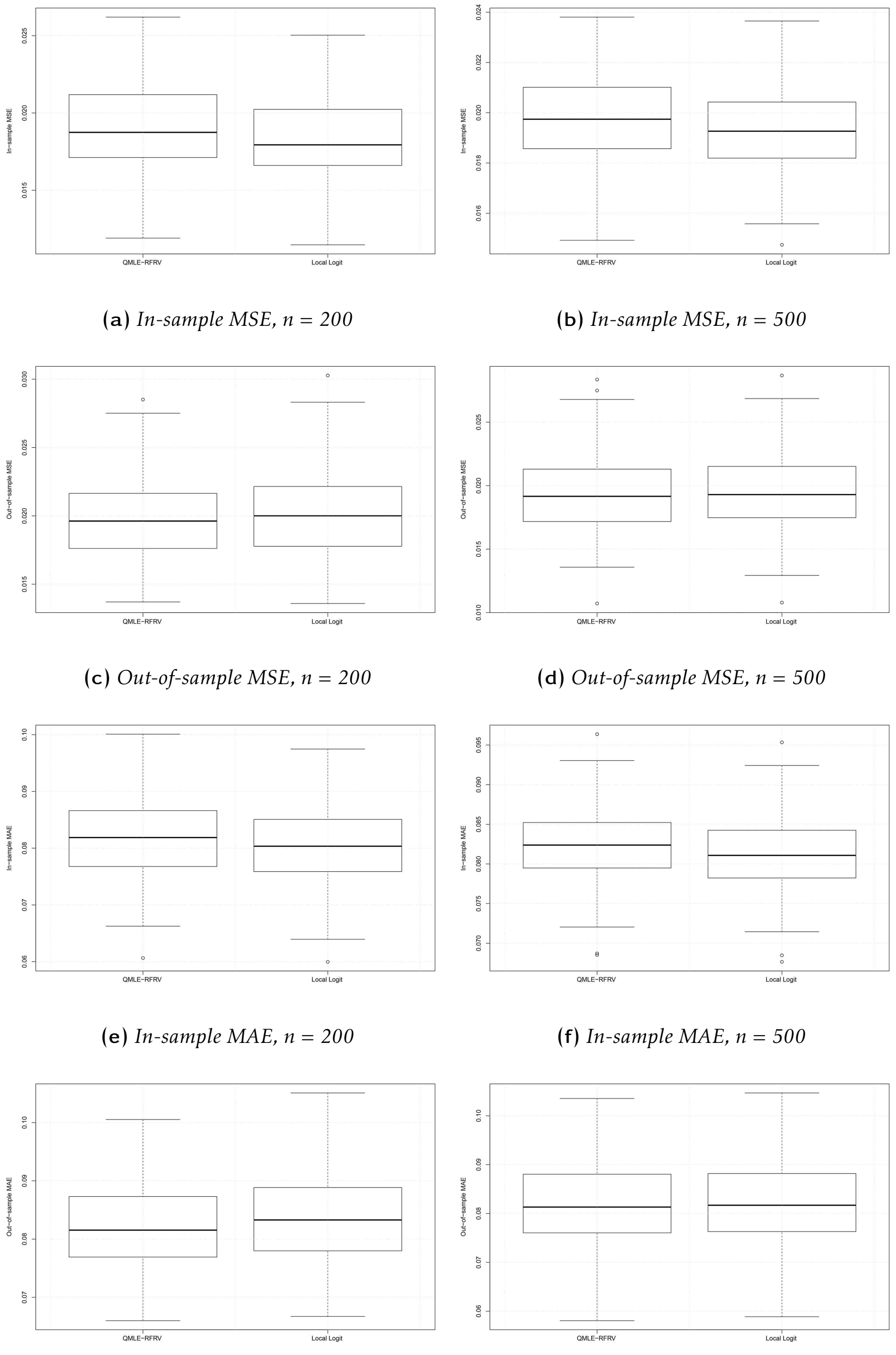
Note: The figures illustrate local logit marginal effect estimates of all five types of loan as a function of the level of the stress index as specified in (5.5.1): (a) to (e) represent the marginal effect estimates $\beta_{Type=2}(x)$, $\beta_{Type=3}(x)$, $\beta_{Type=4}(x)$, $\beta_{Type=5}(x)$, and $\beta_{Type=6}(x)$, respectively, conditional on level of SI.



(**a**) *Stress index*

Figure 5.15: *Interaction effect estimate of the collateral status conditional on the level of the stress index*

Note: The figure illustrates the local logit marginal effect estimates of the collateral status, which is $\hat{\beta}_{Col}(x)$, as a function of the stress index.



(a) *In-sample MSE, n = 200*

(b) *In-sample MSE, n = 500*

(c) *Out-of-sample MSE, n = 200*

(d) *Out-of-sample MSE, n = 500*

(e) *In-sample MAE, n = 200*

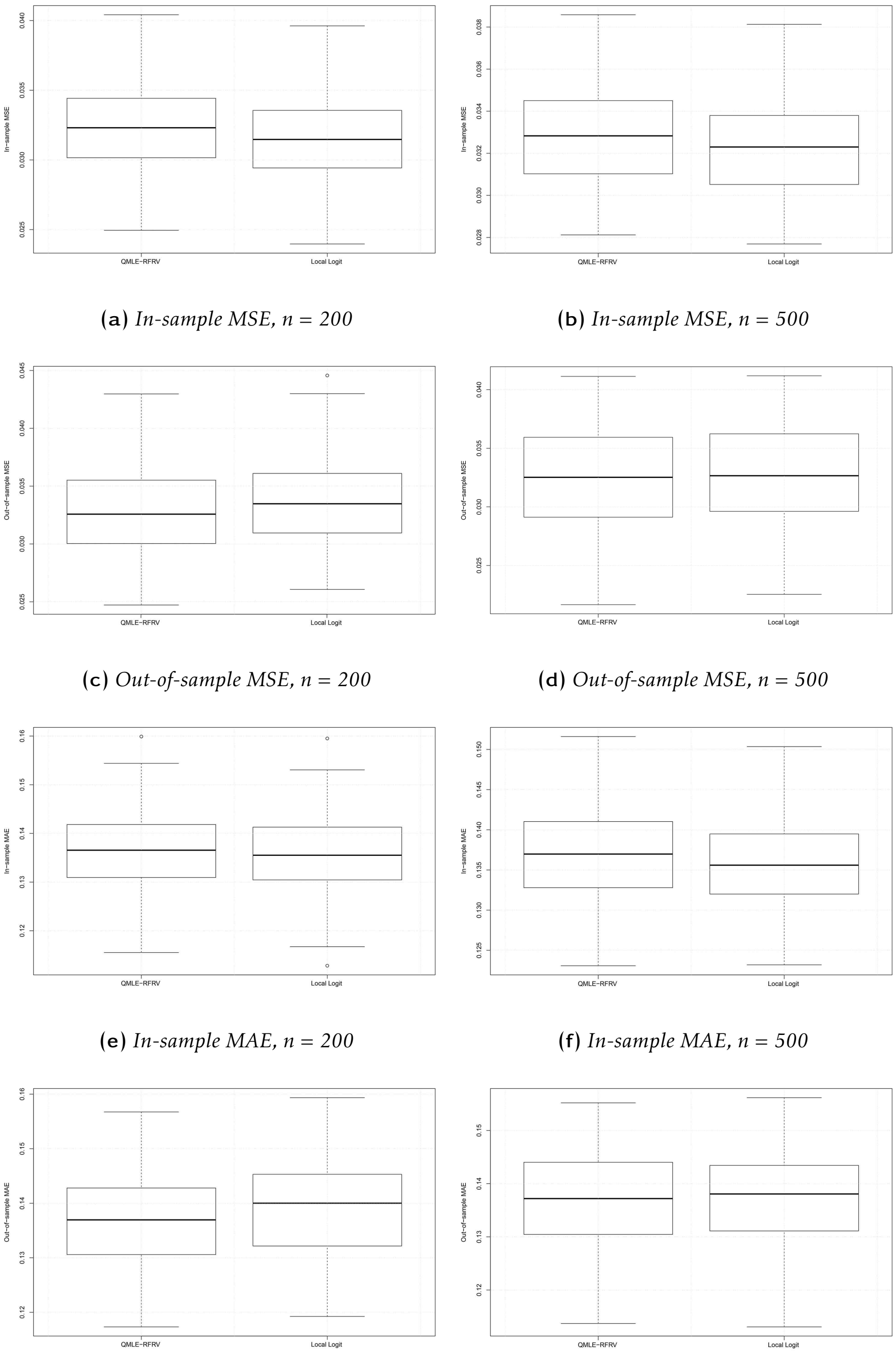
(f) *In-sample MAE, n = 500*

(g) *Out-of-sample MAE, n = 200*

(h) *Out-of-sample MAE, n = 500*

Figure 5.16: *Predictive performances under simulation U2*

Note: The figures compare the in-sample and out-of-sample predictive performances of QMLE-RFRV and local logit model for small (n = 200) and moderate (n = 500) sample sizes. Figures (a) to (d) report MSE, while Figures (e) to (f) report MAE.



(a) *In-sample MSE, n = 200*

(b) *In-sample MSE, n = 500*

(c) *Out-of-sample MSE, n = 200*

(d) *Out-of-sample MSE, n = 500*

(e) *In-sample MAE, n = 200*

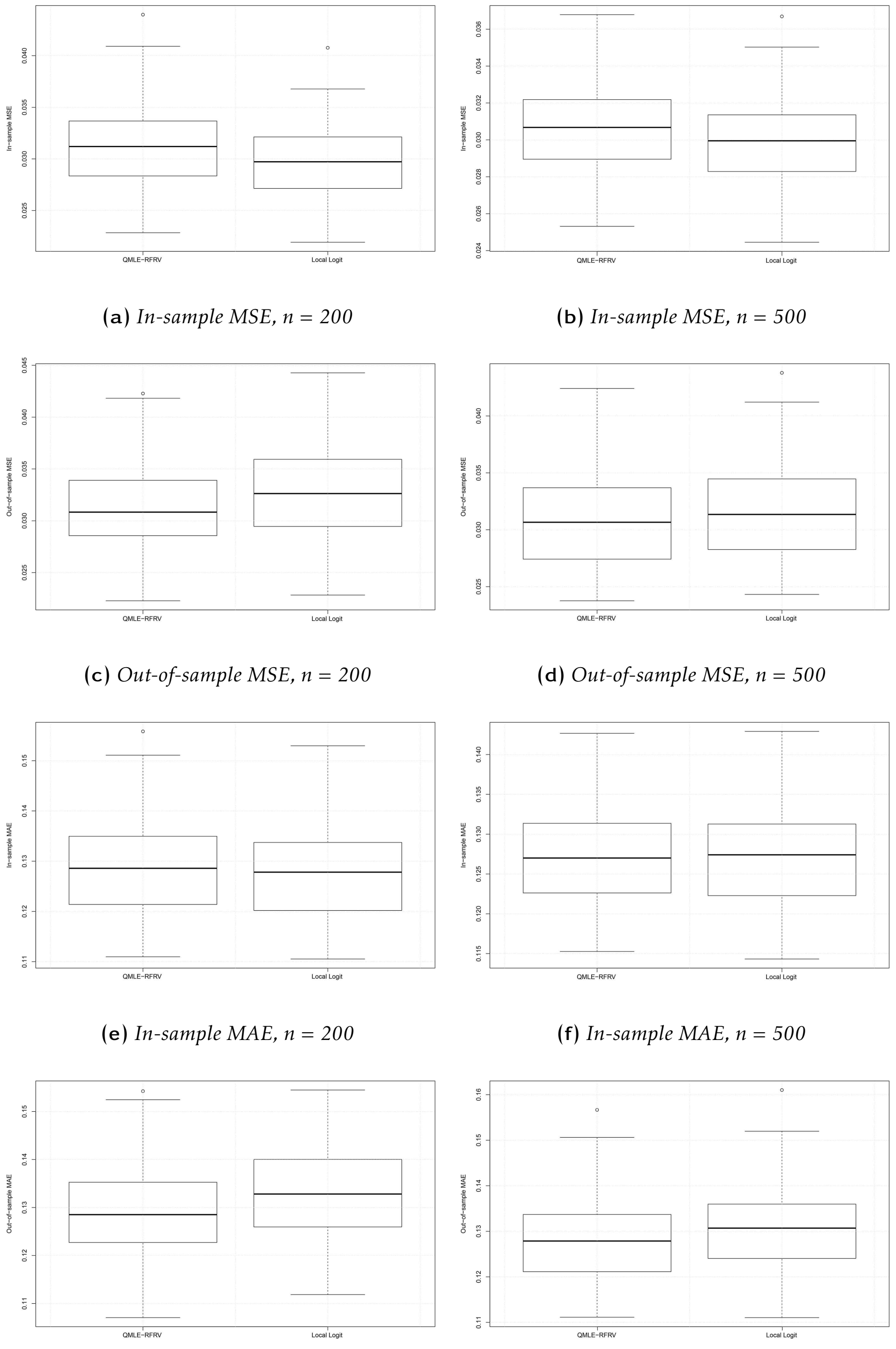
(f) *In-sample MAE, n = 500*

(g) *Out-of-sample MAE, n = 200*

(h) *Out-of-sample MAE, n = 500*

Figure 5.17: *Predictive performances under simulation U3*

Note: The figures compare the in-sample and out-of-sample predictive performances of QMLE-RFRV and local logit model for small (n = 200) and moderate (n = 500) sample sizes. Figures (a) to (d) report MSE, while Figures (e) to (h) report MAE.



(a) *In-sample MSE, n = 200*

(b) *In-sample MSE, n = 500*

(c) *Out-of-sample MSE, n = 200*

(d) *Out-of-sample MSE, n = 500*

(e) *In-sample MAE, n = 200*

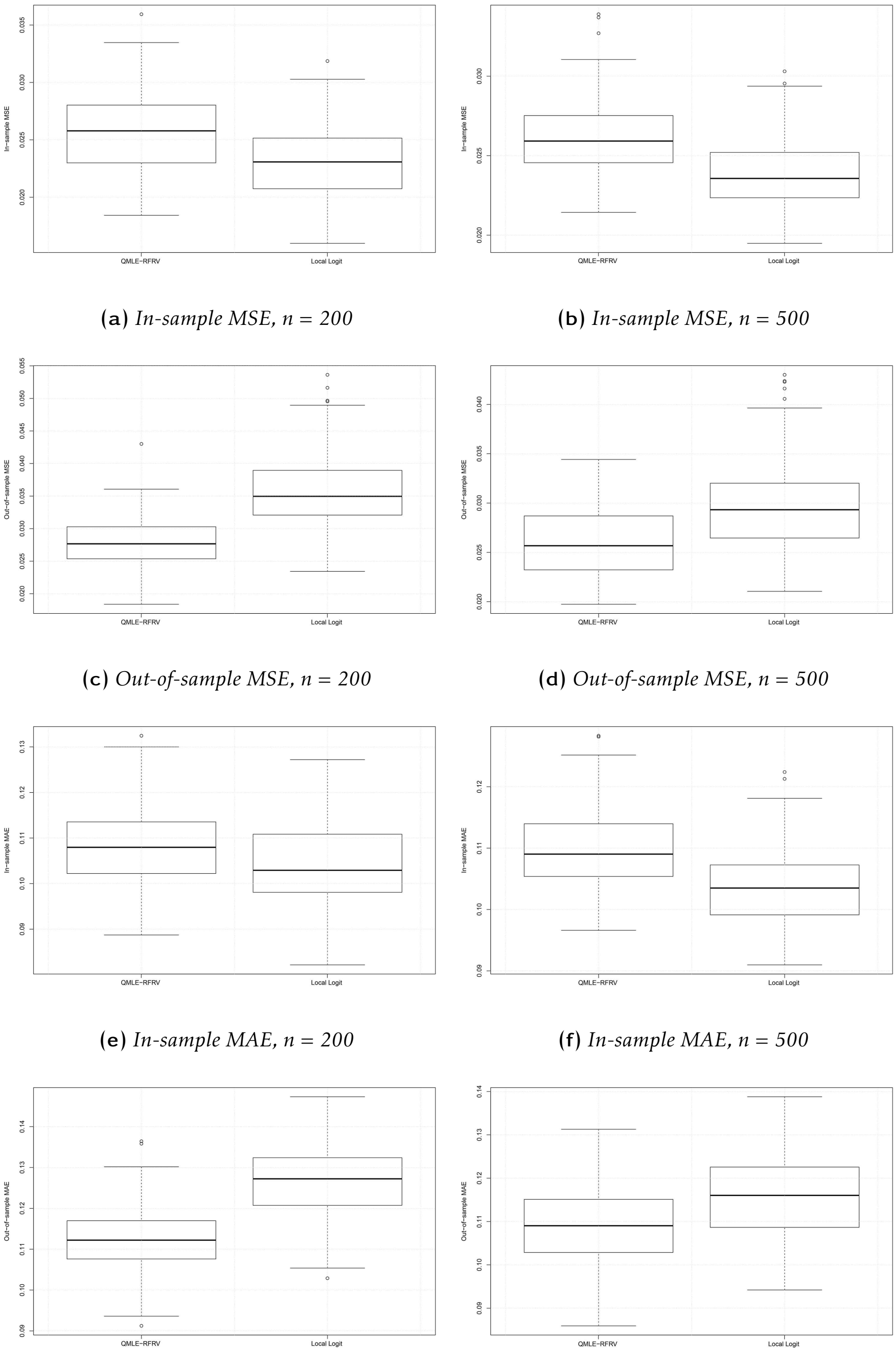
(f) *In-sample MAE, n = 500*

(g) *Out-of-sample MAE, n = 200*

(h) *Out-of-sample MAE, n = 500*

Figure 5.18: *Predictive performances under simulation B2*

Note: The figures compare the in-sample and out-of-sample predictive performances of QMLE-RFRV and local logit model for small (n = 200) and moderate (n = 500) sample sizes. Figures (a) to (d) report MSE, while Figures (e) to (h) report MAE.



(a) *In-sample MSE, n = 200*

(b) *In-sample MSE, n = 500*

(c) *Out-of-sample MSE, n = 200*

(d) *Out-of-sample MSE, n = 500*

(e) *In-sample MAE, n = 200*

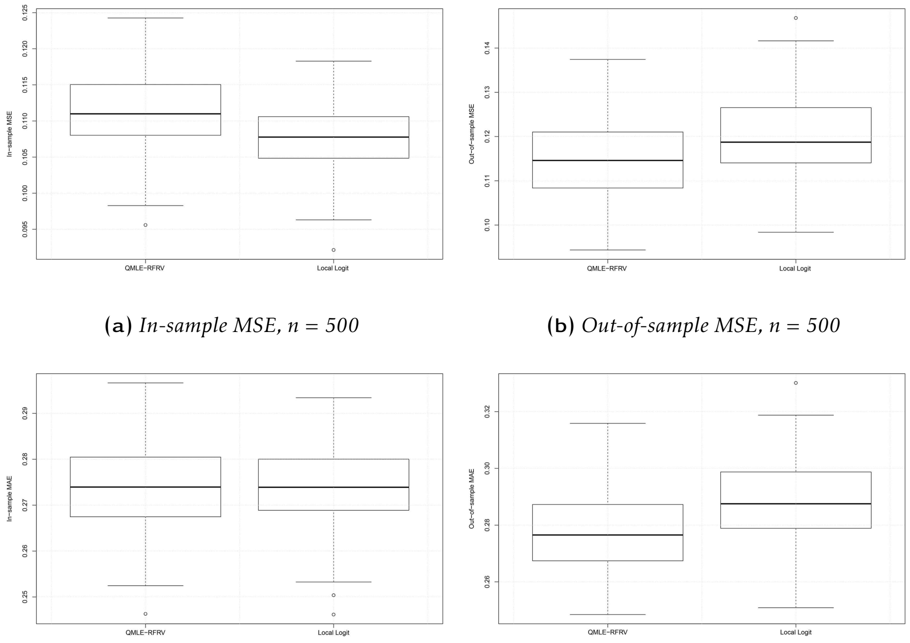
(f) *In-sample MAE, n = 500*

(g) *Out-of-sample MAE, n = 200*

(h) *Out-of-sample MAE, n = 500*

Figure 5.19: *Predictive performances under simulation B3*

Note: The figures compare the in-sample and out-of-sample predictive performances of QMLE-RFRV and local logit model for small (n = 200) and moderate (n = 500) sample sizes. Figures (a) to (d) report MSE, while Figures (e) to (h) report MAE.



(a) *In-sample MSE, n = 500*

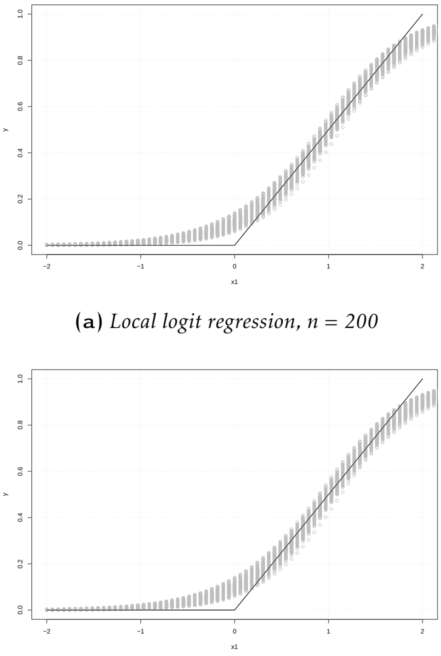
(b) *Out-of-sample MSE, n = 500*

(c) *In-sample MAE, n = 500*

(d) *Out-of-sample MAE, n = 500*

Figure 5.20: *Predictive performances under simulation M1*

Note: The figures compare the in-sample and out-of-sample predictive performances of QMLE-RFRV and local logit model for small (n = 200) and moderate (n = 500) sample sizes. Figures (a) to (d) report MSE, while Figures (e) to (h) report MAE.



(a) *Local logit regression, n = 200*

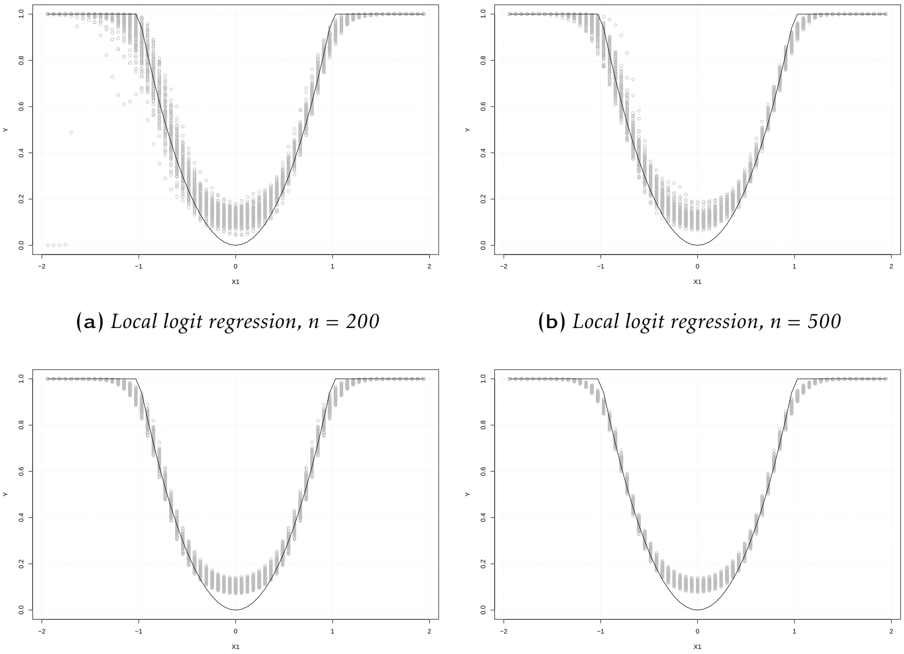
(b) *Local logit regression, n = 500*

(c) *QMLE-RFRV, n = 200*

(d) *QMLE-RFRV, n = 500*

Figure 5.21: *Nonlinear function estimates under simulation U1*

Note: The figures illustrate the nonlinear function estimate of the simulated bounded response variable under U1. The dark solid line is the true function, whereas the dotted plots are the estimates with two sample sizes as described below the sub-figures. Figures (a) and (b) are the estimates using the local logit model, and the remaining figures are those of the QMLE-RFRV with the correct functional form.



(a) *Local logit regression, n = 200*

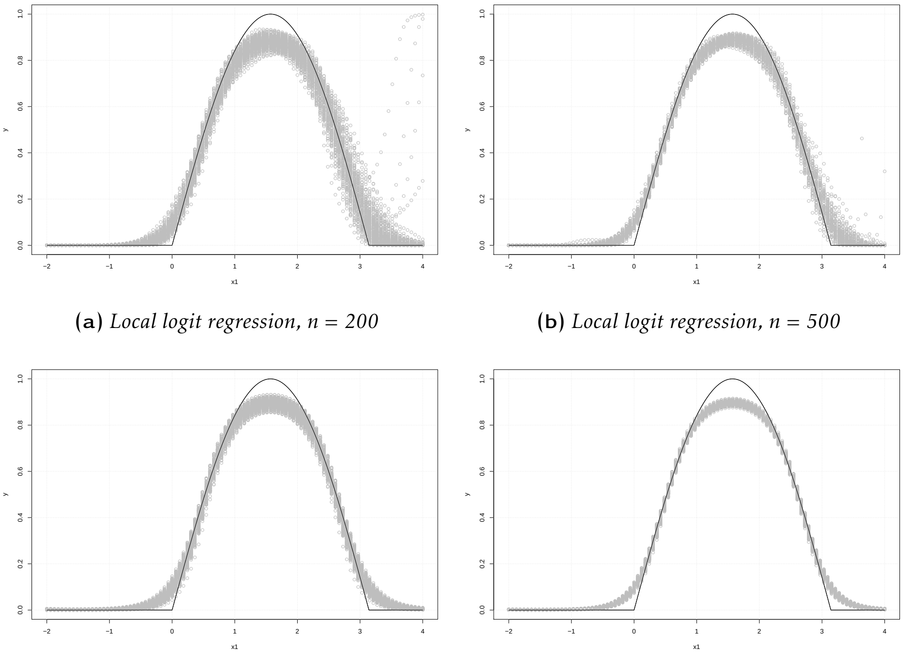
(b) *Local logit regression, n = 500*

(c) *QMLE-RFRV, n = 200*

(d) *QMLE-RFRV, n = 500*

Figure 5.22: *Nonlinear function estimates under simulation U2*

Note: Note: The figures illustrate the nonlinear function estimate of the simulated bounded response variable under U2. The dark solid line is the true function, whereas the dotted plots are the estimates with two sample sizes as described below the sub-figures. Figures (a) and (b) are the estimates using the local logit model, and the remaining figures are those of the QMLE-RFRV with the correct functional form.



(a) *Local logit regression, n = 200*

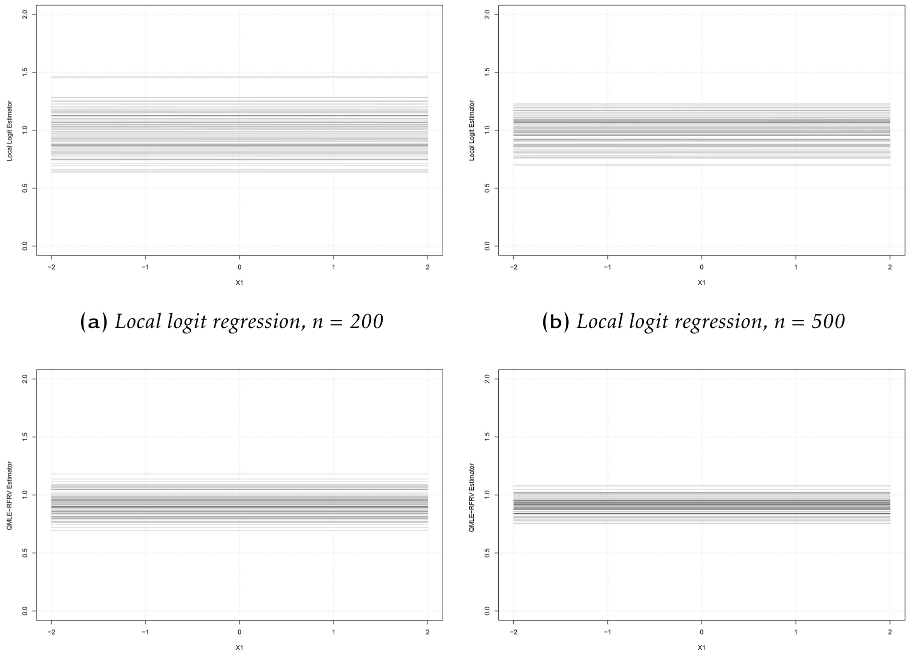
(b) *Local logit regression, n = 500*

(c) *QMLE-RFRV, n = 200*

(d) *QMLE-RFRV, n = 500*

Figure 5.23: *Nonlinear function estimates under simulation U3*

Note: The figures illustrate the nonlinear function estimate of the simulated bounded response variable under U3. The dark solid line is the true function, whereas the dotted plots are the estimates with two sample sizes as described below the sub-figures. Figures (a) and (b) are the estimates using the local logit model, and the remaining figures are those of the QMLE-RFRV with the correct functional form.



(a) *Local logit regression, n = 200*

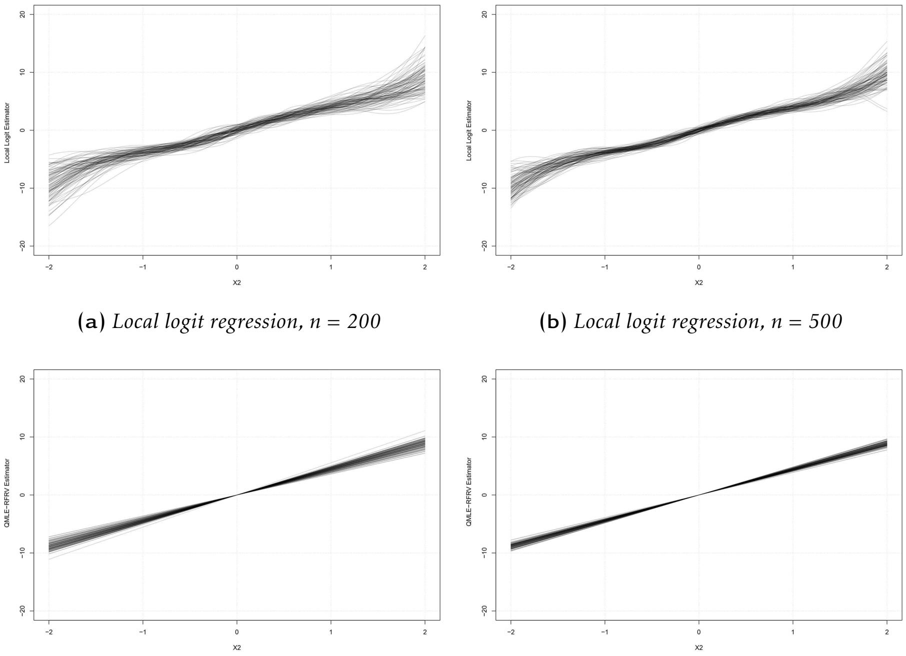
(b) *Local logit regression, n = 500*

(c) *QMLE-RFRV, n = 200*

(d) *QMLE-RFRV, n = 500*

Figure 5.24: *Marginal effect estimate of x_1 under simulation B2*

Note: The figures compare the marginal effect estimates of the proposed local logit model and QMLE-RFRV. (a) and (b) are the local logit marginal effect estimate $\beta_1(x)$ in (5.7.1) as a function of x_1 for small and moderate sample sizes, respectively. (c) and (d) represent the parametric QMLE-RFRV estimate γ_1 as the marginal effect estimate of x_1 in (5.7.2).



(a) *Local logit regression, n = 200*

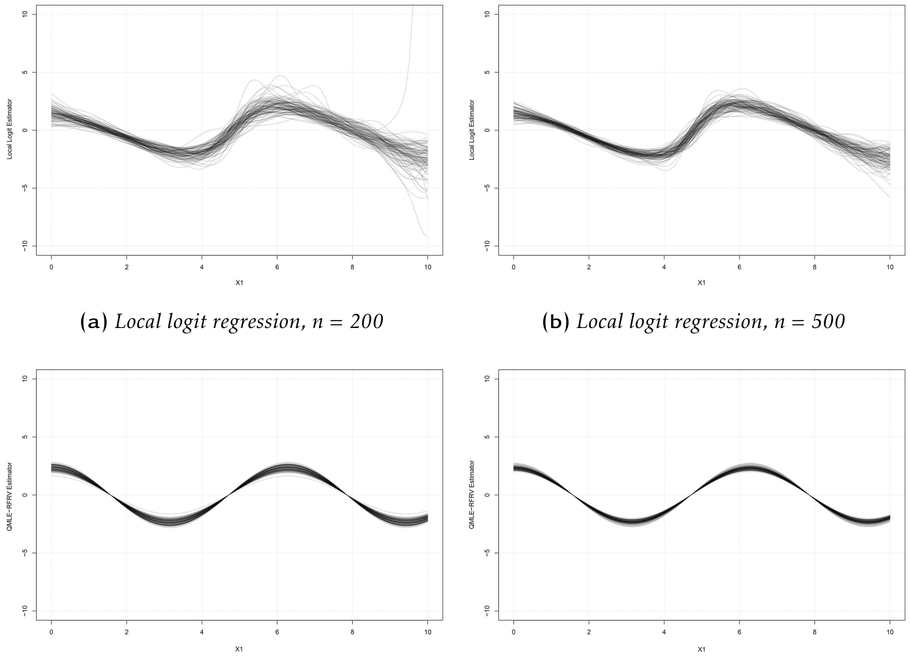
(b) *Local logit regression, n = 500*

(c) *Fractional regression, n = 200*

(d) *Fractional regression, n = 500*

Figure 5.25: *Marginal effect estimate of x_2 under simulation B2*

Note: The figures compare the marginal effect estimates of the proposed local logit model and QMLE-RFRV. (a) and (b) are the local logit marginal effect estimate $\beta_2(x)$ in (5.7.1) as a function of x_2 for small and moderate sample sizes, respectively. (c) and (d) represent the parametric QMLE-RFRV estimate $2\gamma_2 x_2$ as the marginal effect estimate of x_2 in (5.7.2).



(a) *Local logit regression, n = 200*

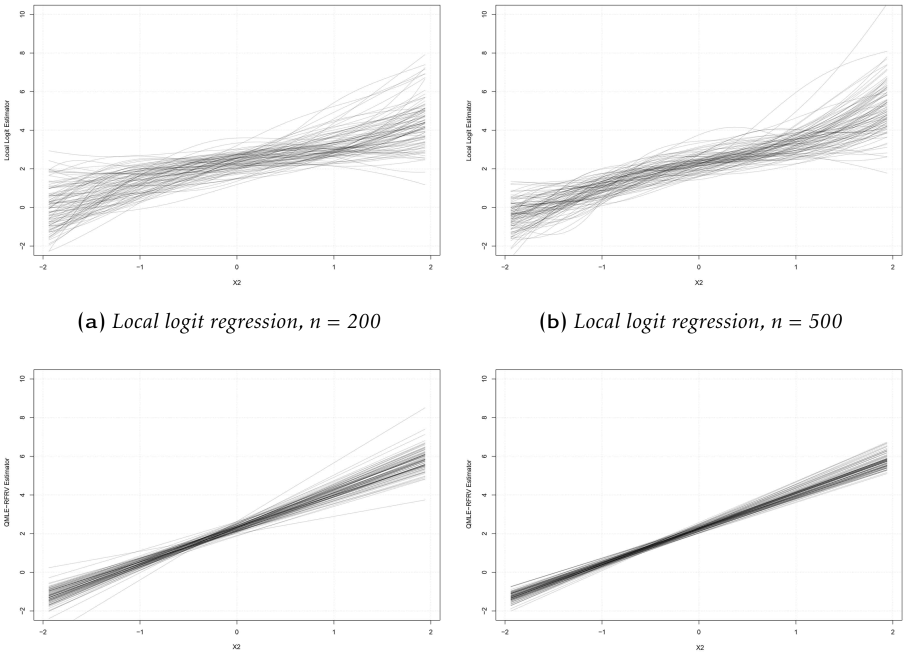
(b) *Local logit regression, n = 500*

(c) *QMLE-RFRV, n = 200*

(d) *QMLE-RFRV, n = 500*

Figure 5.26: *Nonlinear marginal effect estimate of x_1 under simulation B3*

Note: The figures compare the marginal effect estimates of the proposed local logit model and QMLE-RFRV. (a) and (b) are the local logit marginal effect estimate $\beta_1(x)$ in (5.7.1) as a function of x_1 for small and moderate sample sizes, respectively. (c) and (d) represent the parametric QMLE-RFRV estimate $\gamma_1 \cos(x_1)$ as the marginal effect estimate of x_1 in (5.7.3).



(a) *Local logit regression, $n = 200$*

(b) *Local logit regression, $n = 500$*

(c) *QMLE-RFRV, $n = 200$*

(d) *QMLE-RFRV, $n = 500$*

Figure 5.27: *Marginal effect estimate of x_2 under simulation B3*

Note: The figures compare the marginal effect estimates of the proposed local logit model and QMLE-RFRV. (a) and (b) are the local logit marginal effect estimate $\beta_2(x)$ in (5.7.1) as a function of x_2 for small and moderate sample sizes, respectively. (c) and (d) represent the parametric QMLE-RFRV estimate $2\gamma_2 x_2$ as the marginal effect estimate of x_2 in (5.7.3).

Chapter 6

Conclusion and future research direction

6.1 Contributions of the thesis

Quantifying and understanding credit default exposures are fundamentals of credit risk management. The RR is one of the key credit risk parameters. It indicates the risk of losing the amount invested due to debtors having defaulted. In recent decades, a number of studies have paid considerable attention to modelling the RR, mostly for the purpose of generating accurate predictions. A wide range of conventional to more sophisticated statistical models were proposed and are available as predictive models in the literature. This was due to several key empirical features of RRs that challenge applications of standard econometric models. First, RR is continuous, fractional, and bounded in [0,1]. Second, its empirical density is bimodal and asymmetric, with high proportions of recoveries at the boundaries zero and one. Third, in the presence of observations at 0 and 1, trimming and *transformation* as well as *back-transformation* of recoveries are needed for the use of valid statistical theory, despite such transformation introduces bias in the model estimates, resulting in unreliable statistical inference. Lastly, although there are some existing evidences of the nonlinearity in the recovery-covariate relationship, little attempt has been made in the literature to improve the specification of the RR functional form. This thesis proposes non- and semi-parametric regressions, which are data-driven approaches to accommodate the aforementioned properties. Our research aims to offer alternative perspectives and statistical models to analyse the RR-covariate relationship and to overcome the limitations of RR modelling in the literature.

This study makes several significant contributions to the central areas in the credit risk literature in its documentation of the merits of the applications of non- and semi-parametric regressions in RR modelling. The first contribution is that the proposed models do not rely on standard distribution assumptions or prespecified functional form. Thus, they mitigate the misspecification problem and inadequate parametric assumptions that arise in most parametric models. Furthermore, we develop a nonparametric local logit regression specifically for [0,1] RR data, which directly solves the boundary problem of this data.

The second contribution is that we provide a new direction for RR modelling by identifying and uncovering not only the nonlinear marginal effects in chapter 3, but also by showing how the effects vary across conditional quantiles in chapter 4. The third contribution is that, we improve the specification of RR modelling by incorporating the outcomes of the marginal and interaction effects analysis of nonparametric regression to improve and "calibrate" the functional form of semiparametric and parametric regressions in chapters 3 and 5, respectively, which have not been explored in the existing literature. Such an exploration leads to an improvement in the out-of-sample predictive accuracies of these "calibrated", semiparametric and parametric models. Comparing with fully nonparametric regression, these models are more convenient to estimate and interpret the results, and their out-of-sample predictive performances are comparable. These will be useful to applied researchers and industry professionals working in the risk management area who are unfamiliar with nonparametric machinery.

Chapter 2 provides a review of the literature in RR modelling. It began with discussions of the general role and the importance of the RR in credit risk management and the Basel accord. The stylised facts as well as empirical findings regarding specific features of the RR in practice were presented. This was followed by a detailed discussion of the recent developments in RR modelling. Three main approaches were presented: *transformation* regression, conventional regressions for bounded [0,1] response regressions, and data-driven approaches. The limitations of the methodology studied in the literature as well as a gap in it provided the motivation for this thesis, in which we attempt to overcome the limitations of existing approaches and address the practical applications of our proposed models in credit risk modelling. To this end, the fundamental concepts of the existing approaches and their limitations were discussed, which are mainly the misspecification problems in parametric models and the black-box problems in machine learning algorithms. In doing so, we identified several limitations of the RR-covariate model specifications, and estimation methods, which provided us motivations for the research endeavor of this thesis. In what follows, the outcomes of our research will become apparent.

Chapter 3 proposes local constant and local linear kernel estimation methods for nonparametric regressions as well as semiparametric partially linear regressions to predict and analyse RRs. In terms of the effect of recovery covariates, we consider both idiosyncratic and systematic variables, including debt cushion, type of loan, instrumental rank, collateral status, and economic stress index. The local linear method provides the marginal effect estimates of the covariates on RR, which illustrate the nonlinear effect of debt cushion and the approximately linear effect of the economic stress index. This information further enable us to specify the improved functional form for the partially linear regression, in which only debt cushion is assigned to the nonparametric component to capture its nonlinear effect. This study also compared the predictabilities of the proposed models and the existing alternative models such as QMLE regression for fractional response variable, two-sided censored Tobit model, inverse Gaussian linear regressions, mixture distribution, and the regression tree algorithm. The results show that the partially linear regression with the improved functional form outperforms others in the out-of-sample predictions, followed by the nonparametric regression with a local constant method. In addition, the nonparametric regression with local linear methods shows a high proportion of the predictions exceeding the [0,1] boundary. Hence, we proposed two-sided censoring regressions to ensure that the boundary requirement is met, but its predictive performance is not what was expected.

In chapter 4, we analyse the effect of the recovery covariates at the conditional quantiles of the RR distribution. To gain additional information, we applied a fully nonparametic quantile regression and a partially linear additive regression and estimated the marginal effects of the covariates on RR at the various conditional quantiles. Our study indicated the presence of heterogeneous effects. At the 0.25 quantile, the RR is less responsive to an increase in debt cushion at the lower levels compared with at the 0.5 and 0.75 quantiles, especially for unsecured bonds with rank 4, which are high-risk characteristic loans. On the other hand, the RR at the lower quantile decreases substantially with an increase in economic stress in comparison to the higher quantiles. This negative effect is prominent for low- and medium-risk characteristics loans. This result offers more complete picture of the downturn RR at various quantiles. Furthermore, our results also indicate that the proposed nonparametric quantile regression generates the most accurate RR Value-at-Risk prediction at the lower 0.25, 0.20, 0.15, 0.10, and 0.05 quantiles, followed the partially linear additive model. This suggests that the proposed models would be appropriate for estimating distressed downturn RR. Overall, the findings of this chapter could be used to enhance loan recovery strategies for defaulted loans depending on the quantile of the conditional RR distribution. Thus, banks and regulators could manage the credit risk exposure more efficiently by using this analytical framework.

Chapter 5 proposed a local logit regression for [0,1] bounded response variables which integrates the binary nonparametric regression introduced by Frölich (2006) with the parametric QMLE regression for fractional response variable (Papke & Wooldridge, 1996). In doing so, we introduce the nonparametric regression for fractional response variable, which, by its construction, fully addresses the boundary property of RR data. We conducted an extensive simulation study with various data-generating process such as complex nonlinear functions and asymmetric error distributional assumptions. The results show that the model is robust, and it can accurately estimate the nonlinear marginal and interaction effects. The proposed model was then applied on the empirical RR data for marginal effect analysis and prediction purposes.

The results show that the proposed model can capture the nonlinear effects of both debt cushion and stress index on RR. Furthermore, we also find the interaction effects between debt cushion and other loan characteristics, which include the type of loan and collateral status. The functional forms of the models studied in the previous chapters do not include some key interaction effects among covariates, which are essential ingredients for banks to design treatment programs for borrowers. These findings were later incorporated to calibrate the linear parametric QMLE regression for fractional response variable. In particular, the quadratic function and discretization of variables were employed to capture the nonlinear effect of the debt cushion and the stress index. We also improve parametric functional form to capture the interaction effects. To test the superiority of the "calibrated" model, we apply the specification testing as well as several out-of-sample predictive criteria, which they confirm that the parametric model with an improved functional form is comparable to the result of the proposed local logit model. Moreover, as far as the downturn RR is concerned, the out-of-sample predictive evaluation highlights that our models significantly outperform the conventional linear model in estimating RR during the recent financial crisis.

6.2 Future research direction

This thesis showed that applications of non- and semi-parametric regressions were highly useful for RR modelling, as they overcame several challenges addressed in the existing literature and provide new directions for RR analysis. Some aspects of the model can be further improved in the future research, which are outlined below.

In chapter 3, we have extended the one-sided censored nonparametric regression proposed by Lewbel and Linton (2002) to accommodate two-sided censoring, conducted a small scale simulation study, and applied the model to RR modelling with limited success. We will consider a further research investigation of the extended model in the future, which will involve derivations of some analytical properties, and a large scale simulation study of the model. The result of which will improve the understanding and practical applicability of the model.

Moreover, the univariate RR models extensively studied in this thesis and elsewhere can be further extended to panel data modelling. The empirical RR data used in this thesis is in fact cross-sectional data, but the time dimension can be added. The data has 3,742 defaulted loans over a 12-year window, where these loans are from 20 different industries. Hence, the data can be aggregated based on its industry, which will lead to a balanced RR panel data structure of 20 industries over time. Papke and Wooldridge (2008) proposed estimation method for a panel data with fractional response variables, which could be extended to the RR dataset. A number of non- and semi- parametric panel data models are also proposed by Henderson, Carroll, and Li (2008); Horowitz and Markatou (1996), although they are not specific to fractional response data. The outcomes of the RR panel data anaylsis, which has not been explored in the literature, would be undoubtedly add value to credit risk management analysis.

References

Acharya, V. V., Bharath, S. T., & Srinivasan, A. (2007). Does industry-wide distress affect defaulted firms? evidence from creditor recoveries. *Journal of Financial Economics*, *85*(3), 787–821.

Altman, E. I. (2006). Default recovery rates and lgd in credit risk modeling and practice: an updated review of the literature and empirical evidence. *New York University, Stern School of Business*.

Altman, E. I., & Kalotay, E. A. (2014). Ultimate recovery mixtures. *Journal of Banking & Finance*, *40*, 116–129.

Araujo, A., Kubler, F., & Schommer, S. (2012). Regulating collateral-requirements when markets are incomplete. *Journal of Economic Theory*, *147*(2), 450–476.

Arias, O., Hallock, K. F., & Sosa-Escudero, W. (2002). Individual heterogeneity in the returns to schooling: instrumental variables quantile regression using twins data. In *Economic applications of quantile regression* (pp. 7–40). Springer.

Basel committee on banking supervision. (2001). *Consultative document: The internal ratings-based approach*. Bank for International Settlements.

Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance*, *34*, 2510–2517.

Baur, D. G., Dimpfl, T., & Jung, R. C. (2012). Stock return autocorrelations revisited: A quantile regression approach. *Journal of Empirical Finance*, *19*(2), 254–265.

Bayes, C. L., & Valdivieso, L. (2016). A beta inflated mean regression model for fractional response variables. *Journal of Applied Statistics*, *43*(10), 1814–1830.

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, *28*, 171–182.

Bijak, K., & Thomas, L. C. (2015). Modelling lgd for unsecured retail loans using bayesian methods. *Journal of the Operational Research Society*, *66*(2), 342–352.

Billger, S. M., & Goel, R. K. (2009). Do existing corruption levels matter in controlling corruption?: Cross-country quantile regression estimates. *Journal of Development Economics*, *90*(2), 299–305.

BIS. (2004). *International convergence of capital measurement and capital standards: A revised framework*. Bank for International Settlements.

Bohn, J. R., & Stein, R. M. (2011). Loss given default. *Active Credit Portfolio Management in Practice*, 253–287.

Bonini, S., & Caivano, G. (2016). Estimating loss-given default through advanced credibility theory. *The European Journal of Finance*, *22*(13), 1351–1362.

Bottai, M., Cai, B., & McKeown, R. E. (2010). Logistic quantile regression for bounded outcomes. *Statistics in medicine*, *29*, 309–317.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Bruche, M., & González-Aguado, C. (2010). Recovery rates, default probabilities, and the credit cycle. *Journal of Banking & Finance*, *34*(4), 754–764.

Brunnermeier, M. K. (2009). Deciphering the liquidity and credit crunch 2007–2008. *The Journal of Economic Perspectives*, *23*(1), 77–100.

Buchinsky, M. (2002). *Quantile regression with sample selection: Estimating womens return to education in the us*. Springer.

Calabrese, R. (2012). Predicting bank loan recovery rates with a mixed continuous-discrete model. *Applied Stochastic Models in Business and Industry*, *30*, 99–114.

Calabrese, R. (2014). Downturn loss given default: Mixture distribution estimation. *European Journal of Operational Research*, *237*(1), 271–277.

Calabrese, R., & Zenga, M. (2010). Bank loan recovery rates: Measuring and nonparametric density estimation. *Journal of Banking & Finance*, *34*, 903–911.

Carey, M., & Gordy, M. (2004). Measuring systematic risk in recoveries on defaulted debt i: Firm-level ultimate lgds. *Federal Reserve Board, Washington*.

Chalupka, R., & Kopecsni, J. (2008). *Modelling bank loan lgd of corporate and sme segments: A case study* (Tech. Rep.). IES Working Paper.

Chellathurai, T. (2017). Probability density of recovery rate given default of a firms debt and its constituent tranches. *International Journal of Theoretical and Applied Finance*, *20*(04), 1750023.

De Gooijer, J. G., & Zerom, D. (2003). On additive conditional quantiles with high-dimensional covariates. *Journal of the American Statistical Association*, *98*(461), 135–146.

Dermine, J., & De Carvalho, C. N. (2006). Bank loan losses-given-default: A case study. *Journal of Banking & Finance*, *30*, 1219–1243.

De Servigny, A., & Renault, O. (2004). *Measuring and managing credit risk*. McGraw Hill Professional.

DiNardo, J., & Tobias, J. L. (2001). Nonparametric density and regression estimation. *Journal of Economic Perspectives*, *15*(4), 11–28.

Doksum, K., & Koo, J.-Y. (2000). On spline estimators and prediction intervals in nonparametric regression. *Computational Statistics & Data Analysis*, *35*(1), 67–82.

Dwyer, D., & Korablev, I. (2009). Moodys kmv losscalc v3. 0. *Moodys Analytics, manuscript*.

Fan, J., Zhang, C., & Zhang, J. (2001). Generalized likelihood ratio statistics and wilks phenomenon. *Annals of Statistics*, 153–193.

Fattouh, B., Scaramozzino, P., & Harris, L. (2005). Capital structure in south korea: a quantile regression approach. *Journal of Development Economics, 76*(1), 231–250.

Fenske, N., Kneib, T., & Hothorn, T. (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association, 106*(494), 494–510.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*, 799–815.

Fischer, M., Köstler, C., & Jakob, K. (2016). Modeling stochastic recovery rates and dependence between default rates and recovery rates within a generalized credit portfolio framework. *Journal of Statistical Theory and Practice, 10*(2), 342–356.

Fitzenberger, B., Koenker, R., & Machado, J. A. (2013). *Economic applications of quantile regression*. Springer Science & Business Media.

Fong, H. G. (2006). *The credit market handbook: advanced modeling issues* (Vol. 340). John Wiley & Sons.

Frölich, M. (2006). Non-parametric regression for binary dependent variables. *The Econometrics Journal, 9*(3), 511–540.

Frontczak, R., & Rostek, S. (2015). Modeling loss given default with stochastic collateral. *Economic Modelling, 44*, 162–170.

Frye, J. (2000). Depressing recoveries. *Risk, 13*(11), 108–111.

Gao, Q., Liu, L., & Racine, J. S. (2015). A partially linear kernel estimator for categorical data. *Econometric Reviews*, *34*(6-10), 959–978.

Gilson, S. C., John, K., & Lang, L. H. (1990). Troubled debt restructurings: An empirical study of private reorganization of firms in default. *Journal of Financial Economics*, *27*(2), 315–353.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Grunert, J., & Weber, M. (2009). Recovery rates of commercial lending: Empirical evidence for german companies. *Journal of Banking & Finance*, *33*(3), 505–513.

Gupton, G., & Stein, R. (2005). Losscalc v2: Dynamic prediction of losses-given-default modeling methodology. *Moodys KMV*.

Gürtler, M., & Hibbeln, M. (2013). Improvements in loss given default forecasts for bank loans. *Journal of Banking & Finance*, *37*, 2354–2366.

Hall, P., Racine, J., & Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*.

Han, C., & Jang, Y. (2013). Effects of debt collection practices on loss given default. *Journal of Banking & Finance*, *37*(1), 21–31.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, *143*(1), 29–36.

Härdle, W., Liang, H., & Gao, J. (2012). *Partially linear models*. Springer Science &

Business Media.

Hartmann-Wendels, T., Miller, P., & Töws, E. (2014). Loss given default for leasing: Parametric and nonparametric estimations. *Journal of Banking & Finance*, *40*, 364–375.

Henderson, D. J., Carroll, R. J., & Li, Q. (2008). Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics*, *144*(1), 257–275.

Hinloopen, J., Wagenvoort, R., & van Marrewijk, C. (2012). A k-sample homogeneity test: the harmonic weighted mass index. *International Econometric Review*, *4*(1), 17–39.

Hlawatsch, S., & Reichling, P. (2010). A framework for loss given default validation of retail portfolios. *The Journal of Risk Model Validation*, *4*(1), 23.

Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics* (Vol. 12). Springer.

Horowitz, J. L. (2012). *Semiparametric methods in econometrics* (Vol. 131). Springer Science & Business Media.

Horowitz, J. L., & Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, *100*, 1238–1249.

Horowitz, J. L., & Markatou, M. (1996). Semiparametric estimation of regression models for panel data. *The Review of Economic Studies*, *63*(1), 145–168.

Hoshino, T. (2014). Quantile regression estimation of partially linear additive models. *Journal of Nonparametric Statistics*, *26*, 509–536.

Hu, Y.-T., & Perraudin, W. (2002). The dependence of recovery rates and defaults. *Birkbeck College*.

Huang, X., & Oosterlee, C. W. (2011). Generalized beta regression models for random loss given default. *The Journal of Credit Risk*, *7*(4), 45.

Jacobs Jr, M., & Karagozoglu, A. K. (2011). Modeling ultimate loss given default on corporate debt. *The Journal of Fixed Income*, *21*, 6–20.

Jokivuolle, E., & Peura, S. (2003). Incorporating collateral value uncertainty in loss given default estimates and loan-to-value ratios. *European Financial Management*, *9*(3), 299–314.

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, *91*(433), 401–407.

Kalotay, E. A., & Altman, E. I. (2016). Intertemporal forecasts of defaulted bond recoveries and portfolio losses. *Review of Finance*, rfw028.

Khieu, H. D., Mullineaux, D. J., & Yi, H.-C. (2012). The determinants of bank loan recovery rates. *Journal of Banking & Finance*, *36*, 923–933.

Kieschnick, R., & McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical modelling*, *3*(3), 193–213.

Kliesen, K. L., Smith, D. C., et al. (2010). Measuring financial market stress. *Economic Synopses*.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, *46*, 33–50.

Koenker, R., & Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, *94*, 1296–1310.

Koop, G., Poirier, D. J., & Tobias, J. L. (2007). *Bayesian econometric methods*. Cambridge University Press.

Koopman, S. J., Lucas, A., & Schwaab, B. (2012). Dynamic factor models with macro, frailty, and industry effects for us default counts: the credit crisis of 2008. *Journal of Business & Economic Statistics*, *30*(4), 521–532.

Krüger, S., & Rösch, D. (2017). Downturn lgd modeling using quantile regression. *Journal of Banking & Finance*, *79*, 42–56.

Leow, M., & Mues, C. (2012). Predicting loss given default (lgd) for residential mortgage loans: A two-stage model and empirical evidence for uk bank data. *International Journal of Forecasting*, *28*, 183–195.

Lewbel, A., & Linton, O. (2002). Nonparametric censored and truncated regression. *Econometrica*, *70*(2), 765–779.

Li, P., Qi, M., Zhang, X., & Zhao, X. (2016). Further investigation of parametric loss given default modeling. *Economics Working Paper*.

Li, Q., Lin, J., & Racine, J. S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics*, *31*, 57–65.

Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Li, Q., & Racine, J. S. (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, *26*, 423–434.

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, *28*, 161–170.

Ma, L., & Pohlman, L. (2008). Return forecasts and optimal portfolio construction: a quantile regression approach. *The European Journal of Finance*, *14*(5), 409–425.

Meligkotsidou, L., Panopoulou, E., Vrontos, I. D., & Vrontos, S. D. (2014). A quantile regression approach to equity premium prediction. *Journal of Forecasting*, *33*(7), 558–576.

Miller, P., & Töws, E. (2017). Loss given default adjusted workout processes for leases. *Journal of Banking & Finance*.

Nazemi, A., Fatemipour, F., Heidenreich, K., & Fabozzi, F. J. (2017). Fuzzy decision fusion approach for loss-given-default modeling. *European Journal of Operational Research*.

Okada, K., & Samreth, S. (2012). The effect of foreign aid on corruption: A quantile regression approach. *Economics Letters*, *115*(2), 240–243.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics*, *11*(6), 619–632.

Papke, L. E., & Wooldridge, J. M. (2008). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics*, *145*(1), 121–133.

Peter, C. (2011). Estimating loss given default: Experience from banking practice. In *The basel ii risk parameters* (pp. 151–183). Springer.

Qi, M., & Yang, X. (2009). Loss given default of high loan-to-value residential mortgages. *Journal of Banking & Finance*, *33*, 788–799.

Qi, M., & Zhao, X. (2011). Comparison of modeling methods for loss given default. *Journal of Banking & Finance*, *35*, 2842–2855.

Racine, J., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*(1), 99–130.

Ramalho, E. A., Ramalho, J. J., & Murteira, J. M. (2011). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, *25*(1), 19–68.

Renault, O., & Scaillet, O. (2004). On the way to recovery: A nonparametric bias free estimation of recovery rate densities. *Journal of Banking & Finance, 28*(12), 2915–2931.

Resti, A. (2002). *The new basel capital accord: Structure possible changes and micro-and macroeconomic effects* (No. 30). Ceps.

Risk Management Group the Basel Committee on Banking Supervision. (1999). Principles for management of credit risk. *Basel Committee on Banking Supervision*.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Rösch, D., & Scheule, H. (2014). Forecasting probabilities of default and loss rates given default in the presence of selection. *Journal of the Operational Research Society, 65*(3), 393–407.

Saunders, A., & Allen, L. (2010). *Credit risk management in and out of the financial crisis: new approaches to value at risk and other paradigms* (Vol. 528). John Wiley & Sons.

Scandizzo, S. (2016). Loss given default models. *The Validation of Risk Models: A Handbook for Practitioners*, 78-92.

Schuermann, T. (2004). What do we know about loss given default?

Shalizi, C. (2013). *Advanced data analysis from an elementary point of view.* Citeseer.

Siao, J.-S., Hwang, R.-C., & Chu, C.-K. (2015). Predicting recovery rates using logistic quantile regression with bounded outcomes. *Quantitative Finance, 16*, 1–16.

Sigrist, F., & Stahel, W. A. (2011). Using the censored gamma distribution for modeling fractional response variables with an application to loss given default. *ASTIN Bulletin: The Journal of the IAA, 41*(2), 673–710.

Tang, Q., & Yuan, Z. (2013). Asymptotic analysis of the loss given default in the presence of multivariate regular variation. *North American Actuarial Journal, 17*(3), 253–271.

Tanoue, Y., Kawada, A., & Yamashita, S. (2017). Forecasting loss given default of bank loans with multi-stage model. *International Journal of Forecasting, 33*(2), 513–522.

Thomas, L. C., Matuszyk, A., So, M. C., Mues, C., & Moore, A. (2016). Modelling repayment patterns in the collections process for unsecured consumer debt: A case study. *European Journal of Operational Research, 249*(2), 476–486.

Tibshirani, R., & Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association, 82*(398), 559–567.

Tobback, E., Martens, D., Van Gestel, T., & Baesens, B. (2014). Forecasting loss given default models: impact of account characteristics and the macroeconomic state. *Journal of the Operational Research Society, 65*(3), 376–392.

Tong, E. N., Mues, C., & Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting, 29*, 548–562.

Wang, J., & Yang, L. (2009). Efficient and fast spline-backfitted kernel smoothing of additive models. *Annals of the Institute of Statistical Mathematics*, *61*, 663–690.

Wei, L., & Yuan, Z. (2016). The loss given default of a low-default portfolio with weak contagion. *Insurance: Mathematics and Economics*, *66*, 113–123.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Yang, B. H., & Tkachenko, M. (2012). Modeling exposure at default and loss given default: empirical approaches and technical implementation. *The Journal of Credit Risk*, *8*(2), 81.

Yao, X., Crook, J., & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research*, *240*, 528–538.

Yuan, Z. (2016). An asymptotic characterization of hidden tail credit risk with actuarial applications. *European Actuarial Journal*, 1–28.

Zambom, A. Z., & Dias, R. (2012). A review of kernel density estimation with applications to econometrics. *arXiv preprint arXiv:1212.2812*.

Zhang, J., & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling lgd. *International Journal of Forecasting*, *28*, 204–215.

Zietz, J., Zietz, E. N., & Sirmans, G. S. (2008). Determinants of house prices: a quantile regression approach. *The Journal of Real Estate Finance and Economics*, *37*(4), 317–333.