COPYRIGHT AND CITATION CONSIDERATIONS FOR THIS THESIS/ DISSERTATION

o Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

o NonCommercial — You may not use the material for commercial purposes.

o ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

How to cite this thesis
THE APPLICATION OF REDUCED-FORM MODELS FOR MANAGING CONSUMER CREDIT RISK

by

FREDERIK CHRISTOFFEL VAN DER WALT

THESIS

submitted in fulfilment of the requirements for the degree

PHILOSOPHIAE DOCTOR

in

MATHEMATICAL STATISTICS

in the

FACULTY OF SCIENCE

at the

UNIVERSITY OF JOHANNESBURG

SUPERVISOR: PROF. F. LOMBARD

2014
AFFIDAVIT: MASTER AND DOCTORAL STUDENTS

TO WHOM IT MAY CONCERN

This serves to confirm that I, Frederik Christoffel van der Walt (ID number 800530 5154 081, student number 200837974) enrolled for the qualification Philosophiae Doctor in Mathematical Statistics in the Faculty of Science, herewith declare that my academic work is in line with the Plagiarism Policy of the University of Johannesburg with which I am familiar.

I further declare that the work presented in the thesis is authentic and original unless clearly indicated otherwise and in such instances full reference to the source is acknowledged and I do not pretend to receive any credit for such acknowledged quotations, and that there is no copyright infringement in my work. I declare that no unethical research practices were used or material gained through dishonesty. I understand that plagiarism is a serious offence and that should I contravene the Plagiarism Policy notwithstanding signing this affidavit, I may be found guilty of a serious criminal offence (perjury) that would amongst other consequences compel the University of Johannesburg to inform all other tertiary institutions of the offence and to issue a corresponding certificate of reprehensible academic conduct to whomever requests such a certificate from the institution.

Signed at Johannesburg on this fifth day of September 2014.

Signature  ________________________________________
Print name  ________________________________________

STAMP: COMMISSIONER OF OATHS

Affidavit certified by a Commissioner of Oaths

This affidavit conforms with the requirements of the JUSTICES OF THE PEACE AND COMMISSIONERS OF OATHS ACT 16 OF 1963 and the applicable Regulations published in the GG GNR 1258 of 21 July 1972; GN 903 of 10 July 1998; GN 109 of 2 February 2001 as amended.
Abstract

This thesis considers the modelling and prediction of consumer credit risk events. We model consumer credit risk events (like a missed payment, a repayment or a default) by means of a discrete, real time, staggered entry counting process.

Merton’s (1974) structural approach is the foundation of numerous credit-risk models, as well as the Basel capital accords. The underlying assumptions of this approach are that both liability and asset levels are observable to some extent and that default, which occurs if liability levels are larger than asset levels, can occur only once. These assumptions are inappropriate for consumer credit risk, where asset and liability levels are not observable and multiple defaults may occur. We find that the so-called reduced-form models initially developed by Artzner and Delbaen (1995) and Jarrow and Turnbull (1995), which impose no structure on the default event, are better suited to model and predict consumer credit risk.

All reduced-form models can be represented as counting processes. Counting processes are continuous in nature, so we discretize these processes before applying them to the regularly spaced, discrete monthly data. We show that the use of survival analysis techniques such as Cox’s (1972) proportional hazard model, which is a special case in counting processes, are not well suited to model credit risk. This is because survival analysis is mostly concerned with the prediction of the time until a single event occurs. Accordingly, in survival analysis the time domain used is ‘event time’. Hence, all observations need to be aligned to some starting time. We prefer to work in ‘calendar time’ and are concerned with the timing (in calendar time) of multiple events.

We identify and implement a dynamic, discrete statistical model based on calendar time that accounts for staggered entries, multiple entries into and exits from the portfolio, as well as multiple default events on an account level. This approach, from Arjas and Haara (1987), makes use of both idiosyncratic and systematic covariates, which facilitates stress-testing. This approach has, to our knowledge, never been applied to credit risk before and we apply it to a mortgage loan portfolio of a major bank in South Africa. We confirm the theoretical asymptotic properties of the estimated parameters in this model empirically via simulation.

The current delinquency level contributes greatly to default events, so we develop a
model to predict these delinquency levels. We predicted the delinquency level by combining a ‘failure’ counting process that predicts a missed payment with a ‘recovery’ counting process that predicts the size of the repayment made. We compare this approach to the Markov chain approach by Grimshaw and Alexander (2010). We modify their approach to include static coefficients and dynamic covariates instead of dynamic coefficients. Both the combined counting process and Markov chain approach are dependent on idiosyncratic and systematic covariates that may be static and/or dynamic.

Throughout this study we use various diagnostic tests. We use marginal-model plots (Cook & Weisberg 1997) on cross-sectional data that we have modified to include a bootstrapped confidence band. We contribute a composite marginal-model plot to perform heuristic diagnostic tests on longitudinal type data.
Acknowledgments

I do not think I fully comprehended what I was getting myself into when I started this endeavor. I now have a deep appreciation for what Nelson Mandela said: “It always seems impossible until it is done”. Additionally, I also now fully comprehend what the difference is between success (as people think it is) and success (as it really is). This is illustrated vividly by the figure below.

I would like to acknowledge that this Ph.D. could not have been done without the following people, each of whom in his/her own way (whether he/she liked it or not) played a part in the successful completion of this thesis.

To my supervisor, Professor Freek Lombard (affectionately referring to himself as "ouFreek" and to me as "jongFreek"): You are the best teacher I have ever had. Thank you for your dedication, hard work, wisdom and enthusiasm. I always feel inspired after being in contact with you. I have a saying that goes: you need to know so much to be of so little use. Keeping this in mind, Professor Lombard is of tremendous use. You are truly a great inspiration and example, you are my hero. Thank you also to your wife Chrystal for the support, broodjies and kindness.

To my peers who inspired me by example: Wickes Robbertse, Herrie van Rooy and Gideon Griebenow. I guess I am now also part of the club . . .

To many other family, friends and colleagues who all continually enquired after and
supported me (special mention to Pierre Venter).

To my dance partner, then girlfriend, then betrothed, then wife and now the mother of our first-born (all during the course of this drawn-out Ph.D.): Thank you Charlotte for your love and support. I am looking forward to our lives together p-Ph.D. (post-Ph.D.). You will not be a Ph.D. widow again! To my parent-in-law: thank you for supporting us, and especially Charlotte when I couldn’t.

To my mother, Coreen, and father, Sarel: Thank you for your continued support, encouragement and understanding. Thank you for always being there, especially when the going started to get tough. Thank you for being good parents and setting a good example. I am eternally grateful to you and would not have completed this endeavour (Ph.D. and life) without your wisdom, insight and support. You are anchor points in my life. To my brother, Petrus, and sisters, Maria and Hester: I had a great experience growing up with all of you, let’s hope growing old together is also fun. To my whole family: you all have a special place in my heart, I love you very much.

To God in heaven for talent, opportunities and time. None of this could have been possible without Him. The following Psalm aptly summarises my feelings: Psalm 16 v. 5,6: "Here, U is my lewe, U sorg vir my. Wat ek ontvang, kom alles van U af. 'n Pragtige deel is vir my afgemeet, ja, wat ek ontvang het, is vir my mooi."

Fredrik van der Walt
Contents

1 Introduction: credit risk models 4
   1.1 Structural approach . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
   1.2 Reduced-form approach . . . . . . . . . . . . . . . . . . . . . . . . . . 8
   1.3 A common framework . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
   1.4 Consumer credit risk and model requirements . . . . . . . . . . . . . . 11
   1.5 Layout of the thesis . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

2 The data 19
   2.1 Summary statistics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
   2.2 Covariates . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
      2.2.1 Idiosyncratic static covariates . . . . . . . . . . . . . . . . . . . . 23
      2.2.2 Idiosyncratic dynamic covariates . . . . . . . . . . . . . . . . . . 25
      2.2.3 Systematic covariates . . . . . . . . . . . . . . . . . . . . . . . . 26
   2.3 Appendix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30
      2.3.1 Derivation of $CLTV$ . . . . . . . . . . . . . . . . . . . . . . . . . 30

3 Modelling static credit events: classification techniques for cross-sectional data 33
   3.1 Classification: an overview . . . . . . . . . . . . . . . . . . . . . . . . . 34
   3.2 Generalized linear models . . . . . . . . . . . . . . . . . . . . . . . . . . 38
   3.3 Variable selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45
   3.4 Model checking . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 47
      3.4.1 Goodness-of-fit tests . . . . . . . . . . . . . . . . . . . . . . . . . . 49
      3.4.2 Diagnostic tests . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
      3.4.3 Marginal-model plots . . . . . . . . . . . . . . . . . . . . . . . . . . 57
3.4.4 Power of discrimination- ROC analysis ........................................ 73
3.5 Appendix ....................................................................................... 80
3.5.1 Generalized cross-validation technique for estimating the lasso tuning parameter in logistic regression ........................................ 80

4 Modelling dynamic credit events: counting processes 85
  4.1 Survival analysis ........................................................................ 86
  4.2 Counting processes: preliminaries ............................................. 89
  4.3 Survival analysis via counting processes .................................... 91
  4.4 Counting-process model requirements ...................................... 95
    4.4.1 Calendar time and delayed entry ......................................... 95
    4.4.2 Discrete-time regression model .......................................... 98

5 The models: theory 102
  5.1 The Shumway and DTYG models ............................................ 102
  5.2 The Arjas and Haara real-time logistic regression model ............. 107
  5.3 Asymptotic results for the AH model: a simulation study .......... 115
    5.3.1 Normality of the estimated parameters ................................ 119
    5.3.2 The appropriateness of the information matrix ..................... 120
    5.3.3 Determining the normalizing term $c_t$ ................................ 120
  5.4 Appendix .................................................................................... 125
    5.4.1 Various hypothesis tests .................................................... 125

6 The models: application with terminal default events 128
  6.1 Fitting the AH model to data ................................................... 128
  6.2 Fitting the DTYG model to data .............................................. 136
  6.3 Diagnostics ................................................................................ 141
    6.3.1 Prediction validation ......................................................... 141
    6.3.2 Model discrimination ....................................................... 144
  6.4 Appendix .................................................................................... 146
    6.4.1 Computation of the partial derivative $\frac{\partial}{\partial \beta} [\pi_{i,t,j}]$ ......... 146
    6.4.2 Maximizing the likelihood in the DTYG model .................... 147
7 A failure model: the AH model applied to our data
  7.1 The first-order model ........................................... 153
  7.2 Diagnostic tests on the initial model .......................... 154
  7.3 The modified model .............................................. 159
  7.4 Diagnostic tests on the modified model and model comparison 161

8 Modelling delinquencies ............................................. 169
  8.1 Combined counting processes to model delinquency .......... 170
    8.1.1 A truncated distribution approach .......................... 172
    8.1.2 Modelling repayments: a real-time AH-type truncated geometric
          model .............................................................. 174
    8.1.3 Parameter estimation .......................................... 178
  8.2 Markov chains to model delinquency ............................ 179
    8.2.1 Stationary transition matrices ................................. 180
    8.2.2 The Grimshaw and Alexander approach ....................... 182
  8.3 Model prediction comparison ..................................... 187
  8.4 Appendix ............................................................ 191
    8.4.1 The truncated geometric distribution ......................... 191
    8.4.2 Parameter estimation of the truncated geometric distribution 192

9 Summary and suggestions for further research .................. 194

10 List of abbreviations, covariates, notation and definitions ... 197
Chapter 1

Introduction: credit risk models

Credit risk is the potential financial loss that financial institutions may incur owing to unexpected changes in the credit quality of obligors. Changes in the credit quality can be, to name a few examples, a decrease in the credit rating of an obligor, the non-payment of an installment that is due or the default of an obligor. In essence, a financial institution uses credit-risk models to quantify credit risk. An estimation of the distribution of credit losses (the loss distribution) is driven by many factors, the most prominent of which are the probability of default, the resulting monetary losses incurred by the financial institution and default contagion. Default contagion is when the defaults of obligors are dependent on one another. To a large extent developments in credit-risk modelling have been spurred by regulatory requirements, such as the Basel capital accords. The academic literature on credit-risk modelling is extensive and diverse.

Credit-risk models are either used for credit-risk management or for the analysis of credit-risky securities. In this study we focus on credit-risk management. Credit-risk management is mostly concerned with estimating a loss distribution, computing risk measures, stress-testing portfolios and allocating provisional capital. A loss distribution facilitates the calculation of risk measures such as the value at-risk (VaR) or the expected shortfall (ES). In turn, these risk measures are used to calculate capital allocation, price loans and actively manage risk. Stress-testing is the activity of determining the robustness of credit-risk measurements by using specific hypothesized macro-economic scenarios to determine their impact on the risk measures. Typically a stress test would answer questions such as what effect a 100-basis-point increase in interest rates would
have on the default rate of a certain credit-risky portfolio. The whole credit management
process could be applied to corporate entities, small and medium enterprises (SMEs) and
individuals (also known as consumer/retail risk).

Most credit-risk models have two central tasks: to predict and explain the probability
of a default occurring and to estimate the severity of the resulting losses. In this study
we focus on the estimation of the probability of default. More specifically, we emphasize
the modelling of the occurrence of what we term ‘credit events’. A credit event can be,
among other things, a default or an increase or decrease in the current delinquency level
of an obligor. A credit-risk model is a model that estimates the probabilities of credit
events occurring.

Credit-risk models can be divided into several classes. A model can either be a two-
state default-no-default model or a multi-state credit rating migration model. Credit-risk
models are also classified as either static or dynamic models, and as either structural or
reduced-form models.

Default-no-default models are concerned only with the estimation of a default credit
event. Rating migration models extend this binary approach by modelling multiple credit
classes or ‘ratings’. Examples of credit ratings would be those published by Moody’s
KMV, Fitch or Standard and Poor’s.

Static models predict the loss distribution at a fixed point in time and are typically
used for credit-risk management. Extending this approach, dynamic models predict de-
defaults more or less continuously over time. The latter models have particularly found
favour in the field of credit derivative modelling, since the timing of the default is crucial
to the pricing of credit derivatives.

By far the greatest distinction between models is whether a model is a structural
or reduced-form model. Structural models are also known as option-theoretical mod-
els, value-of-firm models or threshold models. Reduced-form models are also known as
intensity-based, actuarial or hazard rate models. In this chapter we will explore the basic
assumptions underlying each of these model types. This chapter will not by any means
constitute a full account of all the credit-risk models and approaches currently available.
We simply highlight some important and influential models as a background. Although
we use a reduced-form model to predict the probability of credit events, a study of credit-
risk management would be incomplete without the mention of the structural approach. In the next section we discuss the structural approach to credit-risk modelling. We then discuss the reduced-form of credit-risk modelling in Section 1.2 with a comparison of the two approaches in Section 1.3. After the introduction to credit-risk models we conclude this chapter with the layout of this thesis.

1.1 Structural approach

The structural approach to credit-risk modelling was spurred by Merton’s (1974) seminal paper on defaultable bond pricing. This approach, often referred to as the ‘Merton model’, attempts to predict default events by modelling an obligor’s assets and liabilities. Here, the underlying assumption is that an obligor will default if the value of its assets falls below its liability level. Hence, as the name implies, some structure is given to the default event. Merton’s original model was developed strictly for corporate (usually listed) firms. Structural models make use of option pricing theory, where the liability level can be seen as the strike price and the underlying instrument is the asset value. Therefore, the financial institution’s liability is an option on the assets of the obligor. In the classical Merton-type setup, the assets and liabilities of an obligor are assumed to be known and default can only occur at the end of a certain time horizon, as in the case of European options. Asset values are assumed to follow a geometric Brownian motion and the volatility of the obligor’s asset value is assumed to be known. Liability levels are assumed to be fixed over the whole time span of the loan. In Figure 1-1 we depict three realizations of the geometric Brownian motion that represents the asset value of a specific firm. The liability level is fixed at 98. At the end of the time horizon $t = 1$, we see that scenario 1 is a case of default, as the asset value is smaller than the liability level, but not the other two scenarios. In scenario 2 there was never a default, but in scenario 3 there was a default (at $t = 0.1$) and then a recovery (at $t = 0.85$). The classical Merton model regards scenario 3 as a non-default since it was not in default at $t = 1$. If, for instance, the time horizon were to be $t = 0.7$ (where the vertical line is), then we would classify scenario 3 as a default and not the other two scenarios.

Merton’s pragmatic and intuitive approach is popular even today, when the deficiencies of the model are well known.
These deficiencies are, among other things, that: default is only allowed at the end of the time period, asset values and liability levels are mostly unobservable, obligors have fixed liability levels over the whole duration and the asset values are driven by geometric Brownian motion. Several credit-risk models still use Merton’s approach as their foundation. Merton’s model spurred the development of models that relax some of its more stringent, and often unrealistic, assumptions. The best-known extensions are the first-passage-time model and the excursion approach.

First-passage-time models attempt to relax the assumption of the Merton model that default may only occur at maturity. In reality, default may occur at any time. This can clearly be seen in Figure 1-1: scenario 3 defaulted before expiry. In first-passage-time models, default is recorded at the first time that the asset value falls below the liability level. In terms of the language of options, this approach would be similar to pricing a barrier option. First-passage models paved the way for the valuation of debt instruments that are sensitive to default timing. The first-passage-time credit-risk model is, of course, a dynamic model. In the excursion approach, default does not take place at the exact time when the asset value falls below the liability level. Here, the asset value
must remain below the liability level for a predefined duration. In practice a financial
institution often extends due dates to enable an obligor to reorganize its structures or
finances. Consequently, obligors will eventually either recover or default. The total
excursion time may be defined either as the sum of all previous excursions or as the
length of the current excursion. In option pricing theory, these respective excursion
models are akin to the so-called Parasian and Parisian options (Giesecke 2004). Several
other extensions to the classical Merton model exist. Some of these models incorporate
tax effects and bankruptcy costs, as well as recapitalization and/or renegotiation effects
(Giesecke 2004).

Several ‘off-the-shelf’ models purchased by financial institutions employ the structural
 approach. Among these are Portfolio Manager by the KMV corporation, CreditMetrics\textsuperscript{TM}
Portfolio Manager, which is a ratings based model, extends Merton’s model by estimating
the unobservable asset level as a function of publicly traded equity prices. Although the
exact methodology is proprietary, a good indication of the model’s technical detail can
be found in Crouhy et al. (2000). CreditMetrics\textsuperscript{TM} is also a ratings-based approach
and makes use of credit-migration matrices (Gupton et al. 1997). CreditPortfolioView,
often referred to as an econometric model, makes use of a logistic regression model that
incorporates macro-economic factors instead of the latent factors that are often assumed
to drive credit risk (Crouhy et al. 2000 and Wilson 1997a, b).

To summarize, the structural approach to credit-risk modelling assumes some degree
of structure to explain and predict default events. The postulated structure is a function
of the true asset value and liability level of an obligor. The implementation of structural
models is hampered by the unobservability of the asset value and liability level. For this
reason, reduced-form models have found favour as of late.

1.2 Reduced-form approach

In contrast to the structural approach, in the reduced-form approach a default event
is simply modelled as a function of some covariates. These models were pioneered by
Artzner and Delbaen (1995) and Jarrow and Turnbull (1995). Reduced-form models do
not attempt to explain why a firm defaults, i.e. no default structure or default mecha-
nism is postulated. Instead, default is defined and default events are modelled by using regression-type models. Reduced-form models are typically formulated in terms of counting processes. The occurrence of a default results in a jump of the counting process. Counting processes are driven by some intensity function, which is generally a stochastic process and/or a function of some covariates. The challenge faced in building reduced-form models is to identify an appropriate intensity function, the driving covariates and checking the model’s goodness of fit.

Reduced-form models may also be default-no-default or rating migration models. Sometimes, an intensity matrix is assumed to govern the migration rate of borrowers between rating classes. The fixed-horizon static versions of reduced-form models are often referred to as mixture models. Default events are assumed to be independent, conditional on some common underlying economic or systematic factor. Binomial mixture models are often used, since they intuitively fit the default-no-default idea, while the mixing distribution of the intensity is normally chosen to be a conjugate prior for mathematical tractability. Poisson mixture models are also often used: here the default event is any realization of a Poisson random variable larger than zero. In credit derivative pricing the exact timing of the default is crucial, so these fixed horizon models are inappropriate. However, static reduced-form models are often used for portfolio management. An off-the-shelf model that uses a reduced-form approach is CreditRisk+ by Credit Suisse Financial Products (CSFP). CreditRisk+ uses the Poisson distribution with its conjugate prior (the Gamma distribution) as mixing distribution (Gordy 2000 and CSFP 1997).

Reduced-form models seem to be favoured over structural models once a dynamic credit-risk model is required. Dynamic reduced-form models are also known as dynamic intensity models (Bluhm et al. 2002 p. 77). For these models it is assumed that each default is associated with some counting process driven by a time-dependent default intensity function. These intensity models are akin to the stochastic differential equations used extensively in interest rate modelling. Affine counting processes are used extensively, as these tend to have closed-form solutions. Since the intensities are functions of time, these models may be seen as dynamic, continuous versions of the static CreditRisk+ model. Rating migration reduced-form models are driven by rating transition intensities.

This section was an overview of reduced-form models, in Chapter 4 we discuss counting
processes (which underlie reduced-form models) in detail.

1.3 A common framework

The debate on whether the structural or the reduced-form approach is better suited to modelling credit risk events is on-going. The structural approach is favoured by many practitioners for its econometric interpretability and intuitive appeal. Intuition dictates that the evolution of a firm’s assets and liabilities, although mostly unobservable, can be modelled to some extent. The reduced-form approach is favoured for its mathematical elegance, tractability and the availability of statistical methods to estimate default intensities (Duffie & Lando 2001). Models that combine these approaches can be thought of as a link between the two paradigms. These models are known as incomplete information models or simply hybrid models. ‘Incomplete information’ refers to the lack of information on asset and/or liability values (Giesecke 2004). Hybrid models are normally some type of reduced-form model, with the addition that one of the model’s covariates is the default probability calculated by means of the structural approach.

Structural models assume that complete knowledge of each obligor’s asset values and liability levels is available. This is also known as ‘perfect information’. A problem associated with perfect information is that it results in the default event being predictable. We use the term ‘predictable’ in the technical sense of the word: a predictable process is some process for which the value at time \( t + dt \) is already known at time \( t \), where \( dt \) is a small time increment. Default events are predictable, since asset prices are modelled by predictable geometric Brownian motion. Intuitively, if we were to know the exact content of the financial statements of each borrower at all times, the occurrence of a default event will not be a complete surprise. To counter this limitation, structural models with jumps in the asset price model (using Levy processes instead of geometric Brownian motion) are often used. Another problem with the structural approach is data availability. Typically, not all of the financial statements of a firm will be accurate. Financial statements may also be delayed and/or updated irregularly. This limitation is overcome by using market-observable data to approximate accounting data. Often, even observed market data is hard to come by, especially default data. This might result in the structural approach, despite its known deficiencies, being the only viable model.
In contrast, reduced-form models incorporate market prices far more easily than structural models. The use of the reduced-form approach is much more feasible when large data sets are available for parameter estimation. In reduced-form models the default event is a true surprise, making these models more appealing. A reduced-form model assumes only that observed market data is available.

Structural and reduced-form models are not entirely disconnected. Gordy (2000) and Koyluoglu and Hickman (1998) do some groundwork in comparing these approaches. They find that static reduced-form models and static structural models have a similar mathematical framework, specifically through their use of mixture models, which are special cases of generalized linear mixed models (GLMMs). A GLMM is a broad regression framework that encompasses several regression models. GLMMs facilitate the use of well-developed inference techniques for estimation and diagnostic tests. Typically, the mixture variable in the GLMM is a latent random variable or process used to explain unobservable underlying factors. For instance, the Basel accord prescribes a one-factor model, where the random factor is latent and called the ‘state of the economy’.

As far as practitioners are concerned, the choice of a credit-risk modelling approach is often dictated by data structures, data integrity and data availability. In this study, we follow the generalized linear model (GLM) route. We do, however, extend this approach to be dynamic. We make use of reduced-form models with intensities dependent on both systematic and obligor-specific covariates. In the next section, we discuss some specific issues and problems posed by the modelling of consumer (also known as retail) credit risk.

1.4 Consumer credit risk and model requirements

Consumer credit risk encompasses numerous different types of loans to the ‘man in the street’ (lending to consumers). Consumer loans are typically small loans to unrated borrowers or individuals who are not required to have financial statements. Consumer loans play a significant role in the economy. In the USA in 2007 consumer loans accounted for 46% of the $28 trillion in loans in all sectors (Thomas 2009 p. 2). Of the $13 trillion in retail loans, 75% of loans were mortgage loans. In this section we give an overview of the specific problems faced when modelling consumer credit risk. Consumer risk management
is typically associated with banks, although other lending institutions may also engage in consumer-type lending. Thus, we henceforth refer to all consumer lending institutions as simply ‘banks’. After the sub-prime mortgage loan crisis in 2007, it is now widely understood that consumer banking practices can influence the financial system significantly. Despite this evidence, consumer lending is under-researched when compared to corporate lending (Malik & Thomas 2012). Consumer credit-risk management has many facets, extending the modelling of consumer credit risk beyond the prediction of default probabilities only. Consumer credit-risk management is concerned with, among other things, delinquency, default and loss rates. We define the following events that are of interest to us:

**Definition 1 (Delinquency)** Delinquency is the state that exists when an obligor is in arrears.

**Definition 2 (Failure)** A failure occurs when there is an increase in an obligor’s level of delinquency.

**Definition 3 (Default)** A default occurs when an obligor is in a state of delinquency for three months.

**Definition 4 (Recovery)** A recovery occurs when there is a decrease in the delinquency level of an obligor.

**Definition 5 (Write-off)** A write-off occurs when an obligor’s loan account is written off the balance sheet of the bank.

**Definition 6 (Credit event)** A credit event can be a failure, recovery, default or a write-off.

‘Failure’ and ‘recovery’ are broad terms, but throughout this study we will use these terms as defined above. We use the term ‘credit event’ to describe the occurrence of any of the following four events: failure, recovery, default or write-off. We develop a model to predict and describe the delinquency level (and hence implicitly the default event) of obligors. Any increase in an obligor’s delinquency will be seen as a failure, but the obligor is only seen to have defaulted when the number of months for which the amount
has been due is equal to three months of delinquency. So, by modelling their delinquency level we would know whether and when an obligor defaulted. Our definition of ‘default’ is a simplified version of the prescribed definition by the Basel II accord (BIS 2006 p. 100).

Retail credit risk can be roughly divided into two parts: the credit scoring process and the credit portfolio management process.

Credit scoring is, among other things, the initial process of assessing the creditworthiness of a potential obligor. An applicant will apply for a loan and be scored with some scorecard model, yielding an application score. This application score is used to discriminate between ‘good’ and ‘bad’ loan applicants. The credit scoring process has replaced individual assessment of the credit risk of each applicant by credit ‘experts’. These experts used the ‘five Cs’ approach: character, capital, capacity, collateral and cycle. Credit scoring is a statistical attempt to quantify these five characteristics. Application scorecards are well known and widely used. The purpose of these scorecards is to prevent bad loans from entering a bank’s balance sheet. Effective scorecards should discriminate between obligors and prevent potential future losses. Scorecards are inseparable from retail credit-risk management. Although we are not concerned with credit scoring, we use some credit-scoring techniques for model diagnostic tests.

In addition to application scorecards there are also behavioural scorecards that track the performance of a loan while it is in the portfolio of the bank. Behavioural scorecards are based on similar principles to application scorecards. However, they are only concerned with the performance of the existing loans in the portfolio. Behavioural scorecards are similar to credit ratings for commercial portfolios and are often linked to default probabilities. These scorecards are typically updated monthly and are used to track the performance of the loan and portfolio. Should a loan perform badly, the behavioural score should ideally flag an obligor so that banks are able to initiate corrective measures. Methods used in credit scoring include, to name a few, discriminant analysis, neural networks, classification and regression trees and logistic regression.

Application and behavioural scorecards are methodologies for classifying cross-sectional data. Cross-sectional data is data collected for all obligors at a specific point in time, in contrast to time series data (or longitudinal data), where covariates and responses of
obligors are time-dependent.

As mentioned earlier, we focus on the management of credit risk and not the pricing of credit risk. Banks need to manage their existing loan portfolio to minimize losses. Retail credit portfolio management entails, among other things, estimation of a loss distribution, calculation of risk measures and stress-testing the bank’s current loan portfolios. Banks mostly estimate a default model to (1) predict the occurrence of default given the current prevailing macro-economic environment, and (2) to perform stress-testing. A default model will facilitate the estimation of a loss distribution, which in turn will lead to the calculation of risk measures and capital allocation. Capital allocation, in turn, is a crucial component of the pricing of products.

The existing research on developing credit-risk models is mostly focused on corporate- and SME-oriented credit-risk models. Corporate and SME obligors are perceived to be more highly correlated with the macro-economic environment than retail loans. The risk of a high level of correlation with the prevailing macro-economic environment is known as systematic risk. Also, corporate entities and SMEs are often interdependent and hence they are prone to the default contagion effect. This is also known as ‘counterparty risk’. In contrast to corporate and SME loans, retail loans are often seen to be less dependent on systematic factors and more dependent on the individual obligor’s risk profile. Individual risk profiles are known as ‘idiosyncratic risk’. However, in an analysis of retail, SME and corporate loans in two Swedish banks, Jacobsen et al. (2005) found that retail and SME loans are consistently more systematically risky than corporate bonds. Hence, they challenge the belief that retail loans have less systematic risk than corporate loans.

In retail credit-risk management we assume that there is systematic risk, but no counterparty risk. Since the credit events of individuals are seen as conditionally independent (conditional on the current prevailing economic environment), this assumption seems reasonable. Hence, while all individuals are influenced by common macro-economic factors, no intrinsic interdependency exists between individuals.

Consumer loans may be one or more of many types: mortgage loans, credit cards, short-term loans, student loans, asset finance or vehicle finance. Considering this, it is important to note that a ‘one-size-fits-all’ approach to consumer credit-risk modelling is not always appropriate (Risk Management Association 2000). The highly differenti-
ated nature of each of the loan products mentioned above requires that each product be modelled according to its own characteristics.

The field of consumer credit risk is known for the availability of ample amounts of data. This might seem to be a ‘statistician’s paradise’, but large data sets require substantial computing power, even by today’s standards. Large data sets often result in statistical rather than practical significance. Additionally, sometimes complicated numerical methodologies need to be simplified to be able to churn through the millions of records.

Although most banks do not publish their consumer credit-risk models, some research has been published on this topic, such as that by Allen et al. (2004), which surveys these models. Most attempts have simply been based on adjusting or tweaking well-known corporate credit-risk models. The structural approach is very popular owing to the Basel capital accord prescribing it via a ‘one-factor model’. A detailed description of the Basel accord’s infinitely divisible one-factor model approach can be found in Schönbucher (2000). Essentially, if a bank decides to follow the Basel accord’s internal ratings based (IRB) approach, it is permitted to estimate its own default probabilities. These are substituted into the Basel formulae to calculate risk-weighted assets and the minimum capital requirement. The basic assumption behind the structural approach, i.e. that a company defaults when debt level exceeds asset level, is not the reason why consumers default. Retail consumers usually default because of cash flow problems, financial naïveté or fraud (Malik & Thomas 2012). Generally, consumer credit-risk models are developed internally by banks and are therefore proprietary. This makes the study and comparison of these models problematic.

It is extremely difficult to obtain the asset values and liability levels of retail obligors. These variables are essentially unobservable. In South Africa, retail loans are rarely traded among banks, making credit options and mark-to-market pricing impossible. Assuming that each obligor has the same debt threshold, as in the Basel accord, seems implausible. Hence, employing the market value structural approach directly to quantify retail credit risk is inappropriate. However, a structural approach with multiple factors was implemented by Bucay and Rosen (2000). They use principal components to orthogonalize factors using a model akin to Wilson’s (1997a,b) corporate model. Bucay and
Rosen (2000) model the default probability as a function of macro-economic factors and derive a loss distribution by simulation. Rösch and Scheule (2004) extend this model by assuming the liability threshold level is a linear function of macro-economic factors, rather than a constant. They use the consumer price index (CPI), deposit interest rate, gross domestic product (GDP) and industrial production index as macro-economic co-variates. Sometimes the asset value in Merton’s model is seen as a ‘creditworthiness’ factor. Should an obligor’s creditworthiness fall below a certain level (Merton’s liability level), the obligor is assumed to have defaulted. Again, all obligors are assumed to have the same creditworthiness threshold, and this model seems to oversimplify reality. Andrade and Thomas (2007) describe a structural model where the behavioural score or credit bureau score is a proxy for the creditworthiness of an obligor.

Interestingly, there has been relatively little work done on the reduced-form model in the retail environment. Allen et al. (2004) state:

"No formal model for retail credit risk has yet used the reduced-form approach, although one model, Credit Risk Plus, could be used for retail credit."

We follow a reduced-form approach in the modelling of credit events for retail credit portfolios. Specifically, we make use of counting process methodology.

1.5 Layout of the thesis

This study is focused on consumer credit risk, specifically the modelling of credit events. We apply this to a mortgage loan portfolio. The introduction set the stage for further research into the prediction of credit events. The layout of this thesis is discussed in this section. In Chapter 2 we will describe our real-world data through some summary statistics and describe the covariates that we deem important or potentially significant in predicting credit events. These covariates include idiosyncratic and systematic covariates and may either be time-dependent (dynamic) or time independent (static). In Chapter 3 we will focus on the classification of cross-sectional data. We specifically focus on the use of generalized linear models (GLMs) as a classification methodology. We briefly discuss variable selection methodologies. We also describe diagnostic tests called marginal-model
plots, which were developed by Cook and Weisberg (1997), and we contribute a new method of deriving confidence bands for these plots. Other diagnostic tests such as the receiver operating characteristic curve and other goodness-of-fit tests will also be discussed. Although our data is not cross-sectional (it is longitudinal) the techniques discussed in Chapter 3 can be extended to apply to our data. The topic of Chapter 4 is the use of counting processes as a way of modelling longitudinal data. We begin by summarizing survival analysis and show how it links to the greater family of models, namely counting processes. Both counting processes and survival analysis are mostly used in continuous time, so we discretize these models. We show that the use of survival analysis techniques is not well suited to model credit risk. Briefly, this is because survival analysis is mostly concerned with the prediction of the time until a single event. The time domain used is event time, so all observations need to be aligned to some starting time. We prefer to work in the calendar-time domain and are concerned with the timing (in calendar time) of multiple events. In Chapter 5 we identify two models that show potential in modelling and predicting a default event. The first model is from Ding et al. (2012) and the second from Arjas and Haara (1987). The Ding et al. (2012) model only allows for a single default to occur, which leads us to rather employ Arjas and Haara’s model. This model has, to our knowledge, never been applied to credit risk before. We show empirically by simulation how well the theoretical asymptotic estimated parameter properties of the Arjas and Haara model hold. We then apply the theory of Chapter 5 by implementing both of the identified models with real-world data in Chapter 6. Here we model the occurrence of a once-off default event. In both cases we found that the default event is highly dependent on the current delinquency level, and so we develop a model to predict the delinquency level in Chapters 7 and 8. In Chapter 7 we spend time on developing a model that predicts a failure event. We also make use of the Arjas and Haara approach, but this time with repeated credit events allowed. Combining the failure process derived in Chapter 7 with another marked-point process that predicts the size (mark) and timing (point) of the repayments made, we obtain a delinquency model in Chapter 8. Chapter 8 also develops a transition-matrix approach where we apply Arjas and Haara’s (1987) assumptions to modify the approach of Grimshaw and Alexander (2010) to model delinquency. We conclude the chapter by performing a comparative
analysis on the two proposed models. Lastly, we summarize the findings and suggest further research in the last chapter. The list of abbreviations, covariates, notation and definitions is given at the end of this thesis.
Chapter 2

The data

In this chapter we describe the mortgage loan data kindly disclosed to us by a major South African bank. We also discuss potential covariates that may be used in the prediction of credit events.

2.1 Summary statistics

We received monthly data for mortgage loans for 60 months from January 2006 to December 2010. The model was applied to mortgage loans that originated between January 1984 and November 2010. After removing obvious data errors (like the closing date being before the opening date) we had 638,287 accounts. Furthermore, we excluded all loans with terms less than 48 months and only used the most common installment frequency: monthly. The observation window was 48 months (from January 2007 to December 2010), one year less than the period for which we received data. This truncation occurred because of the lags of certain covariates (sometimes lagged by a year). This observation window can, of course, be extended to include the five-year data window as prescribed by the Basel accords, since at the time of writing more data has become available. We did not include the latest data in this thesis, as this is an academic publication and the emphasis is not on the interpretation of the results, but rather on the techniques and methodologies applied/developed. We excluded extremely small or large mortgage loans (higher than R5,000,000 or less than R10,000). We removed all data that did not have a delinquency level history (a data error) and assumed that all accounts must be less
than or equal to six months delinquent. If any account was delinquent for more than six months, then we would assume that it has been written off.

After these exclusions were made, there were 492,588 accounts. Over the 48 months, a total of 19,525,220 ‘obligor month’ observations were made. Not all accounts were observed over all 48 months, as some accounts closed during the 48-month period and others may have opened during that time. On average we observed each obligor for about 40 months \((19,525,220/492,588 = 39.64)\). Since the occurrence of a failure event (a missed payment) has a low probability and we have limited computing power, we performed a retrospective study: we only considered accounts that experienced a failure event within the window of January 2006 to December 2010. So, a biased sample was taken, with oversampling from the ‘failure accounts’. The number of obligors decreased to (a still large) 112,676 obligors or 4,523,811 obligor months. The average time for which we observed an obligor remained more or less the same \((4,523,811/112,676 = 40.15)\).

This retrospective study represents a surprisingly high 22.87% of the total population. From now on, all analysis will be done on this retrospective data sample. It is important to note that although we do not concern ourselves too much with the absolute level of the overall portfolio default rate, it should be kept in mind that this retrospective sample only represents 22.87% of the total population. Therefore, to obtain the correct overall default level, adjustments should be made to the default probabilities that are derived.

The payment history of each loan enables us to determine the number of installments an obligor has in arrears and makes it possible to derive failure, recovery and write-off indicators. Since we will later be concerned with the modelling of default in Figure 2-1 we plot the default rate (for our retrospective sample) over time. The peak around July 2009 was the effect of the global recession. We are also interested in the modelling of delinquency levels. So, in Figure 2-2 we plot the percentage (or rate) of obligors that experienced a failure event from one month to the next. The recovery rate is superimposed on the same figure. It is interesting to note that the failure rate is almost always higher than the recovery rate. Again, the peak in 2009 is evident. In this thesis we develop models to predict default, failure and delinquency levels. These events are driven by, among other things, obligor-specific covariates, business decisions and external covariates. In the next sections we discuss the covariates we deem important to predict credit events.
Figure 2-1: The default rate over time. The economic crises only started to affect the mortgage loan portfolio in January 2008 and peaked at July 2009.

Covariates are split between obligor-specific (or idiosyncratic) covariates and external (or systematic) covariates. The number of idiosyncratic covariates may be very large (for behavioural or application scoring, the number is typically well above 200). We were, however, limited to the few covariates that were supplied to us. Systematic covariates are macro-economic variables. We discuss both idiosyncratic and systematic covariates in the next sections.

Credit events are also driven by certain influencing factors that are impossible to observe or estimate, such as the strategic business decisions made by portfolio managers. We do not know what decisions were made or when, and as a result we do not account for these decisions in this study.

Sometimes when modelling a specific outcome, no real covariates with practical causality can be found. For this reason the use of latent variables is popular in credit-risk management. These are used as a type of additional error term that explains all unobservable events. Although latent covariates may improve the prediction quality (the Vasicek distribution used in the Basel accord uses a latent covariate model), this comes at the price
Figure 2-2: The recovery and failure rate. There seems to be a positive correlation (the Pearson's correlation coefficient was 0.34).

of increased mathematical complexity and decreased interpretability. In this study we omit the use of latent covariates, but the use of these covariates should be investigated in further research.

2.2 Covariates

We identified some covariates that might be useful in predicting credit events. We will discuss both idiosyncratic and systematic covariates. The idiosyncratic covariates are specific to each obligor, and may be dynamic or static. Static covariates (like the obligor’s date of birth) are fixed at the outset of the loan, while dynamic covariates (like the obligor’s age) are updated monthly. Systematic covariates influence all obligors and are assumed to be dynamic. We start out by discussing the idiosyncratic covariates that are static, and then progress to the dynamic idiosyncratic covariates. Lastly, we will discuss the systematic covariates.
2.2.1 Idiosyncratic static covariates

Here we describe and analyse some static idiosyncratic covariates. Typically banks would have abundant idiosyncratic static covariates to derive the credit score of an application. Unfortunately, in this study we were not supplied with the mortgage loan application scores, which in theory should combine all the idiosyncratic static covariates into a useful ranking index. We were only supplied with some of the covariates that are used to derive the application score. Table 2.1 describes the static idiosyncratic covariates we received.

Table 2.1: Static idiosyncratic covariates

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLT V</td>
<td>The loan-to-value ratio at the commencement of the loan.</td>
</tr>
<tr>
<td>Purchase price</td>
<td>The original purchase price of a property.</td>
</tr>
<tr>
<td>Ream ABC code</td>
<td>The Ream value: An ordinal classification of the quality of the neighbourhood. It determines the loss severity from 1 (best neighbourhoods) to 4 (worst neighbourhoods).</td>
</tr>
<tr>
<td>Security value</td>
<td>The total valuation of the property (land, building, alterations, swimming pool, and so on).</td>
</tr>
<tr>
<td>Term</td>
<td>The full period over which the loan is to be repaid.</td>
</tr>
<tr>
<td>Total bond amount</td>
<td>The sum of all the bond amounts registered on a specific loan account.</td>
</tr>
<tr>
<td>Total loan granted</td>
<td>The sum of all loans granted to the client up to the current date (this would include further loans, i.e. every time the client signs a new contract).</td>
</tr>
</tbody>
</table>

Collectively the covariates in Table 2.2 should capture the creditworthiness of an obligor. The loan-to-value covariate at origination should give a good indication of the ability of the obligor to repay his or her loan. The Ream ABC code covariate indicates the quality of the suburb and hence the potential recovery amount should an obligor default. The purchase price, total loan amount, total loan granted and security value covariates are indications of the size of the loan, and the term covariate indicates how soon an obligor plans on repaying the loan. After standardization we transformed the covariates to get their distributions as close to a normal distribution as possible. This was done by using a Box-Cox transformation with only an intercept term as a covariate. Whenever the Box-Cox suggested a log transformation, we transformed the covariate to \( x^* = \log(x + 1) \) to prevent a log of zero from occurring. The descriptive statistics and
transformation of each covariate are given in Table 2.2.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Mean</th>
<th>Std.</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLT V</td>
<td>0.71</td>
<td>0.28</td>
<td>0.00183</td>
<td>2.99</td>
<td>-</td>
</tr>
<tr>
<td>Purchase price</td>
<td>498.054</td>
<td>518.335</td>
<td>10,003</td>
<td>4,995,000</td>
<td>log(x + 1)</td>
</tr>
<tr>
<td>Ream abc code</td>
<td>2.86516</td>
<td>0.82993</td>
<td>1</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>Security value</td>
<td>699.896</td>
<td>59.8451</td>
<td>10,001</td>
<td>4,995,000</td>
<td>log(x + 1)</td>
</tr>
<tr>
<td>Term</td>
<td>242.06</td>
<td>50.74</td>
<td>59</td>
<td>720</td>
<td>square root</td>
</tr>
<tr>
<td>Total bond amount</td>
<td>508510</td>
<td>483.516</td>
<td>10,093</td>
<td>4,965,000</td>
<td>log(x + 1)</td>
</tr>
<tr>
<td>Total loan granted</td>
<td>478355</td>
<td>448.126</td>
<td>10,012</td>
<td>4,960,000</td>
<td>log(x + 1)</td>
</tr>
</tbody>
</table>

An approximation of the interaction between the static covariates can be made through a correlation analysis. We calculate Spearman’s correlation coefficients for the various covariates in Table 2.3. Owing to the large sample size, all correlations were significantly different from zero at a level of $\alpha = 0.001$. We noticed a high correlation between the security value, total bond amount, total loan granted and purchase price, which may lead to multicollinearity. Multicollinearity has been shown to be an unimportant issue for prediction purposes in logistic regression with large sample sizes (de Jongh et al. 2013). However, in our effort to keep the model as parsimonious and interpretable as possible, we replaced these covariates with a single covariate: the average of these four loan size variables. We call this covariate ‘bond’.

Table 2.2: Static idiosyncratic covariates: Descriptive statistics.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Mean</th>
<th>Std.</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLT V</td>
<td>0.71</td>
<td>0.28</td>
<td>0.00183</td>
<td>2.99</td>
<td>-</td>
</tr>
<tr>
<td>Purchase price</td>
<td>498.054</td>
<td>518.335</td>
<td>10,003</td>
<td>4,995,000</td>
<td>log(x + 1)</td>
</tr>
<tr>
<td>Ream abc code</td>
<td>2.86516</td>
<td>0.82993</td>
<td>1</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>Security value</td>
<td>699.896</td>
<td>59.8451</td>
<td>10,001</td>
<td>4,995,000</td>
<td>log(x + 1)</td>
</tr>
<tr>
<td>Term</td>
<td>242.06</td>
<td>50.74</td>
<td>59</td>
<td>720</td>
<td>square root</td>
</tr>
<tr>
<td>Total bond amount</td>
<td>508510</td>
<td>483.516</td>
<td>10,093</td>
<td>4,965,000</td>
<td>log(x + 1)</td>
</tr>
<tr>
<td>Total loan granted</td>
<td>478355</td>
<td>448.126</td>
<td>10,012</td>
<td>4,960,000</td>
<td>log(x + 1)</td>
</tr>
</tbody>
</table>

Table 2.3: Spearman’s coefficient of correlation for static idiosyncratic covariates.

<table>
<thead>
<tr>
<th>Spearman’s Correlation</th>
<th>OLT V</th>
<th>Purchase price</th>
<th>Ream abc code</th>
<th>Security value</th>
<th>Term</th>
<th>Total bond amount</th>
<th>Total loan granted</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLT V</td>
<td>1</td>
<td>0.24845</td>
<td>0.01519</td>
<td>−0.10652</td>
<td>0.18285</td>
<td>0.37482</td>
<td>0.40644</td>
</tr>
<tr>
<td>Purchase price</td>
<td>0.24845</td>
<td>1</td>
<td>0.017</td>
<td>0.72696</td>
<td>0.18369</td>
<td>0.79947</td>
<td>0.80891</td>
</tr>
<tr>
<td>Ream abc code</td>
<td>0.01519</td>
<td>0.017</td>
<td>1</td>
<td>0.01366</td>
<td>−0.01008</td>
<td>0.0235</td>
<td>0.02146</td>
</tr>
<tr>
<td>Security value</td>
<td>−0.10652</td>
<td>0.72696</td>
<td>0.01366</td>
<td>1</td>
<td>0.09453</td>
<td>0.81887</td>
<td>0.81321</td>
</tr>
<tr>
<td>Term</td>
<td>0.18285</td>
<td>0.18369</td>
<td>−0.01008</td>
<td>0.09453</td>
<td>1</td>
<td>0.18526</td>
<td>0.20317</td>
</tr>
<tr>
<td>Total bond amount</td>
<td>0.37482</td>
<td>0.79947</td>
<td>0.0235</td>
<td>0.81887</td>
<td>0.18526</td>
<td>1</td>
<td>0.98285</td>
</tr>
<tr>
<td>Total loan granted</td>
<td>0.40644</td>
<td>0.80891</td>
<td>0.02146</td>
<td>0.81321</td>
<td>0.20317</td>
<td>0.98285</td>
<td>1</td>
</tr>
</tbody>
</table>
2.2.2 Idiosyncratic dynamic covariates

Dynamic idiosyncratic covariates are specific to individual obligors and change over time. A behavioural score is a creditworthiness ranking index determined by, among other things, the payment history of an obligor. Ideally, a behavioural score should capture all dynamic idiosyncratic variability. Again, we were not supplied with such a score. Although all of these covariates change over time, we suppressed the subscript \( t \) to make the presentation more concise. We add the following potential covariates as substitutes for the behavioural score:

1. **Age of the loan (Age)**: The age of the loan from origination, which is a predictable covariate. Loans close to maturity should be less prone to default, whereas other loans should have a higher risk of failure/default.

2. **The current loan-to-value ratio (CLTV)**: The CLTV is a ratio that should determine an obligor’s willingness to repay the loan. A CLTV greater than one implies that the loan’s value is higher than the value of the mortgage. This can occur if house prices have declined significantly or if the loan value increases. The latter can occur when an obligor extends his loan by refinancing his house. The value of the CLTV is random, since interest and house-price growth rates are unknown. Additionally, early repayments or extensions of loans cannot easily be modelled. However, conditional on interest rates and the house-price index and assuming that no additional payments or refinancing were made, the CLTV is deterministic. We will derive a proxy for the CLTV by approximating the current house value with an index, \( H_t \), (we use the First National Bank [FNB] house-price index) and assuming that interest rates, \( r_t \), are the prime interest rate minus a one percent concession for \( t = 1, \ldots, T \). Additionally, we suppose all obligors repay their loans over the full term of the loan (\( T \)). The full derivation of the CLTV index is described in the appendix at the end of this chapter.
3. **The total months that the account was in arrears in the last six months** ($TD$): We use the total number of months an account was delinquent in the last six months as an indication of delinquency and willingness to pay.

4. **Cumulative failure events up to time $t$ of the loan** ($CF$): We add the total number of failure events that an obligor incurred up to time $t$.

5. **The current delinquency level** ($D$): We use the delinquency level in the previous month as a potential covariate.

6. **Maximum delinquency rate** ($MD$): The maximum delinquency amount that the account was delinquent from origination to time $t$.

7. **Average long term delinquency rate** ($AD$): The average delinquency level from origination up to time $t$.

8. **Time since last failure** ($TF$): The number of months since the last failure occurred.

9. **Time since last recovery** ($TR$): The number of months since the last repayment was made.

10. **The total number of repayments made during the last six months** ($TP$): We use the total number of repayments (additional to the normal installment due) made during the previous six months.

11. **Cumulative number of recovery events up to time $t$ of the loan** ($CR$): We use the cumulative number of repayments made up to time $t$.

We summarize the dynamic idiosyncratic covariates and the transformations used in Table 2.4. After transformation, all covariates were standardized to have zero mean and unit standard deviations.

### 2.2.3 Systematic covariates

In this section we will discuss some macro-economic covariates that we consider as potential covariates in modelling default events. The effects of macro-economic covariates on
Table 2.4: Dynamic idiosyncratic variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Age(t)$</td>
<td>The age of the loan.</td>
<td>$\log(x + 1)$</td>
</tr>
<tr>
<td>$CLTV(t)$</td>
<td>Current loan to value.</td>
<td>$-$</td>
</tr>
<tr>
<td>$TD(t)$</td>
<td>The total months that the account was in arrears in the last six months.</td>
<td>$-$</td>
</tr>
<tr>
<td>$CF(t)$</td>
<td>The total number of failure events that an obligor incurred up to time $t$.</td>
<td>$\log(x + 1)$</td>
</tr>
<tr>
<td>$D(t)$</td>
<td>The delinquency level at time $t$.</td>
<td>$-$</td>
</tr>
<tr>
<td>$MD(t)$</td>
<td>The maximum delinquency amount that the account was delinquent from origination to time $t$.</td>
<td>$-$</td>
</tr>
<tr>
<td>$AD(t)$</td>
<td>Average long term delinquency rate up to time $t$.</td>
<td>$-$</td>
</tr>
<tr>
<td>$TF(t)$</td>
<td>Time since last failure.</td>
<td>$-$</td>
</tr>
<tr>
<td>$TR(t)$</td>
<td>Time since last recovery.</td>
<td>$\log(x + 1)$</td>
</tr>
<tr>
<td>$TP(t)$</td>
<td>The total number of repayments made during the last six months.</td>
<td>$\log(x + 1)$</td>
</tr>
<tr>
<td>$CR(t)$</td>
<td>Cumulative number of repayments made up to time $t$.</td>
<td>$\log(x + 1)$</td>
</tr>
</tbody>
</table>

default events will also be discussed. We begin by exploring the macro-economic variables that were found to be significant in other empirical studies.

Couderc and Renault (2005) group macro-economic variables into three groups: financial market information, business cycle indicators and credit market information. Financial information includes all of the information that is observable in the market. In order to represent the financial market information they use the return and volatility of S&P500, the 10-year treasury yield and a variable defined as the "slope of term structure": the 10-year interest rate minus a one-year interest rate. The business cycle is an indicator of business health obtained from real gross domestic product ($GDP$), industrial production growth (an approximation for $GDP$ that is updated more often), personal income growth, $CPI$ growth and other macro-economic factors. For credit market information they use the ratio of BBB-rated companies’ yield to treasury yield and BBB rated companies’ yield to AAA rated companies’ yield (using the ratings by Moody’s), as well as net issues of treasury securities and money lending such as the M2 money supply. Couderc and Renault (2005) perform a principal component analysis and find that five factors explain 71% of the total variation captured by their data set.

Wilson (1997a, b) uses overall unemployment, $GDP$ growth rate, rate of government spending and the regional housing price index to estimate the credit-risk measures used
in his mortgage loan portfolios.

In line with Wilson (1997a, b) and Couderc and Renault’s (2005) practices, we identified the following list of potential macro-economic covariates:

1. **Interest rates** (*Prime*). The main driver of the property market and defaults may well be interest rates. High interest rates slow property price growth and increase the monthly installments owed by obligors. Should distressed obligors attempt to sell their houses the availability of potential buyers will be impaired by higher interest rates. Additionally, higher supply and lower demand will deflate house prices. In South Africa, most home loan rates are linked to the prime interest rate, which in turn is linked to the repurchase rate (repo). Banks borrow at the repo rate and lend at the prime rate. The repo rate is announced quarterly by the monetary committee, whose mandate it is to curb inflation. For this reason we find that the prime interest rate is extremely highly correlated to the consumer price index (*CPI*). Hence, we omit the *CPI* from our analysis. An increase in interest rates should result in an increase in failures, defaults and write-offs.

2. **Property price index** (*PPI*). We used the property price index in the calculation of the current loan-to-value ratio in the previous section. We will use this index as an indication of property capital growth. One would assume that if the value of a property rises, an obligor would be less inclined to miss a payment or to default. Higher property prices imply that a distressed obligor can, rather than default, effortlessly sell their properties. In this study, we use the FNB house-price index. This index is highly correlated with the Absa house-price index. If the year-on-year change in the property price index is negative, this could lead to a situation in which an obligor owes more on the mortgage than the value of the house. This would result in a current loan-to-value ratio greater than one. These obligors are sometimes referred to as ‘upside-down’ homeowners and should have a higher probability of defaulting.

3. **Leading business-cycle indicator** (*Lead*). This is an indicator of the future state of the economy. An increase in this indicator would imply a positive expectation for economic growth. This would in turn lead to more employment opportuni-
ties and hence better bond repayments. This index is highly correlated to the G7 leading indicator and the property price index.

4. **The gross domestic product** \((GDP)\) should be an indicator of a country’s overall financial health. This macro-economic variable is included in several empirical studies. We found it to be highly correlated to M3 money supply. An increase in the GDP should result in a lower default rate.

5. **Personal savings** \((PSav)\). The value of personal savings should be an indicator of retail clients’ propensity to save. An increase in this index should result in lower defaults.

6. **Personal disposable income** \((PDispInc)\). We deem the amount of disposable income crucial to the ability to repay loans. Hence we suspect that a higher personal disposable income level would lead to fewer defaults.

7. **Rand-dollar exchange rate** \((R\$)\). This is an indication of the globally perceived creditworthiness or investment potential of South Africa. Like \(GDP\), it is an indicator of the country’s economic health. An increase in the rand–dollar exchange rate means that the global community sees South Africa as being a risky investment, and hence that our economy is under pressure. We expect to find a positive relationship between the rand–dollar exchange rate and the probability of default.

8. **Total number of insolvencies.** \((TIns)\). This index is an indicator of the number of defaulted or insolvent obligors across South Africa. This index should be positively correlated with the default rate.

9. **Liquidations** \((Liq)\). This indicates the total number of liquidations in a specific month. We expect that this index will be a lagging indicator of default.

10. **M3 money supply** \((M3)\). \(M3\) money supply includes: notes and coins (currency) in circulation, traveler’s cheques from non-bank issuers, demand deposits, savings deposits, long-term deposits, institutional money market funds, short-term repurchase operations and other larger liquid assets. The \(M3\) should be an index representing the business cycle. We found it to be highly correlated to the \(GDP\).
Table 2.5: Systematic covariates.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prime</td>
<td>Interest rates</td>
</tr>
<tr>
<td>PPI</td>
<td>Property price index</td>
</tr>
<tr>
<td>Lead</td>
<td>Leading business-cycle indicator</td>
</tr>
<tr>
<td>GDP</td>
<td>The gross domestic product</td>
</tr>
<tr>
<td>PSav</td>
<td>Personal savings</td>
</tr>
<tr>
<td>PDispInc</td>
<td>Personal disposable income</td>
</tr>
<tr>
<td>R$</td>
<td>Rand-dollar exchange rate</td>
</tr>
<tr>
<td>TIns</td>
<td>Total number of insolvencies</td>
</tr>
<tr>
<td>Liq</td>
<td>Liquidations</td>
</tr>
<tr>
<td>M3</td>
<td>M3 money supply</td>
</tr>
</tbody>
</table>

The consumer price index (CPI) was not included in our list because of the extremely high correlation it had to the prime interest rate. A rising CPI index implies higher cost of living that makes it increasingly difficult for obligors to service debt. Basic needs like food form a larger part of household expenditure for lower-income households. Rising food prices affect low-income households more than medium- to high-income households. We also did not include the unemployment rate: this rate is only published quarterly and was very flat over the estimation window of the study.

Some macro-economic factors have a delayed influence on default events, and hence we include the lagged values of all 10 macro-economic variables. We include 12-, six- and three-month lagged macro-economic variables. We summarize the systematic covariates in Table 2.5. Again, all macro-economic covariates were standardized.

2.3 Appendix

2.3.1 Derivation of CLTV

We know the maturity term $T$ and the principal loan amount $P_0$. Without the loss of generality, we may assume that the principal loan amount is $P_0 = 1$. The value of the loan, $V$, at the end of $t = 1$ is

$$V_1 = P_1 - A_1,$$

where $P_1$ is the outstanding amount at the end of time $t = 1$ and $A_1$ is the monthly installment made at the end of the month $t = 1$. Note that $P_1 = P_0 \exp(r_1) = \exp(r_1)$.
since the prevailing interest rate for the first month was $r_1$. The monthly installment is calculated so that the loan is always repaid in full at the original maturity date $T$, so $V_T := 0$. Using geometric series and $V_T = 0$ we get

$$A_1 = e^{Tr_1} \left[ \frac{1 - e^{r_1}}{1 - e^{r_1 T}} \right].$$

The value of the loan at the end of $t = 1$ is

$$V_1 = e^{r_1} - e^{Tr_1} \left[ \frac{1 - e^{r_1}}{1 - e^{r_1 T}} \right].$$

The value of the loan at the end of $t = 2$ would then be

$$V_2 = V_1e^{r_2} - A_2,$$

where the installment $A_2$ is given by

$$A_2 = V_1e^{(T-1)r_2} \left[ \frac{1 - e^{r_2}}{1 - e^{r_2(T-1)}} \right].$$

Hence

$$V_2 = V_1e^{r_2} - V_1e^{(T-2)r_2} \left[ \frac{1 - e^{r_2}}{1 - e^{r_2(T-1)}} \right]$$

$$= e^{r_1} - e^{Tr_1} \left[ \frac{1 - e^{r_1}}{1 - e^{r_1 T}} \right] \left[ e^{r_2} - e^{(T-1)r_2} \left[ \frac{1 - e^{r_2}}{1 - e^{r_2(T-1)}} \right] \right]$$

$$= \prod_{j=1}^{2} \left[ e^{r_j} - e^{(T-j+1)r_j} \left[ \frac{1 - e^{r_j}}{1 - e^{r_j(T-j+1)}} \right] \right].$$

Continuing in this manner we get the loan value at time $t$ to be

$$V_t = \prod_{j=1}^{t} \left[ e^{r_j} - e^{(T-j+1)r_j} \left[ \frac{1 - e^{r_j}}{1 - e^{r_j(T-j+1)}} \right] \right].$$
Therefore, the CLTV at time $t$ can be estimated ($\hat{C}_{LTV_i}$) by dividing the current outstanding loan amount by the value of the house

$$\hat{C}_{LTV_i} = V_t / \exp(\sum_{i=1}^{t} H_t).$$

(2.1)

Note that the value for each factor $V_t$ and $\exp(\sum_{i=1}^{t} H_t)$ in (2.1) will be different for most obligors, since they have different maturity terms and different origination dates. These time scales influence the prevailing interest rates and house-price indexes used in the ratio.
Chapter 3

Modelling static credit events: classification techniques for cross-sectional data

This is the first of two chapters that describe various methods of modelling credit events. Credit event models not only describe the occurrence of credit events, but also facilitate their prediction. Respectively, these abilities will enhance understanding of the drivers of default and enable the stress-testing of a credit portfolio. We distinguish between two broad classes of credit event modelling, namely the analysis of cross-sectional and the analysis of longitudinal data. A simple way to model the occurrence of credit events is to classify cross-sectional data. Cross-sectional classification models do not allow for repeated defaults or for any correlation within obligors’ credit events. Cross-sectional techniques also tend to be more static in nature. We use dynamic longitudinal data in this study. The reason we discuss cross-sectional techniques in this chapter is because cross-sectional analysis forms the foundation of more advanced longitudinal-type analysis. These models are static in the sense that the response variables (in this case, default or not) are always assumed to be independent between obligors and not dependent on time. A natural extension of these models is a type of longitudinal analysis that incorporates repeated events and is time-dependent. We extend the cross-sectional classification techniques discussed in this chapter to longitudinal analysis in Chapter 4.
Throughout this chapter we will describe the models by keeping the following framework in mind:

1. Model identification: choosing a specific model within one of the model classes;
2. Parameter estimation: the selection of variables and the fitting of the model; and
3. The determination of the model quality (prediction accuracy and discrimination) or model checking: here we will focus on diagnostic tests and remedial measures.

These steps are not strictly sequential. It is often necessary to retrace some steps after making some adjustment to the model. We summarize the relevant literature and contribute some improved diagnostic techniques. Throughout this brief theoretical literature study we mostly focus on binary data, i.e. where the response variable can only be one of two outcomes.

### 3.1 Classification: an overview

The classification of cross-sectional data is central to discriminating between different response groups. The objective of the classification of cross-sectional data, hereafter simply referred to as ‘classification’, is to use cross-sectional data to derive a rule that classifies events. Classification techniques will be used to attempt to derive some rule that classifies obligors as either having a credit event within a certain time period or not. Classification analysis is widely used for any type of classification, but lacks the ability to incorporate time-dependent data. Hence, classification is similar to the static credit-risk approach. As mentioned in the model requirements, we would like to develop a dynamic model that can predict a credit event at any time within a certain timeframe. To this end, this chapter will mostly serve as a foundation for the chapters to follow. The classification techniques discussed here will serve as the fundamental building blocks of more advanced models and can also serve as naïve comparison models. We discuss the broad idea of classification and then one useful example of classification: generalized linear models.

Classification techniques are widely used in retail applications of credit-risk scoring. Application scorecards have static predictor variables, and the response variable
is whether or not an obligor has defaulted within a certain timeframe. Discriminating
between performing and non-performing obligors thus involves the use of static models.
Therefore, application scorecards are in the domain of cross-sectional data classification.

The classification of binary response variables can normally be expressed as a function
of certain covariates. For each obligor, these covariates may be univariate scalars or
multivariate vectors. Suppose that for an obligor we have a default indicator variable

\[ Y = 1 \{ \text{The obligor experienced a credit event} \}, \]

where \( 1 \{ \cdot \} \) is the indicator function. Further suppose that the obligor has a \((1 \times p)\)
static explanatory variable vector \( \mathbf{X} = (X_1, X_2, \ldots, X_p) \). We want to derive a classification
rule that is some function \( h : \mathbf{X} \rightarrow Y \) such that when we observe \( \mathbf{X} \), we can use \( h(\mathbf{X}) \)
to predict the value of \( Y \). There are several well-developed methods to classify a ran-
don response variable: we may use, to name a few, discriminant analysis, classification
trees, logistic regression, linear classifiers, odds ratios, random forests or neural networks.
Some classifiers are statistical procedures or models, while others have been supported
by advances in machine learning and ever-increasing computational power. In this sec-
tion we will discuss the underlying fundamental classification problem and highlight some
solutions to this problem. We follow the presentation by Wasserman (2003 p. 349).

The principle of classification is captured by the Bayes classification rule. Most of the
statistical classification rules are, in essence, an attempt to estimate Bayes’ rule. Bayes’
rule is an intuitive way to approach classification by using conditional expectation. If
we observe the data \( \mathbf{x} \), we are interested in the function \( r(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x}) \).
From Bayes’ theorem we get

\[
\begin{align*}
    r(\mathbf{x}) &= P(Y = 1|\mathbf{X} = \mathbf{x}) \\
    &= \frac{f(\mathbf{x}|Y = 1)P(Y = 1)}{f(\mathbf{x}|Y = 1)P(Y = 1) + f(\mathbf{x}|Y = 0)P(Y = 0)} \\
    &= \frac{f(\mathbf{x}|Y = 1)\pi}{f(\mathbf{x}|Y = 1)\pi + f(\mathbf{x}|Y = 0)(1 - \pi)},
\end{align*}
\]

where \( f(\mathbf{x}|Y = y) \) is the marginal density function of \( \mathbf{x} \) conditioned on \( Y = y \) for
$y = \{0, 1\}$ and $\pi = P(Y = 1)$. The Bayes classification rule is defined as

$$h^*(x) = \begin{cases} 1 & \text{if } r(x) \leq b \\ 0 & \text{otherwise} \end{cases},$$

(3.1)

where $b$ is some cut-off value, $b \in [0, 1]$. Inspecting (3.1) the function of covariates $r(x)$ is a good discriminatory rule if the defaults are all above the threshold $b$. So, a perfect discrimination will be obtained if the density of the random variable for the defaults ($f(x|Y = 1)$) and the density of the random variable for non-defaults ($f(x|Y = 0)$) do not overlap.

Before we extend the idea of classification rules, we need some measure to determine which classification rule is the best classifier within a group of classifiers. Classification rules may be compared by their accuracy in predicting classes correctly. Their accuracy can be measured by the so-called ‘true error rate’. The true error rate of classifier $h$ is

$$L(h) = P(h(X) \neq Y).$$

(3.2)

Estimating the true error rate can be done by using the empirical- or training error rate, $h$, defined as $L_n(h) = \frac{1}{n} \sum_{i=1}^{n} I \{h(X_i) \neq Y_i\}$ for observed values $x_i$ and $y_i$ for each obligor $i = 1, 2, ..., n$. It is known that the Bayes classification rule $h^*$ in (3.1) is optimal in the sense that no other classification rule will have a smaller true error rate:

$$L(h^*) \leq L(h)$$

for any $h$ or $L$ (Wasserman 2003 p. 352). So the Bayes classification rule is the optimal classifier but depends on the unknown parameter set $\theta = \{\pi, f(x|Y = y); y = 0, 1\}$. Since the elements of $\theta$ are difficult to estimate, several estimation methods have been proposed. The estimation methods can be grouped into two rough approaches:

1. Density estimation: estimate $f(x|Y = 1)$ by $\hat{f}(x|Y = 1)$ and $f(x|Y = 0)$ by $\hat{f}(x|Y = 0)$; or

2. Regression: estimate $r(x)$ in (3.1) by $\hat{r}(x)$ using some type of regression directly without estimating $f(x|Y = 1)$ and $f(x|Y = 0)$ separately.
Density estimation for the multivariate densities \( f(x|Y = y) \), \( y = 0, 1 \) may conveniently, though often incorrectly, be done by assuming the multivariate normality of \( x \). In fact, when we assume the multivariate density functions \( f(x|Y = y) \), \( y = 0, 1 \) are multivariate normal with covariance matrices \( \Sigma_y \) for \( y = 0, 1 \) and \( \Sigma_0 = \Sigma_1 \), we are in the well-known realm of linear discriminant analysis. Quadratic discriminant analysis is also based on the assumption of multivariate normality, but the covariance matrices are assumed to be unequal. Discriminant analysis is well-known in credit-risk scorecards and was introduced by Altman in 1968, who used the term 'Z-scores' (Altman 1968). Unfortunately, the assumption of normality is often unrealistic and restrictive. The high dimensionality of the density functions could also pose a problem. This problem can be overcome by reducing the dimension of \( X \) by replacing \( X \) with a linear combination of its components, say \( V = \sum_{i=1}^{p} w_i X_i \), where \( (w_1, w_2, ..., w_p) \) are weights chosen to "separate the data best" (Wasserman 2003 p. 355). Here a \( p \)-dimensional problem is reduced to a one-dimensional problem in a technique known as Fisher’s linear discriminant analysis. The density functions \( f(x|Y = y) \), \( y = 0, 1 \) may also be estimated by non-parametric methods such as multivariate kernel density estimation (Silverman 1986). Should a data set have an immense size and a large number of covariates, which is typical in the case of retail credit risk, this approach, known as non-parametric discriminant analysis, does not seem to be computationally viable. Another density estimation method is to assume that the elements of the predictor matrix \( X \) are independent. The estimation of \( f(x|Y = y) \) is now reduced to estimating the product of the conditional marginal density functions \( \prod_{i=1}^{p} f(x_i|Y = 1) \). The conditional marginal densities can be estimated by non-parametric kernel density estimation. This approach is known as the naïve Bayes classifier.

The second, often-used classification approach is regression modelling. Here, the parameter set \( \theta \) may be estimated by either a simple linear generalized linear model (GLM) or other variants such as generalized estimation equations (GEE) or GLMM regression. In contrast to the density estimation procedures discussed above, here we estimate \( r(x) = E[Y|X = x] \) directly without estimating the conditional densities \( f(x|Y = y) \), \( y = 0, 1 \). Additionally, we do not make any normality, covariance or independence assumptions. Linear regression is not ideal, since the response variable has range \((-\infty, \infty)\),...
but may result in sufficient classification and is, surprisingly, still widely used. The response variable of logistic regression, a GLM, has range \((0, 1)\), which makes it suitable and popular for classification.

In comparing discriminant analysis and logistic regression, Press and Wilson (1978) found that parametric discriminant analysis is better suited to discriminate data if the data has a multivariate normal distribution. Otherwise, logistic regression is preferred. Logistic regression has the advantage of being much easier to interpret than discriminant analysis. The disadvantage of logistic regression is that a cut-off value \(b\) in (3.1) needs to be estimated, which may be very subjective.

We prefer GLMs to discriminant analysis, which has restrictive assumptions such as the normality of the covariates. We have identified GLMs as the most promising methodology for classification. Hence, the focus of the rest of this chapter will be solely on GLMs. We will now review some relevant GLM literature and also extend some diagnostic tests used in GLMs. Should the reader be familiar with GLMs, we recommend skipping the next section and continuing directly with Section 3.3.

### 3.2 Generalized linear models

This section contains an overview of generalized linear models (GLMs). The generalized linear model framework is a general framework that encompasses, among other things, linear regression, logistic regression, analysis of variance, count models and survival data models. Nelder and Webberburn (1972) propose this framework and McCullagh and Nelder (1989) discuss it in detail. GLMs with binary response variables seem to hold the most potential among classification techniques where binary classification is required.

Cross-sectional data analysis and GLMs assume that observations are independent. As a result, time series data with some auto- or cross-correlation is not addressed by GLMs. Although credit-risk data is correlated over time, we need a thorough understanding of GLMs to extend this framework to use in our credit-risk model. The independence assumption may be relaxed by the use of quasi-likelihood or using generalized estimating equations (GEE’s). Another assumption in GLMs is that there is only one random ‘error’ term. This assumption may be relaxed by including a random covariate (or latent covariate, as discussed earlier) through the use of generalized linear mixed models.
In this section we follow the GLM presentation by McCullagh and Nelder (1989), Dobson and Barnett (2008) and Gill (2000). We do not derive, but rather highlight and reference useful results. Define the \((n \times 1)\) vector \(y\) as a realization of an independent random variable \(Y\) which we now define to be any real number, i.e. \(Y \in \mathbb{R}^n\). Although, strictly, classification techniques are only concerned with \(Y \in \mathbb{Z}^n\), we discuss the real number case here and address our binary outcome interest, \(Y \in \{0,1\}\), later.

GLMs are an extension of classical linear regression. They attempt to loosen some of the restrictive assumptions governing linear regression. Basic assumptions for linear regression are that (1) there is a linear relationship between the response and covariates, (2) that residuals are independent with mean zero and a constant variance and (3) that no correlation exists between any explanatory variable and the error term. GLMs eliminate the first two assumptions and hence provide a method to analyse data in a way closely related to methods used in linear regression (Gill 2000 p. 7). Hence, the only assumption that still prevails is that there exists no correlation between any explanatory variable and the error term. Similarly to linear regression, we still encounter systematic- and random element. A model attempts to explain the systematic effects found in the data. The error term is seen as the random or unexplainable effect. The objective is to build a model that explains the most systematic variation in the data, while having a small random effect. GLMs consist of three components:

1. The random or stochastic component, \(Y\), is independently distributed according to any distribution that is part of the so-called exponential distribution family. \(Y\) has expected value \(E[Y] = \mu\).

2. The systematic component is the linear predictor \(\eta\) defined as \(\eta = \mathbf{X}\beta\), where \(\mathbf{X}\) is the \((n \times p)\) design matrix of covariates and \(\beta\) is a \((p \times 1)\) parameter vector.

3. The link function is some invertible, smooth function \(g(\mu) = \eta = \mathbf{X}\beta\) that links the systematic and random components of the model. The covariates \(\mathbf{X}\) only affect the response variable \(Y\) through the link function \(g\).

So, a GLM is of the form

\[ g(\mu) = \eta, \]
with the emphasis being that the linear predictor is connected to the response variable’s *mean value* via the link function. In other words, the response variables are not linked directly to the observed response variable (as is the case for linear regression) but rather to the expected value. To estimate the parameters of a GLM we use maximum-likelihood, although Monte Carlo Markov chains are also used by Bayesians. Before we estimate the parameters, we introduce the exponential family of distributions.

GLMs rely on the fact that the outcome variable is a part of the exponential distribution family. Several probability distributions can be expressed as a special case of the exponential family of distributions. Assuming we have some random variable $Y$ with one unknown parameter $\zeta$, it is part of the exponential family if its density function can be expressed as

$$f_Y(y|\zeta) = \exp \left[ a(y)d(\zeta) - g(\zeta) + h(y) \right],$$

where the functions $a, d, g$ and $h$ are known. A more useful representation of this density function, as will become apparent later, is the ‘canonical form’ (for the random variable $Y$) when $a(y) = y$. So,

$$f_Y(y|\zeta) = \exp \left[ yd(\zeta) - g(\zeta) + h(y) \right]. \tag{3.3}$$

We can further simplify this canonical form by rewriting it in the canonical form of the parameter $\zeta$. This can be done by forcing the transform $\theta = d(\zeta)$ to obtain the formula

$$f_Y(y|\theta) = \exp \left[ y\theta - b(\theta) + c(y) \right], \tag{3.4}$$

where both $b(\theta)$ and $c(y)$ are obtained by substituting $\theta = d(\zeta)$ into $g(\zeta)$ and $h(y)$ respectively (Gill 2000 p. 12). Several other presentations of the exponential family exist, with some accounting for two (often a scale parameter) or more parameters. The results that follow refer only to the canonical presentation (3.4). The canonical or standard form is useful when deriving the moments of the distribution. The first moment of $Y$ plays a crucial role in generalized linear models. The mean of the exponential family of distributions is well-known and given, according to Dobson and Barnett (2008 p. 49), by

$$E[Y] = \frac{\partial}{\partial \theta} b(\theta) := \mu. \tag{3.5}$$
The variance of $Y$ plays a role in the estimation of the parameters of the GLM. The variance of $Y$, according to Gill (2000 p. 26) is

$$V ar[Y] = \frac{\partial^2}{\partial \theta^2} b(\theta) := V.$$ (3.6)

We are now able to estimate the parameters $\beta$ for the GLM using maximum-likelihood estimation. Assume we observe a response variable $y_i$ and corresponding covariates $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,p})$ for each obligor $i = 1, 2, ..., n$. The log-likelihood for a single response $y_i$ is

$$lik_i(\theta|y_i) = y_i \theta_i - b(\theta_i) + c(y_i).$$

To calculate the partial derivative of the likelihood with respect to $\beta_j$ for $j = 1, 2, ..., p$ we use the chain rule. The partial derivative of the log-likelihood for observation $i$ with respect to parameter $j$ gives

$$\frac{\partial lik_i}{\partial \beta_j} = \frac{\partial lik_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j},$$ (3.7)

where $\frac{\partial lik_i}{\partial \theta_i} = y_i - \partial b(\theta_i) / \partial \theta_i = y_i - \mu_i$, $\frac{\partial \theta_i}{\partial \mu_i} = (\partial^2 b(\theta_i) / \partial \theta_i^2) = V_i^{-1}$ from (3.6) and $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$. Summing (3.7) over all $i$ we get the score function $U_j$ for parameter $j$:

$$U_j = \sum_{i=1}^{n} (y_i - \mu_i) \frac{x_{ij} V_i^{-1} \partial \mu_i}{\partial \eta_i},$$

$$= \sum_{i=1}^{n} (y_i - \mu_i) x_{ij} W_i \frac{\partial \eta_i}{\partial \mu_i},$$

where

$$W_i = \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 V_i^{-1}.$$ 

To find the null point of the score function we use Fisher scoring where the $(j, k)^{th}$ element of the $(p \times p)$ information matrix is given, according to Dobson and Barnett (2008 p. 65)
\[
\mathcal{I}_{jk} = E \left[ \frac{\partial l_{jk}}{\partial \beta_j} \frac{\partial l_{jk}}{\partial \beta_k} \right] 
\]
\[= \sum_{i=1}^{n} x_{ij} x_{im} W_i, \tag{3.8} \]

where \(j, k = 1, \ldots, p\). Using matrix notation, let \(\mathbf{U}\) be the \((p \times 1)\) score vector, \(\mathbf{I}\) the \((p \times p)\) information matrix with elements \(\mathcal{I}_{jk}\) and \(\hat{\beta}\) the \((p \times 1)\) parameter estimate of \(\beta\). Fisher scoring yields the iterative equation

\[
\hat{\beta}^{m+1} = \hat{\beta}^m - \mathbf{I}^{-1} \mathbf{U},
\]

with both \(\mathbf{I}\) and \(\mathbf{U}\) evaluated at the \(m\)th iteration estimate \(\hat{\beta}^m\). From (3.8) it follows that we may write

\[
\mathbf{I} = \mathbf{X}^\top \mathbf{W} \mathbf{X},
\]

where \(\mathbf{W}\) is a \((n \times n)\) diagonal matrix with diagonal elements \(W_i\). If \(\mathbf{G}\) is the \((n \times n)\) matrix with diagonal elements \(\partial \eta_i / \partial \mu_i\) the score vector is

\[
\mathbf{U} = \mathbf{X}^\top \mathbf{W} \mathbf{G} (\mathbf{y} - \mathbf{\mu}).
\]

So

\[
\hat{\beta}^{m+1} = \hat{\beta}^m - (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{U} \tag{3.9}
\]

\[
= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{X} \hat{\beta}^m + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{U}
\]

\[
= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \left( \mathbf{X}^\top \mathbf{W} \mathbf{X} \hat{\beta}^m + \mathbf{X}^\top \mathbf{W} \mathbf{G} (\mathbf{y} - \mathbf{\mu}) \right)
\]

\[
= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \left( \mathbf{X} \hat{\beta}^m + \mathbf{G} (\mathbf{y} - \mathbf{\mu}) \right)
\]

\[
= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{z}, \tag{3.10}
\]
with

\[ z = X\hat{\beta}^m + G(y - \mu) \]

\[ = \eta + G(y - \mu), \]

where \( W, \mu, G \) and \( \eta \) are all evaluated at \( \hat{\beta}^m \) (McCullagh & Nelder 1998 p. 43). From equation (3.10) we see that the estimation process is simply the solution to the efficient iterative reweighted least squares problem (IRLS), by iteratively finding \( \hat{\beta}^{m+1} \) until \( \| \hat{\beta}^{m+1} - \hat{\beta}^m \|^2 \) converges to some very small number.

An example

Throughout this chapter we will illustrate concepts by using a sample of about 3.5% of the data. Although this is a small percentage, the sample size is still a large 4000 obligors. We attempt to illustrate a typical binary response classification problem with cross-sectional data. In order to prevent longitudinal data from appearing, we took the sample on one specific month. Credit-risk data with credit events as response variables, are realizations of Bernoulli random variables. Hence, we set the vector \( y \) to be a realization of an independent Bernoulli random vector \( Y \in \{0, 1\}^n \). To model the probability of an event occurring, the outcome variables should be bounded between zero and one. We model the probability of \( n \) Bernoulli random variables, each with individual success probability \( \pi_i \). Many suitable link functions exist for binomial and Bernoulli random variables, although popular link functions are:

1. \( g(\pi) = \log [\pi/(1 - \pi)] \), known as the logit function;

2. \( g(\pi) = \Phi^{-1}(\pi) \), where \( \Phi^{-1}(\ ) \) the inverse normal-density function, known as the probit function or

3. \( g(\pi) = \log (\log(1 - \pi)) \), known as the complementary log-log or cloglog function.

The differences between these link functions are apparent in the tails of the distributions: the cloglog function has longer tails. We use the logit function here as it is the
canonical link to a Bernoulli random variable. We may rewrite the Bernoulli distribution in the exponential family distribution form given in (3.4) as

\[ f_Y(y|\theta) = \pi^y (1 - \pi)^{1-y} \]

(3.11)

\[ = \exp \left[ y \log \left( \frac{\pi}{1 - \pi} \right) - \left( - \log (1 - \pi) \right) \right], \]

where the canonical link function of the parameter is \( \theta = \log \left( \frac{\pi}{1 - \pi} \right) \). Substituting the inverse of the canonical link function into (3.11) we get \( b(\theta) = \log (1 + \exp(\theta)) \). Hence, in canonical form the binomial distribution is

\[ f_Y(y|\theta) = \exp \left[ y\theta - \log (1 + \exp(\theta)) \right]. \]

A Bernoulli random variable’s mean and variance can be confirmed by applying (3.5) and (3.6) respectively. Observing \( y \) and corresponding design matrix \( x \), the solution to the iterative least squares equation in (3.9) is

\[ \hat{\beta}^{m+1} = (X^T WX)^{-1} X^T Wz, \]

where \( W \) is a \((n \times n)\) diagonal weight matrix with entries \( \text{diag}(\pi_i (1 - \pi_i)) \) for \( i = 1, 2, ..., n \). The response variable \( z \) is

\[ z = X\hat{\beta}^m + W^{-1} (y - \pi), \]

where, in this case, \( G_{ii} = \partial \eta_i / \partial \mu_i = (\pi_i (1 - \pi_i))^{-1} = W_{ii}^{-1} \) and \( \mu = \pi \), all evaluated at \( \hat{\beta}^m \). Lastly we may also write \( \hat{\beta} \), the IRLS parameter estimate of \( \beta \), as

\[ \hat{\beta} = \arg \min_{\beta} \left\{ (z - X\beta)^T W (z - X\beta) \right\}. \]  

(3.12)

\( \Delta \)
Having now estimated the GLM parameters we now need to decide which covariates to include in the model.

### 3.3 Variable selection

Thus far we have introduced GLMs and discussed a method to estimate the parameters of a GLM model. In this section we describe some variable selection techniques. The need to reduce the number of covariates is crucial, and is especially relevant when the number of covariates is large (as is common in credit-risk modelling). There are several techniques for variable selection, or dimension reduction techniques. These techniques are all mostly aimed at improving interpretability: models should ideally have a small number of predictors that describe the strongest systematic effects. These techniques then aim at maximizing the model’s prediction accuracy while simultaneously minimizing the number of predictors. This results in a parsimonious model. Classical automatic variable selection methods are: stepwise regression, backward regression, best subset regression and variations of the aforementioned methods. Automatic model selection methods have been criticized for sometimes selecting an erroneous model as the ‘best’ model, while other models may also be ‘good’. When the selection criteria used to select the most significant covariates are different from the criteria used to measure prediction accuracy, selecting a ‘good’ model with ‘bad’ predictability becomes much more likely.

Normally the best model is chosen in a way that minimizes some discrepancy measure like mean-squared errors ($MSE$). Frequently used model-fit statistics, such as the Akaike (1974) information criterion (AIC) and the Bayesian information criterion (BIC) can be written as functions of the $MSE$ for linear models. The AIC and BIC penalize the $MSE$ by some function of the number of predictors. We know that the $MSE$ of an estimator $\hat{\theta}$ of $\theta$ is simply $MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta})^2$, where $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$. Hence, the $MSE$ can be decreased by either decreasing the variance or by decreasing the bias. Thus, requiring an estimator to be unbiased, i.e. with $Bias(\hat{\theta}) = 0$, could result in a higher $MSE(\hat{\theta})$ than would be obtained if an unbiased model was of less importance. In other words, a biased estimator may have a much smaller variance than an unbiased estimator, ultimately reducing the $MSE(\hat{\theta})$. As a result a stable estimator (with low variance) could be preferable over an unbiased estimator. This is, of course,
at the cost of increasing the bias of the estimator. The unbiased ordinary least squares (OLS) estimates in linear regression may be 'shrunk' or set equal to zero: this increases the model bias, but potentially improves the prediction accuracy. Effectively, we want a less complex model, i.e. one with lower variance, at the cost of increasing the model bias. Variable selection methods that shrink the value of the unbiased estimators are known as shrinkage models.

An innovative shrinkage approach developed by Tibshirani (1996) is the 'least absolute shrinkage and selection operator' (lasso) approach. This approach shrinks coefficients to be equal to zero or less than their unbiased estimates, resulting in some model bias, but a lower model $MSE$. Best subset selection will also set some coefficients equal to zero, but is very unstable: some predictor variables will be included in a $p^*$-variate model, but may be removed in a $(p^* + 1)$-variate model. Hence, the prediction accuracy is volatile. Ridge regression, a well-known shrinkage model, shrinks coefficients but never sets the predictors equal to zero completely, so a complex model still remains. The lasso can be seen as a method that takes advantage of the properties of both subset selection and ridge regression. The lasso sets some coefficients equal to zero, while producing a model with high prediction accuracy. The lasso "does a kind of continuous subset selection" (Hastie et al. 2009). Using the same binary response variable and outcome variables as before, the lasso estimate $\hat{\beta}_t^{\text{lasso}}$ for logistic regression is the estimated parameter value of $\beta$ that maximizes the log-likelihood, $lik(\beta)$, subject to the so-called $l^q$ norm, $q > 0$

$$\hat{\beta}_t^{\text{lasso}} = \arg\max_\beta \{lik(\beta)\}$$

$$= \arg\max_\beta \left\{ \sum_{i=1}^n [Y_iX_i\beta - \log(1 + \exp(X_i\beta))] \right\}$$

subject to

$$\sum_{j=1}^p (|\beta_j|^q)^{1/q} \leq t,$$

where $t$ is the so-called 'tuning' or 'lasso' parameter. The lasso has $q = 1$, or uses the $l^1$-norm. Tibshirani (1996) uses three methods to estimate the tuning parameter $t$ for linear models: (1) cross-validation, (2) generalized cross-validation and (3) an analytical unbiased estimate of risk. The analytical unbiased estimate of risk is known as Stein's
unbiased risk estimator (Stein 1981). Stein’s method relies heavily on the fact that the design matrix is orthogonal, i.e. \( \mathbf{X}^\top \mathbf{X} = \mathbf{I} \), where \( \mathbf{I} \) is the identity matrix. This assumption is restrictive and usually not applicable, so we only focus on data-driven methods namely, cross-validation and generalized cross-validation. The lasso was originally only developed for linear models, and we spent quite some time on extending the generalized cross-validation technique to estimate the tuning parameter for logistic regression (see the appendix at the end of this chapter). This has, to our knowledge, not been done before. Eventually, the computing time for the estimation of the tuning parameter was found to be very high and was only really applicable to small data sets, so we did not find it useful in our application. This finding is in line with the critique of the lasso by Lin et al. (2011). An extension of the penalization function to include the \( l^2 \)-norm \((q = 2)\) is the approach used by Zou and Hastie (2005), which is called the ‘elastic net’. In essence, the elastic net is a combination of the lasso and ridge regression. Tuning parameters also need to be estimated, but they use an effective variant of least angle regression (LARS) developed by Efron et al. (2004) to speed up estimation. More recently, Lin et al. (2011) introduced the idea of variance inflation (VIF) regression. Their VIF regression technique does not require any tuning parameter to be estimated and is very fast. They show that the lasso and other \( l^1 \)-norm algorithms lead to biased estimates (which is expected) with large out-of-sample mean-squared errors (which is not desirable). The VIF regression methodology, according to Lin at al. (2011), is "in fact an improved, much faster version of stepwise regression".

Since we have a limited number of covariates, we used forward step-wise regression to perform all variable selection in this thesis. We would recommend that VIF regression be considered for use if data sets are large and a large number of covariates is available.

### 3.4 Model checking

"All models are wrong, some are useful" - G.E.P. Box

When assessing a model’s adequacy, we primarily assess whether the systematic part of the model has been identified correctly. Additionally, we check the assumed error structures, the existence of outliers and influential points (Cox & Snell 1990 p. 70).
Generally, model adequacy can be checked by:

1. Performing a goodness-of-fit test, which is a summary measure of the distance between the observed values of \( Y \) and the fitted values of \( Y \);
2. Comparing the model against specific modified or extended alternative models; and
3. Performing diagnostic tests or analysing the residuals: the individual residuals for the pairs \((Y_i, \hat{Y}_i)\) should be small relative to the model’s error structure and be unsystematic.

In this section we are again concerned with the checking of models with a binary outcome. We discuss the inefficiency of goodness-of-fit measures, diagnostic tests and measures of discriminatory power in the following sections. We choose to focus on logistic regression, but the methods described here also apply to binary outcome models with other GLM link functions. Unlike simple linear regression models, logistic regression models do not assume to have a constant error term variance. As a result, they do not fit well into the residual-based paradigms for model assessment (Cook and Weisberg 1997). In the next sections we address this issue and propose improved methods for analysing logistic model adequacy.

**Example continued**

Continuing the example in Section 3.2, to predict whether an account will increase in delinquency from one month to the next, the following method is used. Denote the \((n \times 1)\) response column vector by \( Y \). Each element, \( Y_i \), is an indicator function that indicates whether obligor \( i \) failed to pay the installment due (resulting in an increase in delinquency). We fitted a logistic regression model to predict the outcome of \( Y \). We arbitrarily chose the covariates summarized in Table 3.1. Without a loss of generality we will omit the time index \( t \), since this is cross-sectional data. The parameter estimates are summarized in Table 3.2. All parameter estimates were significantly different from zero at a level less than 0.001.
Table 3.1: The covariates of the logistic model.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TD</td>
<td>The total number of months an account was delinquent over the previous six months (standardized)</td>
</tr>
<tr>
<td>D</td>
<td>The previous months’ delinquency level (standardized)</td>
</tr>
<tr>
<td>Age</td>
<td>The age of the loan (transformed &amp; standardized)</td>
</tr>
</tbody>
</table>

Table 3.2: The parameter estimates for the logistic regression model.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\beta$</th>
<th>$SE(\beta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.3327</td>
<td>0.0711</td>
</tr>
<tr>
<td>TD</td>
<td>0.4322</td>
<td>0.0579</td>
</tr>
<tr>
<td>D</td>
<td>0.1836</td>
<td>0.0485</td>
</tr>
<tr>
<td>Age</td>
<td>-0.3010</td>
<td>0.0814</td>
</tr>
</tbody>
</table>

3.4.1 Goodness-of-fit tests

In regression analysis, we can fit a model by minimizing some discrepancy or distance measure, such as the mean-squared method, which minimizes $\sum_i \left( Y_i - \hat{Y}_i \right)^2$. The ‘closeness’ of the observed and expected values are used to test the goodness of fit. In a correctly specified model the fitted value $\hat{Y}$ is ‘close’ to $Y$.

The goodness of fit of the model can be assessed by comparing observed and expected event counts. The discrepancy between these counts can be reduced by including any number of additional parameters to the postulated model. In fact, a perfect fit can be obtained by setting the number of parameters equal to the number of observations, $p = n$. This is known as the full or saturated model $\hat{Y}_i = Y_i$. Here, the full model is totally un parsimonious and merely estimates the response variables by using the observed values. On the other extreme, we can fit a model with a single parameter in the minimal model. In the minimal model all variation between the response variables is caused by a random effect, while in the full model all variation is caused by the systematic component.

Although we use either the full or minimal model during hypothesis testing, neither is especially useful for prediction. Hence, we strive to obtain a model that lies between the saturated and minimal models. This model should be a parsimonious model that can improve predictability.

The goodness-of-fit test is a single summary value that indicates whether the expected values accurately represent the observed values.

49
For logistic regression we test the hypothesis

\[ H_0 : \quad E[Y] = (1 + \exp(-X\beta))^{-1} \quad (3.13) \]

vs.

\[ H_1 : \quad E[Y] \neq (1 + \exp(-X\beta))^{-1}. \]

A poor fit may be the result of inappropriate link functions, an inappropriate variance function, omitted covariates or the wrong transformation used on the covariates (McCullagh & Nelder 1989 p. 211). Unfortunately, the goodness-of-fit model test do not provide us with any guidance on the cause of a rejected null hypothesis (Cox & Snell 1990 p. 71). Additionally, a high goodness-of-fit statistic value would indicate substantial lack of fit, but a low value would not necessarily rule out the possibility of a severe lack of fit for a few observations.

The most well-known goodness-of-fit statistics are the Pearson’s chi-squared, \( X^2 \) and the deviance statistic, \( D \). The latter is also known as the likelihood ratio statistic and is sometimes denoted by \( G^2 \). The deviance goodness-of-fit test is analogous to the linear lack-of-fit test. These two statistics rely heavily on repeated measures or repeating covariate patterns (typically found in binomial experiments). Repeated covariate patterns are sets of covariate values that are exactly the same for each observation within a set. Assume we have \( n \) observations and \( c \) unique covariate patterns \( X_j, j = 1, 2, ..., c \). Further assume there are \( n_j \) observations in each covariate pattern group, that the total number of observations is \( n = \sum_{j=1}^{c} n_j \) and that \( \hat{\pi}_j \) is the expected value of an observation with covariate pattern from group \( j \). Let \( y_j \) denote the number of observed successes in the \( j^{th} \) group.

Pearson’s chi-square statistic is defined as

\[
X^2 = \sum_{j=1}^{c} \frac{(y_j - n_j\hat{\pi}_j)^2}{n_j\hat{\pi}_j(1 - \hat{\pi}_j)}.
\]

\( X^2 \) can be shown to be asymptotically equivalent to the deviance statistic (Dobson &
Both Pearson’s statistic and the deviance statistics are (assuming that the null hypothesis is true) asymptotically \( \chi^2_{(c-p-1)} \) distributed. This asymptotic distribution follows from what is known as \( n \) asymptotes: the number of covariate patterns stays fixed, but \( n \to \infty \).

The so-called \textit{classical fixed-cell asymptotic} approach relies on the following assumptions (Osius & Rojek 1992):

1. The sample size increases to infinity (\( n \to \infty \));
2. The number of covariates (\( p \)) stays fixed;
3. The number of outcomes stays fixed (in logistic regression there are two outcomes);
4. The number of covariate patterns (\( c \)) stays fixed; and
5. The observations with a certain covariate pattern increase (\( n_j \to \infty \)). Hence, the expected frequencies increase \( n_j \hat{\pi}_j \to \infty \).

Should these six properties hold, we may use \( n \) asymptotics to derive a parametric test for the goodness of fit. However, the fourth assumption is mostly violated. If there is only a small number of observations in each covariate pattern \( j \), we are dealing with what is called a ‘sparse data’ problem. Sparse data problems are encountered very often, especially if at least one of the covariates is continuous (Kuss 2002). Hosmer and Lemeshow (2000) state that: “The case where \( c = n \) is most frequently encountered in practice. It also presents the greatest challenge in developing distributions of goodness-of-fit statistics.”

If the model contains at least one continuous covariate, then \( c \to n \). Therefore, the \( \chi^2_{(c-p-1)} \) distribution of the test statistic is inappropriate. McCullagh and Nelder (1989 p. 121) state that in the case where \( n_j = 1 \) for all \( j \) (known as binary logistic regression), the deviance function simplifies to

\[
D = 2 \sum_{j=1}^{c} \left[ y_j \log \left( \frac{y_j}{n_j \hat{\pi}_j} \right) + (n_j - y_j) \log \left( \frac{n_i - y_i}{n_i (1 - \hat{\pi}_i)} \right) \right].
\]
where \( \hat{\pi}_i = \left(1 + \exp(-X_i^\top \hat{\beta})\right)^{-1} \) for all \( j \) and \( \hat{\beta} \) is given. \( D \) is asymptotically a degenerate distribution. We do not concern ourselves with the asymptotic distribution of either Pearson’s statistic or the deviance function, but suffice by concluding that neither is well suited as goodness-of-fit measure for sparse data.

Several tests have been proposed to address the sparse data problem. An intuitive approach was developed by Hosmer and Lemeshow (2000 p. 147). Their test groups data in one of two ways: (1) by grouping the expected values \( \hat{\pi}_j \) in percentiles, and (2) by grouping the values of \( \hat{\pi}_j \) according to fixed interval values (Hosmer et al. 2013 p. 157). After a grouping is made, the normal Pearson’s chi-squared test is performed, in which each group is seen as having the same covariate pattern. As a result, all observations that fall within a specific group are assumed to have a similar covariate pattern. Hosmer and Lemeshow find the test statistic

\[
G^2_{HL} = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{E_j(1 - E_j/N_j)}
\]

(3.15)

to be \( \chi^2_{(g-2)} \) distributed, where \( g \) is the number of groups, \( O_j = \sum_i y_{ij} \) is the observed number of successes in group \( j \), \( E_j = \sum_i \hat{\pi}_{ij} \) is the expected number of successes in group \( j \) and \( N_j \) is the number of observations in group \( j \). They did not derive the test statistic’s null distribution rigorously, but rather did so by extensive simulation. The Hosmer-Lemeshow test statistic is widely used, since it is easily implemented and understood. The deficiencies of the test statistic are that the limiting null distribution is not rigorously derived and the test is a conservative test with low power to detect specific lack-of-fit types. The Hosmer-Lemeshow statistic is highly dependent on the number of groups and the algorithms that determine the groups’ cut-off points. Other deficiencies include that the test will almost always indicate a good fit for small values of \( g \), (Xie et al. 2008). Hosmer et al. (1997) found that when they analysed a data set using six different statistical computing packages, the \( p \)-values of the Hosmer-Lemeshow test statistics were unique, varying from 0.02 to 0.159.

Osius and Rojek (1992) derived a goodness-of-fit test that has a test statistic that
is asymptotically normally distributed. Another test we may perform is to compare the fitted model to the minimal model by using likelihood ratio hypothesis tests (Dobson & Barnett 2008 p. 137). This test statistic is still asymptotically chi-square distributed and will facilitate the comparison of different nested models or the testing of specific hypotheses. Measures like the Akaike information criterion, the Bayesian information criterion and the pseudo $R^2$ may also serve as goodness-of-fit statistics and will be useful when comparing different nested models (Dobson & Barnett 2008 p. 137). Several other goodness-of-fit tests have been developed: see Kuss (2002) for a survey. After comparing several goodness-of-fit models, Kuss concludes that a $p$-value that is greater than the significance level in the hypothesis test (3.13) "does not tell you that your model is correct, it just tells you that the lack-of-fit is not large enough for you to reject your model".

Example continued

Continuing the example in Section 3.2, we tested the goodness-of-fit of our sparse data problem by using the Hosmer-Lemeshow test with $j = 10$ (each group had 400 observations). This corresponds to grouping the expected values $\hat{\pi}_j$ in percentiles. The goodness-of-fit test rejected the null hypothesis in (3.13) with $G^2_{HL} = 39.914$ and a $p$-value of $3.3234 \times 10^{-6}$. The contingency table is given in Table 3.3.

Table 3.3: Contingency table for the Hosmer-Lemeshow test.

<table>
<thead>
<tr>
<th>Group</th>
<th>Max $Y$</th>
<th>Group size</th>
<th>$G_j$</th>
<th>$E_j$</th>
<th>$O_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0684</td>
<td>400</td>
<td>2.1253</td>
<td>21.609</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>0.0709</td>
<td>400</td>
<td>2.1499</td>
<td>27.964</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>0.0732</td>
<td>400</td>
<td>1.1343</td>
<td>28.87</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>0.0766</td>
<td>400</td>
<td>1.1354</td>
<td>29.979</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>0.083</td>
<td>400</td>
<td>0.8858</td>
<td>31.792</td>
<td>27</td>
</tr>
<tr>
<td>6</td>
<td>0.0931</td>
<td>400</td>
<td>0.7114</td>
<td>34.981</td>
<td>39</td>
</tr>
<tr>
<td>7</td>
<td>0.1201</td>
<td>400</td>
<td>1.2832</td>
<td>42.122</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>0.1672</td>
<td>400</td>
<td>3.5873</td>
<td>56.089</td>
<td>81</td>
</tr>
<tr>
<td>9</td>
<td>0.279</td>
<td>400</td>
<td>2.7071</td>
<td>85.778</td>
<td>108</td>
</tr>
<tr>
<td>10</td>
<td>0.551</td>
<td>400</td>
<td>2.2506</td>
<td>150.82</td>
<td>129</td>
</tr>
</tbody>
</table>

Inspecting the error terms ($G_j$) in Table 3.3, we see that large discrepancies exist for several groups. In Figure 3-1 we depict the error terms $G_j$ for $j = 1, \ldots, g$. The
horizontal axis has the maximum $\hat{Y}_j$ per group $j$. Here we can clearly see several groups have residual values greater than (a heuristically chosen) 2.5, indicating that there is a problem in the fit in those groups. A heuristic graphical goodness-of-fit test (very similar to the Hosmer-Lemeshow test) can be obtained by grouping the $\hat{\pi}_j$ values by fixed interval values. Assume that the intervals are given by $[\hat{\pi}_0, \hat{\pi}_1), [\hat{\pi}_1, \hat{\pi}_2), ..., [\hat{\pi}_{g-1}, \hat{\pi}_g]$. Then the midpoints of the intervals given by $\hat{\pi}_j^{\text{mid}} = (\hat{\pi}_j - \hat{\pi}_{j-1})/2$ for $j = 1, ..., g$ can be plotted against the observed default percentage for obligors whose estimated default probabilities fall within each group. Should this scatter plot yield a straight line with a unit slope and zero intercept then (heuristically) the model has perfect fit. Ideally, the scatter plot’s points should be close to the straight line or exhibit a random scatter around the line.

In our example we split the data into 10 groups with equally spaced $\hat{\pi}_j$ values. The contingency table is given in Table 3.4. The scatter plot of the midpoints, $\hat{\pi}_j^{\text{mid}}$, against the observed default frequency given by $\hat{\pi}_j^{\text{obs}} = O_j/N_j$ is given in Figure 3-2. Since some intervals contain more observations than others, we draw the size of the bubble relative to the size $N_j$. Clearly the model exhibits some deviation from the ideal line (unit slope

Figure 3-1: The $G_j^{1/2}$ errors for the ten groups in the Hosmer Lemeshow test. Every group with an error term above 2.5 (a heuristic number) should be investigated.

54
and zero intercept) in red. The second to sixth dots seem to have a quadratic trend, which is not what would be expected from a model that fits the data well. This curved trend typically hints at the omission of second-order terms. The small dot on the extreme lower right does not carry much weight as it was only based on 17 observations (0.425% of the total number of observations).

Table 3.4: Contingency table for the heuristic visual presentation of the Hosmer-Lemeshow test.

<table>
<thead>
<tr>
<th>Group number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_j$</td>
<td>2098</td>
<td>901</td>
<td>294</td>
<td>222</td>
<td>115</td>
<td>121</td>
<td>105</td>
<td>76</td>
<td>51</td>
<td>17</td>
</tr>
<tr>
<td>$O_j$</td>
<td>113</td>
<td>118</td>
<td>63</td>
<td>59</td>
<td>37</td>
<td>39</td>
<td>31</td>
<td>31</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>$\hat{\pi}_j^{obs}$</td>
<td>0.054</td>
<td>0.131</td>
<td>0.214</td>
<td>0.266</td>
<td>0.322</td>
<td>0.322</td>
<td>0.295</td>
<td>0.408</td>
<td>0.353</td>
<td>0.059</td>
</tr>
<tr>
<td>$\hat{\pi}_j^{mid}$</td>
<td>0.059</td>
<td>0.111</td>
<td>0.163</td>
<td>0.215</td>
<td>0.267</td>
<td>0.318</td>
<td>0.37</td>
<td>0.422</td>
<td>0.474</td>
<td>0.526</td>
</tr>
</tbody>
</table>

Figure 3-2: The scatterplot of $\hat{\pi}_j^{mid}$ against $\hat{\pi}_j^{obs}$. The size of the bubbles are proportional to the number of observations ($N_j$) in each interval $j$.

Unfortunately the Hosmer-Lemeshow test, or any other goodness-of-fit test, will only indicate for what ranges of the predicted probability there will be a deviation from the predicted model, not what the problem is or how it should be solved.  

Δ
Rather than relying on either $D$, $X^2$, $G^2_{H,L}$ or any other goodness-of-fit test for sparse data, we should rather investigate specific deviations from the postulated models (McCullagh & Nelder 1989 p. 122). In the sparse data there are an extremely high number of observations. If $n \to \infty$, it is known that every postulated null hypothesis will be rejected. In asymptotic theory, the asymptotic power of a test should approach unity as $n \to \infty$, but in practice this is not the case. This is because of the violation of the asymptotic consistency that is assumed for hypothesis tests. Hence, should we have extremely large sparse data sets almost any goodness-of-fit hypothesis will be rejected. Composite hypothesis tests may be of some use here, but are subjective. Because we have extremely large data sets and we lack a sufficient goodness-of-fit test, we will, in the next section, discuss diagnostic tests that will be employed to perform model checking.

### 3.4.2 Diagnostic tests

Goodness-of-fit statistics provide a single measure to assess the fit of a model. The disadvantage of these statistics is that they could miss some deviations from the fit for a small, but sometimes crucial, number of observations. Diagnostic tests and an analysis of the residuals assist in identifying these deviations. In addition, if the goodness-of-fit test rejects the null hypothesis, no or little indication of why the hypothesis was rejected will be given. In contrast to formal goodness-of-fit measures, diagnostic tests are often graphical in nature. Summary statistics cannot replace the insight that a graph provides, as was shown by the well-known Anscombe’s quartet (Anscombe 1973). Hence, since formal goodness-of-fit tests are inadequate for our data, we use graphical diagnostic tests, which are inherently more heuristic in nature. The application of traditional linear residual analysis to binary data is inadequate. Checking the adequacy of a model can either be done by focusing on the model’s weaknesses or on the model’s strengths (Cook 1998 p. 303). In linear regression, residual plots are used to check the model weaknesses. Residual plots normally plot the residual of the model versus some linear combination of the predictors in the model.
The residuals for a binary model are

\[
\epsilon_i = Y_i - \hat{Y}_i = \begin{cases} 
1 - E(Y_i | X_i = x_i) & \text{if } Y = 1 \\
-E(Y_i | X_i = x_i) & \text{if } Y = 0 
\end{cases},
\]  

(3.16)

where

\[
\hat{Y}_i = E(Y_i | X_i = x_i) = \left(1 + \exp\left(-\beta^\top X_i\right)\right)^{-1}
\]

for obligor \(i = 1, 2, \ldots, n\). Residual plots for logistic regression cannot be interpreted as we would normally interpret residual plots for linear regression. In linear regression, the residuals are plotted against some linear combination, say \(Xa\), of the predictors \(X\). Should these plots exhibit some pattern, then the model is incorrectly specified. These plots should also give a visual verification of the underlying assumptions: whether the variance of the error term is constant and whether the errors are normally distributed. The choice of \(a\) was historically mostly a matter of prior information and taste (Cook 1998 p. 304). Plotting the unstandardized residual of logistic regression against \(Xa\) will never result in a random pattern. The residuals, or any variation of the residuals, produce highly non-random patterns, owing to the binary nature of the residuals in (3.16).

**Example continued**

Continuing the example in Section 3.2, in Figure 3-3 we plot the residuals, \(\epsilon_i\) against the \(Xa\), with \(a\) chosen to be the maximum-likelihood estimator \(\hat{\beta}\). To increase the clarity of the graph we perturbed the values of the residuals by adding a normally distributed random variable with small variance to the original residual values. Figure 3-3 has a trend that is artificial, since the response variables are either a one or a zero. Hence, approaching residual analysis in this manner is not useful when applying logistic regression to binary data.

The lack of interpretability of residual plots leads us to rather focus on the model strengths by using marginal-model plots.

### 3.4.3 Marginal-model plots

Several graphical methods of determining the adequacy of a model exist. The marginal-model plot is a promising methodology introduced by Cook and Weisberg (1997). They
find that marginal-model plots are especially useful for logistic regression where traditional residual analysis techniques fail. Marginal-model plots are used to identify incorrect covariate functional forms and, to some extent, the goodness of fit. For binary logistic regression (i.e. logistic regression with no repeating covariate patterns), generally all plots of the residual versus some linear combination of the covariates $Xa$ will be non-random. This property holds for most generalized linear models. We notice this property in the residual plot of our data example in Figure 3-3.

We follow Cook and Weisberg (1997) by attempting to infer the real unknown conditional density function $f(y|X)$ of the response variable $Y$ conditional on a realization of the covariates $X$. This density function defines what is sometimes referred to as the ‘model-free’ distribution (Pan et al. 2001). We assume that the observed data pairs $(Y_i, X_i)$ are independent for $i = 1, ..., n$. The conditional density of the postulated model, denoted by $m(y|\theta, X)$, is assumed to depend on an unknown parameter set $\theta$. In essence we must determine whether $m(y|\theta, X)$ adequately describes the data originating from $f(y|X)$. So, we need to determine how ‘close’ $f(y|X)$ and $m(y|\theta, X)$ are. Since these are
both multivariate densities, the estimation is complex. To simplify the estimation, Cook and Weisberg make use of what is called the *marginal-model-checking condition* to assess model adequacy.

**CONDITION (Marginal-model-checking condition)**

Two conditional distribution functions $G(y|X)$ and $H(y|X)$ are equal if and only if $G(y|Z) = H(y|Z)$ for all $Z = \sum_{j=1}^{p} a_j X_j$ with $\sum_{j=1}^{p} a_j^2 = 1$.

This condition can be derived by using characteristic functions of the two random variables $Y$ and $Y^*$. Below we use the standard notation $f_{Y|Z}(y|z)$ to denote the density function $f$ of $y$ given $z$.

\[
g_{Y|Z}(y|z) = h_{Y^*|Z}(y|z)
\]

\[\iff\]

\[
g_{Y,Z}(y,z) = h_{Y^*,Z}(y,z)
\]

\[\iff\]

\[
E \left[ e^{i(t_0 Y + t_1 Z)} \right] = E \left[ e^{i(t_0 Y^* + t_1 Z)} \right] \quad \forall \ t_0, t_1 \in \mathbb{R}^1
\]

\[\iff\]

\[
E \left[ e^{i(t_0 Y + (t_1 a_1)X_1 + \cdots + (t_1 a_p)X_p)} \right] = E \left[ e^{i(t_0 Y^* + (t_1 a_1)X_1 + \cdots + (t_1 a_p)X_p)} \right]
\]

\[\iff\]

\[
E \left[ e^{i(t_0 Y + \kappa_1 X_1 + \cdots + \kappa_p X_p)} \right] = E \left[ e^{i(t_0 Y^* + \kappa_1 X_1 + \cdots + \kappa_p X_p)} \right] \quad \forall \ (\kappa_0, \ldots, \kappa_p) \in \mathbb{R}^p
\]

\[\iff\]

\[
g_{Y,X_1,\ldots,X_p}(y,x_1,\ldots,x_p) = h_{Y^*,X_1,\ldots,X_p}(y,x_1,\ldots,x_p)
\]

\[\iff\]

\[
g_{Y|X_1,\ldots,X_p}(y|x_1,\ldots,x_p) = h_{Y^*|X_1,\ldots,X_p}(y|x_1,\ldots,x_p).
\]

\[\iff\]

\[
g_{Y|X}(y|x) = h_{Y^*|X}(y|x).
\]

Notice that $\sum_{j=1}^{p} a_j^2 = 1$ is a simple norming condition used to avoid replication of linear combinations. Using the marginal-model-checking condition we effectively reduce a $p$-dimensional problem to a one-dimensional problem. As a result, multivariate distributions can now be broken down into manageable parts. Henceforth, we will use the notation $M$ and $F$ to refer to the distribution functions of $m(y|\theta, X)$ and $f(y|X)$ respectively. We can make use of two-dimensional graphs to visually assess the model’s adequacy. In
other words, we use a linear combination \( Xa \) to assess the adequacy of \( M(Y|\theta, X) \) in describing \( F(Y|X) \). One way to do this is by comparing all the moments of \( M(Y|\theta, Xa) \) and \( F(Y|Xa) \) for all values of \( a \in \mathbb{R}^p \) with \( \|a\| = 1 \). This is a formidable task. Focusing only on the first two moments, the marginal-mean and marginal-variance functions, and some carefully chosen values of \( a \), we reduce the problem to a manageable size. We now discuss methods of choosing \( a \) and estimating the marginal-mean and -variance functions.

Cook and Weisberg (1997) propose five different methods of choosing \( a \) (\( a \) is sometimes referred to as the plotting direction). We pursue the following three ways: (1) choose the plotting direction \( a = \hat{\beta} \); (2) select \( a \) to be each predictor in turn, i.e. \( a \) should be a vector of zeros bar one position where it is a one (this will show individual covariate deficiencies or highlight potential transformations to be considered for covariates); (3) choose \( a \) as a random vector.

The marginal-mean function for the postulated model is derived by using the law of iterated expectations. To keep the notation concise we suppress the \( \theta \) from the notation. The marginal mean of \( Y \) given \( Xa \) is

\[
E_M[Y|Xa] = E[E_M(Y|X)|Xa]. 
\]

To calculate \( E_M(Y|Xa) \) exactly we need to know the marginal distribution of \( X \). This distribution will generally be unknown, and consequently from equation (3.17) we can estimate \( E_M(Y|Xa) \) by smoothing a scatter plot of the fitted model values \( \hat{E}_M(Y|X) \) against \( Xa \). That is, we smooth the scatter plot of the fitted values \( \hat{Y} \), which are a proxy for \( E_M[Y|X] \), against \( Xa \). We denote this smoothed estimator of \( E_M(Y|Xa) \) by \( \hat{E}_M(Y|Xa) \). We estimate the marginal mean of the non-parametric true distribution \( E_F(Y|Xa) \) by simply smoothing the scatter plot of the observed values \( Y \) against \( Xa \), and denoting the estimator by \( \hat{E}_F(Y|Xa) \).

We make use of non-parametric smoothers. Specifically, we use the loess smoothers proposed by Cleveland (1979) and developed further by Cleveland and Devlin (1988). Although many methods exist to estimate the smoothing or nearest-neighbour fraction (Silverman 1986), we follow Cook and Weisberg’s (1997) method by estimating it visually.

The marginal-model plot is obtained by plotting both estimated marginal-mean functions, \( \hat{E}_F(Y|Xa) \) and \( \hat{E}_M(Y|Xa) \) against \( Xa \). The scatter plots of \( Y \) and \( \hat{Y} \) against \( Xa \)
may also be superimposed. After estimating the two marginal mean functions, we perturb (also known as ‘jiggling’) the values of $Y$ by adding a normal random variable with small variance. This is to enhance the graphical presentation and interpretability. Note that smoothing of the scatter plots is done on the original (‘unjiggled’) data. Jiggling is essential if a covariate is categorical. Caution should be taken when interpreting the jiggled scatter plots: the jiggled data points form scatter ‘clusters’ on both one and zero. The darkness and width of these clusters does not necessarily reflect the true amount of data points underlying these points. We found that the graphs are misleading since our proportion zeros to ones is extremely high. Hence, the estimated curves $\tilde{E}_F(Y|Xa)$ and $\tilde{E}_M(Y|Xa)$ are much lower than they might seem from the scatter plot of the jiggled data.

To interpret the marginal-model plot, the two curves $\tilde{E}_F(Y|Xa)$ and $\tilde{E}_M(Y|Xa)$ should exhibit ‘closeness’. Should this be the case for all values of $a$, there is reason to believe that the model describes the data well. Marginal-model plots can be drawn for covariates that are continuous as well as discrete. Should $\tilde{E}_F(Y|Xa)$ and $\tilde{E}_M(Y|Xa)$ on the marginal-model plot have a slope coefficient near to zero for a specific covariate, it would indicate that this particular covariate does influence the occurrence of the response variable.

**Example continued**

Continuing the example in Section 3.2, we fit a binary logistic regression. The conditional expected value of the model is

$$E_M(Y|X) = [1 + \exp(-(Xa))]^{-1},$$

with fitted values

$$\hat{Y} = [1 + \exp(-(Xa))]^{-1},$$

for any value of $a$. In Figure 3-4 we plot the marginal-model plot when $a = \hat{\beta}$. So,

$$Xa = X\hat{\beta} = \logit^{-1}(Y)$$
Figure 3-4: Marginal-model plot for the linear component $\mathbf{X}\beta$. Superimposed is the scatterplot of $(Y_i, X_i\beta)$. The smoothing parameter was 0.6.
is the linear component of the logistic regression model. To increase interpretability we included two histograms. The top histogram is the histogram of the realized one values, whereas the bottom histogram is the histogram of the realized zero values. This makes the interpretation easier: we can now clearly see that the proportion of zeros to ones is high (whereas only interpreting the jigged clusters might have lead to another conclusion).

Here we see that the two estimates $\hat{E}_F(Y|X_\alpha)$ and $\hat{E}_M(Y|X_\alpha)$ are not exceptionally close, especially on the right tail. To some extent this confirms the goodness-of-fit result (the Hosmer-Lemeshow test) we obtained earlier.

Figure 3-5 is the marginal-model plot with $a=(0,0,1,0)$. We are only interested in the effect of the current delinquency level covariate $D$, on the occurrence of a missed payment. For discrete covariates we jiggle both $X_\alpha$ and $Y$. Here we see that the blue solid line $\hat{E}_M(Y|X_\alpha)$ is linear, while the red dotted line representing the estimated true model $\hat{E}_F(Y|X_\alpha)$ is distinctly curved. This strongly suggests that a second-order term, $D^2$, should be included in the model. The marginal-model plot for the covariate $TD$, not depicted here, also suggested that a second-order term for $TD$ should be included. In

![Figure 3-5: Marginal-model plot for $D$. Superimposed is the jigged scatterplot of $(Y_i,D_i)$. The smoothing parameter was 0.3.](image)

63
Figure 3-6 we choose \( a = (0, 1, 0, 0) \), i.e. we analyse the covariate Age. Here it would seem as if the non-parametric model has a constant level for values of Age up to about \(-1\). It would also seem as if the postulated model and the non-parametric model are very close to one another. As a result, a higher-order term might not be needed for this covariate.

![Marginal-model plot for Age](image)

Figure 3-6: Marginal-model plot for Age. Superimposed is the jiggled scatterplot of \((Y_i, Age_i)\). The smoothing parameter was 0.6.

We now have a better idea of how to detect a model’s deficiencies. This would not have been made as apparent with traditional goodness-of-fit methodologies. The marginal-model plots seem to show that second-order terms are missing in the model.

Typically, it is hard to judge the closeness of curves if no measure is available of the uncertainty associated with each curve. Hence, the second moments are used. Cook and Weisberg (1997) derive the marginal-variance function for generalized linear models. They estimate the standard deviation of the true non-parametric distribution \( F(Y|X) \), \( S\hat{D}_F(Y|Xa) \). They extend the marginal-model plot by superimposing the two one-standard-deviation bands \( \hat{E}_F(Y|Xa) \pm S\hat{D}_F(Y|Xa) \) in addition to the marginal mean plots over the scatter plot of \( Y \) versus \( Xa \). For the model distribution \( M(Y|\theta, X) \), they compute the marginal-variance function analytically, and also superimpose these plots...
on the marginal-model plot. The standard deviation bands serve as reference bands that enable the judgment of the closeness of $\hat{E}_F(Y|Xa)$ and $\hat{E}_M(Y|X)$. In binary logistic regression we know that the variance $P(Y = 1|X)[1 - P(Y = 1|X)]$ is non-constant and is a function of the predicted value of $Y$ given $X$. Cook and Weisberg (1997) do not compute or plot these bands for binary data at all, since the probability of default is a sufficient statistic for the binary distribution. However, we would like to obtain some measure or guideline regarding whether the two estimated loess curves are indeed close. When working with binomial data, over-dispersion is often encountered. Over-dispersion occurs when the variance of the data is higher than implied by the parametric distribution. This is in contrast to the idea that the first moment is a sufficient statistic. With this in mind, we revert to non-parametric methods to estimate a confidence band.

We now describe a novel approach to assess whether $\hat{E}_F(Y|Xa)$ is close to $\hat{E}_M(Y|Xa)$ for all types of regression, but specifically binary logistic regression. Judging the closeness of the marginal-mean functions visually gives no indication of the underlying uncertainty associated with the marginal mean functions. In a certain sense, we would like to include some dispersion measure to determine the closeness of the two non-parametric smoothed curves. Determining whether $\hat{E}_F(Y|Xa)$ and $\hat{E}_M(Y|Xa)$ are close is made easier by the use of reference bands. Cook and Weisberg (1997) plot a reference band equal to one marginal standard deviation around the estimated first moments of both the true non-parametric and postulated models for linear regression.

Since Cook and Weisberg did not construct confidence bands for binary data, we propose a confidence band that should serve as a good estimate of the dispersion level given $Xa$ for any $a \in \mathbb{R}^p$. We estimate a point-wise confidence band around the expected value of the true non-parametric model by bootstrapping. We do not concern ourselves too much with whether the correct method to use is a pointwise confidence interval or a simultaneous confidence band. In nearly all cases, the pointwise confidence interval is narrower than the simultaneous confidence band (Härdle et al. 2004 p. 65). The pointwise estimate is therefore more conservative than the simultaneous confidence bands. From now on when a confidence band is referred to, this means a pointwise confidence band. Should $\hat{E}_M(Y|Xa)$ fall within the bootstrapped pointwise confidence band, we will have good reason to believe that the model is adequate. In some sense we are performing a
hypothesis test where the null hypothesis is that $\hat{E}_F(Y|Xa)$ is correct. Hence, should $\hat{E}_M(Y|Xa)$ fall outside the confidence bands of $\hat{E}_F(Y|Xa)$, we know that the postulated model is incorrectly specified. By choosing various values of $a$ we may check whether the postulated model’s marginal-mean function lies within the confidence band of the true model. Should this not be the case, the ‘test’ will point out whether a transformation of covariates or other additional covariates could remedy the problem.

To simplify the notation, let

$$\hat{\theta}_F (Xa) = \hat{E}_F(Y|Xa)$$

and

$$\hat{\theta}_M (Xa) = \hat{E}_M(Y|Xa).$$

We perform a semi-parametric bootstrap by assuming that the non-parametric estimator $\hat{E}_F(Y|Xa)$ is the true probability of default given the linear combination of covariates $Xa$. The collection of covariates $X_1, ..., X_n$ is denoted by $\mathcal{X}$, and we sample with replacement $n$ values $X_1^*, ..., X_n^*$ from $\mathcal{X}$ and compute $p_i = \hat{\theta}_F (X_i^*a)$. Now binary random values are generated:

$$Y_i^* = \begin{cases} 1 & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{cases},$$

for $i = 1, ..., n$. By smoothing the scatter of $Y_i^*$ against $X_i^*a$, we have a bootstrap estimate of $\hat{\theta}_F (Xa)$. We can now estimate the bootstrapped sampling distribution of $\hat{\theta}_F (Xa)$, denoted by $\hat{\theta}_F^* (Xa)$, by using either an empirical density estimator or a kernel density estimator. We used the empirical density estimator to derive the $100(1 - \alpha)\%$ confidence bands around $\hat{\theta}_F (Xa)$.

Pseudo code for this semi-parametric bootstrap is:

1. Using all data calculate $\hat{\theta}_F (Xa)$ and $\hat{\theta}_M (Xa)$;
2. Take a sample $X_1^*, ..., X_n^*$ from $\mathcal{X}$ with replacement and calculate $p_i = \hat{\theta}_F (X_i^*a)$;
3. Generate binary random variables $Y_i^*$ with probability $p_i$;
4. Smooth the scatter plot of $Y_i^*$ against $X_i^*a$ to find the bootstrap estimate $\hat{\theta}_F^* (Xa)$;
5. Repeat steps two to five $B$ times; and
6. Estimate the sampling distribution of \( \hat{\theta}_F(Xa) \) by using either a kernel density or empirical density estimation method.

Normally in inferential statistics the significance level \( \alpha \) is chosen to be \( \alpha \in [0.05, 0.1] \). We opt for the traditional \( \alpha = 0.05 \).

**Example continued**

Continuing the example in Section 3.2, we modify our model to including the covariates \( D^2 \) and \( (DT)^2 \). This model will henceforth be referred to as the modified model, whereas the model with first-order terms only will be referred to as the initial model. The modified model should exhibit better closeness than the initial model. We computed a point-wise confidence band for our example data for the linear part of the GLM, \( X\hat{\beta} \). We depict the marginal-model plot with the reference band, omitting the scatter plot, for the initial model and the modified model in Figure 3-7 and Figure 3-8 respectively. Comparing these figures we see that the modified model is much closer to the non-parametric model than the initial model. In fact, \( \bar{E}_M(Y|X\hat{\beta}) \) falls inside the confidence interval for almost the whole range of \( X\hat{\beta} \). The marginal-model plots for the individual covariates also look promising for the modified model. In Figure 3-9, we draw the marginal-model plots for \( D, DT, Age \) and a random linear combination \( Xa^R \) with

\[
a^R = [0.7252, -0.47761, -0.3031, 0.94568, -0.29383, 0.073964]^T.
\]

We chose \( a^R \) by generating a random number between \(-1\) and \(1\) and standardizing the vector so that \( \sum (a_i^R)^2 = 1 \). Clearly, the marginal-model plot suggests that the fit has improved: all covariates now lie within the reference bands. The random direction remained mostly within the reference bands, except for high values of \( Xa^R \).

The use of bootstrap methods to derive reference bands is not new. Pan et al. (2001) also use the bootstrap method to obtain a reference band. Their goal is to assess the statistical uncertainties when comparing two marginal-model-mean functions. Their reference band is calculated by bootstrapping the estimators \( \hat{E}_F(Y|Xa) \) and \( \hat{E}_M(Y|Xa) \). The middle of their reference band is given by the average \( \left[ \hat{E}_F(Y|Xa) + \hat{E}_M(Y|Xa) \right] / 2 \) and half the width of the reference band is given by the standard error of the difference.
Figure 3-7: Marginal-model plot and bootstrapped confidence band for the initial model \( \logit(Y) = \beta_0 + \beta_1 D + \beta_2 DT + \beta_3 \text{Age} \).

They argue that if \( E_F(y|X_a) \) and \( E_M(y|X_a) \) are close or equal, their estimates \( \hat{E}_F(Y|X_a) \) and \( \hat{E}_M(Y|X_a) \) should lie within the reference band. This approach is technically incorrect: in their paper, a referee pointed out that they did not estimate the standard error function under the null hypothesis (that the model fits well). Our approach is different to the approach by Pan et al. (2001) in the way that we calculate the reference band. Note that in our approach, we are performing a bootstrap hypothesis correctly by bootstrapping from the null hypothesis that \( E_F(y|X_a) \) is the correct model.

Our approach is also different from Cook and Weisberg’s in that we estimate a 100\((1-\alpha)\)% confidence band around \( E_F(Y|X_a) \) and not only a one-standard-deviation reference band.

We now elaborate on the marginal-model plot methodology further. Assuming that \( E_F(Y|X_a) \) is the true model, we are faced with the problem of finding a parametric model that is close enough to the true model. The true mean squared error is given by \( \text{MSE} = E \left[ \left( \theta_F(X_a) - \hat{\theta}_F(X_a) \right)^2 \right] \). The closeness of the postulated model to the true
Figure 3-8: Marginal model plot and bootstrapped confidence band for the modified model

\[ \text{logit}(Y) = \beta_0 + \beta_1 D + \beta_2 DT + \beta_3 \text{Age} + \beta_4 (DT)^2 + \beta_5 D^2. \]

model can be measured numerically by the observed mean-squared error

\[ \overline{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\theta}_F(X_i, a) - \hat{\theta}_M(X_i, a) \right)^2 / n, \]

where the summation index \( i \) denotes the \( n \) obligors in the samples. We need some way to objectively measure whether this quantity is small or large. By computing the bootstrap estimates \( \hat{\theta}_F^*(Xa) \) and we approach this by calculating the mean-squared error for each of the bootstrap samples to obtain

\[ MSE^b = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\theta}_F(X_i, a) - \hat{\theta}_F^*(X_i, a) \right)^2 / n, \]

for \( b = 1, 2, \ldots, B \). Then \( \sum_b MSE^b / B \) is the bootstrap estimate of \( MSE \). The empirical distribution function, \( F_{\overline{MSE}Xa} \) of \( MSE^b \) for \( b = 1, 2, \ldots, B \) is used to estimate the sampling distribution of \( \overline{MSE} \). The reasoning is that if \( \overline{MSE} \) lies close to the center of \( F_{\overline{MSE}Xa} \) it can be argued that the postulated model could just as well have arisen from
Figure 3-9: The marginal-model plots for all covariates and a random direction $\mathbf{Xa}^R$. The smoothing parameter for $D$ was 0.3, $Age$ was 0.6, $DT$ was 0.4 and the random part was 0.8.
one of the non-parametric bootstrap samples. The $p$-value of the test is

$$p = 1 - F_{\widetilde{MSE}|X\cdot}(a),$$

computed at $a = \widetilde{MSE}$. This hypothesis test should be done for all linear combinations of $X\cdot$. However, since there are infinitely many options in choosing $a$, this task is impossible.

**Example continued**

Continuing the example in Section 3.2, we estimate $f_{\widetilde{MSE}|X\cdot}$ by using a kernel density estimator with smoothing parameter 0.1. This is shown in Figure 3-10 for both the initial and the modified models. We also plot $\widetilde{MSE}$ as a vertical red dotted line on each graph. Clearly, the null hypothesis is rejected in the initial model, in contrast to the modified model with its much larger $p$-value.
We drew a heuristic Hosmer-Lemeshow graph similar to Figure 3-2 of both the initial as well as the modified model in Figure 3-11. Here the modified model (in red) is closer to the perfect-fit line than the initial model (in blue). This shows that the addition of second-order terms has improved the prediction ability of the model. The modified model appears to be much closer to the perfect-model line than the initial model.

![Figure 3-11: The heuristic goodness of fit test for both the original model (in blue) as well as the modified model (in red).](image)

We also tested the goodness of fit of the modified model by again using the Hosmer-Lemeshow test with \( j = 10 \). The goodness of fit did not reject the null hypothesis in (3.13) with a test value of \( G^2_{HL} = 9.056 \) and a \( p \)-value of 0.338 (up from \( 3.323e-06 \)). Now the Hosmer-Lemeshow test does not reject the null hypothesis even though the sample size is large. Again, this does not tell us that the model is correct, only that the lack-of-fit is not great enough for us to reject the model (Kuss 2002).

To summarize, we see that marginal-model plots assess the adequacy of a model by checking the functional form and relevance of the current and potential covariates. We developed a simple approach to bootstrap reference bands around an estimated marginal-mean function. The reference bands enable us to judge the closeness of the marginal-mean
functions of the true non-parametric and postulated models. With this technique, the model’s adequacy and the need for covariate transformation or covariate addition may be explored. We now explore methods of determining a model’s discriminatory power.

### 3.4.4 Power of discrimination- ROC analysis

**Determining the optimal cut point and contingency tables**

Say we have binary random variables $Y_1, ..., Y_n$, which are indicators of a credit event. One may use $Y_i = 1$ if a default occurred (we call this a ‘positive’ event) and $Y_i = 0$ if it did not (a ‘negative’ event). Additionally, say we postulate a model that predicts default or positive events with probabilities $\hat{\pi}_1, ..., \hat{\pi}_n$. Note that we have $\sum_{i=1}^{n} Y_i$ positive responses and $n - \sum_{i=1}^{n} Y_i$ negative responses. One way to determine the model’s adequacy is through the use of a contingency table. In order to construct a contingency table, we need to determine an adequate cut point $(c \in [0, 1])$. The chosen $c$ should ‘best’ discriminate between positive and negative events. Choosing the best $c$ can be done in several ways. Here we discuss two ways: (1) accuracy and (2) the expected-cost function.

Let $\hat{Y}_i(c)$ be the predicted indicator variable of random variable $Y_i$ with cut point $c$. We then set $\hat{Y}_i(c) = 1 \{ \hat{\pi}_i \geq c \}$ for each $i$. We can now count the number of correctly estimated positive events $T^+(c) = \sum_{i=1}^{n} \hat{Y}_i(c) Y_i$ (this is widely referred to as the ‘true positive’ number). The number of correctly estimated negative events is given by $T^-(c) = \sum_{i=1}^{n} \left(1 - \hat{Y}_i(c)\right) (1 - Y_i)$ (also known as the ‘true negative’ number). Similarly, the ‘false positive’ or negative responses incorrectly or falsely estimated as positive are given by $F^+(c) = \sum_{i=1}^{n} \hat{Y}_i(c) (1 - Y_i)$. The number of positive responses falsely estimated as negative events is given by $F^-(c) = \sum_{i=1}^{n} \left(1 - \hat{Y}_i(c)\right) Y_i$ (‘false negative’).

We may now estimate the probability of correctly classifying a positive event as positive (the sensitivity or ‘true positive rate’) $P(\pi_i \geq c | Y_i = 1)$ by

$$\hat{P}(\hat{\pi}_i \geq c | Y_i = 1) = \frac{T^+(c)}{T^+(c) + F^-(c)} := Q^+(c),$$

say. $Q^+(c)$ is the total correctly specified positive responses at cut point $c$ divided by the total number of positive responses.
The so-called *specificity* is the estimate for \( P(\hat{\pi}_i < c | Y_i = 0) \) given by

\[
\hat{P}(\hat{\pi}_i < c | Y_i = 0) = \frac{T^-(c)}{T^-(c) + F^+(c)} := Q^-(c).
\]

Similarly to \( Q^+(c) \), \( Q^-(c) \) is the total proportion of correctly specified negative responses.

The optimal cut point \( \hat{c} \) can be estimated by maximizing the accuracy level defined as

\[
P(\text{classifying correctly}|c) = \frac{T^+(c) + T^-(c)}{n}.
\]

When we want to determine the optimal cut point by maximizing accuracy, we find that the accuracy is dominated by the group that has the most observations (Hosmer et al. 2013 p. 171). Hence, accuracy is sensitive to the relative sizes of the groups and favours classification into the larger group. Note that choosing the value of \( c \) to be either 0 or 1 will lead to all observations being classified as either positive or negative, as \( \lim_{c \to 1} Q^+(c) = 1, \lim_{c \to 0} Q^+(c) = 0, \lim_{c \to 1} Q^-(c) = 0 \) and \( \lim_{c \to 0} Q^-(c) = 1 \).

**Example continued**

Continuing the example in Section 3.2, in our example data the maximum accuracy of 0.8725 is achieved at \( c \in [0.531, 1] \). So, in choosing \( c = 1 \) all events are classified as negative. This coincides with the fact that \( \sum_{i=1}^{n} Y_i/n = 0.8725 \). To counter this bias we could maximize the sum of \( Q^+(c) \) and \( Q^-(c) \). Figure 3-12 overlays the sensitivity \( Q^+(c) \) and the specificity \( Q^-(c) \) on one graph for \( c \in (0, 0.6) \). Note that \( c > 0.6 \) was uninformative and has been omitted. Here the sum of \( Q^+(c) \) and \( Q^-(c) \) is maximized at the point where the two curves cross. For each value of \( c \) we can construct a classification or contingency table. For instance, in the example above, with \( c \) chosen to be more or less the maximum of the sum of \( Q^+(c) \) and \( Q^-(c) \), it was found that \( \hat{c} = 0.11 \). We sampled 2,000 observations that were used as a training sample and fitted the modified model to predict \( \hat{\pi}_i \). Using \( \hat{c} = 0.11 \) and fitting the model on the other 2,000 observations in the validation sample, we calculated the contingency table in Table 3.5. Here the specificity was \( 1,355/1,803 = 0.7515 \) and the sensitivity \( 136/197 = 0.69 \). The overall value of the correctly classified observations (also referred to as the accuracy) is \( (152 + 1,308)/2,000 = 0.73 \).
Figure 3-12: The accuracy can be determined by optimizing the sensitivity and specificity. Here an optimal point could be at about 0.11, where the two curves cross.

Table 3.5: The contingency table for the validation dataset.

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th></th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Y = 1</td>
<td>152</td>
<td>472</td>
<td>624</td>
<td></td>
</tr>
<tr>
<td>Y = 0</td>
<td>68</td>
<td>1,308</td>
<td>1,376</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>220</td>
<td>1,780</td>
<td>2,000</td>
<td></td>
</tr>
</tbody>
</table>

The expected-cost function is also useful in determining the optimal $c$. Some penalty or weights are used to discriminate between positive and negative responses. Heuristically, an incorrect classification might cost more than a correct classification. If we were to include a cost of misclassification, then the expected cost is

$$E[Cost|c] = \frac{T^+(c)C(\hat{Y} = 1|Y = 1) + T^-(c)C(\hat{Y} = 0|Y = 0)}{n} + \frac{F^+(c)C(\hat{Y} = 1|Y = 0) + F^-(c)C(\hat{Y} = 0|Y = 1)}{n},$$

(3.18)

where $C(X = x|Y = y)$ is the cost involved for classifying an event that is known to
be $Y = y$ as $X = x$ with $x, y \in \{0, 1\}$ at a chosen $c$. No fixed rule exists for choosing $C(X = x|Y = y)$. In a credit risk scoring application, Bellotti and Crook (2009) assumes that $C(\hat{Y} = 1|Y = 1) = C(\hat{Y} = 0|Y = 0) = 0$, $C(\hat{Y} = 1|Y = 0) = 1$ and that $C(\hat{Y} = 0|Y = 1) = 20$. Hence, no cost is incurred by correctly predicting either a positive or a negative event. There is a cost of one if a negative event was predicted but not observed. A cost of 20 is assumed if a positive event was predicted but a negative event was observed. Therefore, the cost ratio is $20:1$, reflecting the severity of making a wrong decision (in the credit-risk-management context, classifying a default as a non-default will certainly incur more costs than vice versa). Since the cost ratio is heuristic and not set in stone, Bellotti and Crook (2009) perform a robustness test by considering cost ratios of $15:1$ and $25:1$.

Example continued

Continuing the example in Section 3.2, in Figure 3-13, we determine the optimal cut point using the expected cost with various cost ratios on the training data. Clearly the costs are optimal at about $\hat{c} = 0.06$, irrespective of the cost ratio chosen. This answer differs from the optimal cut point that we obtained by maximizing the accuracy ($\hat{c} = 0.11$). The contingency table for the validation data set at $\hat{c} = 0.06$ is given in Table 3.6. The accuracy 0.665 and the expected cost is minimized at 0.7625. The accuracy is now not as high as the accuracy obtained by maximizing the accuracy. This compares favourably with the results from Table 3.5, where the accuracy was 0.73 (about 10% higher), but the expected cost was higher at 0.916 (a cost increase of 20%).

Table 3.6: The contingency table minimizing the expected cost.

<table>
<thead>
<tr>
<th>Expected</th>
<th>Observed $Y = 1$</th>
<th>Observed $Y = 0$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{c}$ = 0.06</td>
<td>175</td>
<td>625</td>
<td>800</td>
</tr>
<tr>
<td>$\hat{c}$ = 0.11</td>
<td>45</td>
<td>1,155</td>
<td>1,200</td>
</tr>
<tr>
<td>Total</td>
<td>220</td>
<td>1,780</td>
<td>2,000</td>
</tr>
</tbody>
</table>
Figure 3-13: Optimal cutpoint for the different cost ratios should be at about 0.06.

**ROC analysis**

Choosing an optimal cut point does not convey any information about the classification ability or power of discrimination of the postulated model. Classification measurements such as sensitivity and specificity are limited to analysis involving a single cut point. The receiver operating characteristic (ROC) curve was developed specifically to address this limitation. ROC-curve analysis was developed to analyse classification accuracy when differentiating signal from noise in radar detection. We use ROC analysis to analyse the classification accuracy or discriminatory power of a postulated default prediction model. The ROC curve and corresponding area under the ROC curve (AUC) are often used in credit-risk scoring and management.

The ROC curve attempts to graphically depict the trade-off pattern between the correctly classified and incorrectly classified obligors. The ROC curve is a scatter plot of $Q^+(c)$ against the $1 - Q^-(c)$ for all values of $c$ (not only one specific $c$). The ROC curve has a domain and range of 0 to 1. A binary classifier that performs perfectly should be a line that starts at $(0,0)$, goes directly to $(0,1)$ and then continues to $(1,1)$. We refer to such a model as the ‘perfect model’. A classifier that is no better than a proverbial
coin flip will be the diagonal line that connects the points \((0, 0)\) and \((1, 1)\). A good model should have a curve above the coin-flip line and close to the perfect line.

The ROC curve can be summarized in a single statistic that describes the model’s ability to discriminate. Note that, as is usual with most statistics, all attempts to summarize the ROC curve in a single statistic lose some information. An attempt to summarize the classification accuracy of the model is given by the AUC: the area between the horizontal line on the \(x\)-axis and the ROC curve.

According to Hosmer et al. (2013 p. 174):

"This measure [the AUC] has now become the standard for evaluating a fitted model’s ability to assign, in general, higher probabilities of the outcome to the subgroup who develop the outcome \((y = 1)\) than does to the subgroup who do not develop the outcome \((y = 0)\)".

A good model has an AUC value close to one. Such a model will have good discrimination, whereas a model with an AUC equal to 0.5 may just as well have been a coin toss. Although the AUC seems like a crude summary measurement, it has a probabilistic interpretation. The AUC can be seen as the probability of correctly classifying a randomly chosen positive event more highly than a randomly chosen negative event. The AUC can be estimated by using trapezoidal approximation or by noting that it is a function of the non-parametric test for stochastic dominance (the Mann–Whitney \(U\) statistic). Specifically, if \(U\) is the Mann–Whitney \(U\) statistic calculated in the normal way, then, according to Brown and Davis (2005),

\[
AUC = \frac{U}{\sum_{i=1}^{n} 1\{Y_i = 1\} \sum_{i=1}^{n} 1\{Y_i = 0\}}.
\]

The ROC curve is insensitive to the costs involved when classifications are wrong, possibly giving rise to misleading conclusions. Bellotti and Crook (2009) prefer not to use the ROC curve for this exact reason. They rather use the expected cost.
Hosmer et al. (2013 p. 177) state that the following rule of thumb should be used when using the AUC as a discrimination measure:

\[
\begin{align*}
AUC = 0.5 & \text{ This suggests no discrimination, so we might as well flip a coin;} \\
0.5 < AUC < 0.7 & \text{ We consider this poor discrimination, not much better than a coin toss;} \\
0.7 \leq AUC < 0.8 & \text{ We consider this acceptable discrimination;} \\
0.8 \leq AUC < 0.9 & \text{ We consider this excellent discrimination; and} \\
AUC \geq 0.9 & \text{ We consider this outstanding discrimination}. \\
\end{align*}
\]

**Example continued**

Continuing the example in Section 3.2, in Figure 3-14 we show the ROC curve for the modified model fitted on the example data. The red dotted line is the coin-flip line. The AUC was 0.7604, denoting ‘acceptable discrimination’.

There are certain problems associated with the use of the AUC, and as a result it
should be used with caution. Lingo and Winkler (2008) state that the AUC can only be used as a discriminatory method of power comparison when the portfolios are exactly the same. Additionally, the AUC of a portfolio cannot be compared to the AUC of the same portfolio at a different time. This is because of the fact that the AUC is random and is dependent on the underlying observations/portfolio. When using the AUC as a comparison method for various models, the AUC should be calculated on the same data. This will give an indication of which model has the best discrimination.

3.5 Appendix

3.5.1 Generalized cross-validation technique for estimating the lasso tuning parameter in logistic regression

Here we derive the generalized cross-validation (GCV) technique for estimating the tuning parameter when applying the lasso in logistic regression. As far as we know this has not been done before. Tibshirani (1996) derives the generalized cross-validation (GCV) formulae for a simple linear regression model. In a later article, Tibshirani (1997) uses GCV to estimate the tuning parameter of the lasso-type Cox (1972) proportional-hazard model. We discuss the simple linear-regression GCV and thereafter derive the GCV for logistic regression.

GCV approximates \( n \)-fold cross-validation (leave-one-out cross-validation), but it is computationally much less expensive. It is especially useful if we only know the average influence of all cases, i.e. we do not know the individual influences of all the observations. Say we have some linear model

\[
\hat{Y} = SY,
\]

with \( S \), some projection matrix. \( S \) might be, among other things, a smoothing matrix or a hat matrix. To perform an \( n \)-fold cross-validation we make use of the Sherman-Morrison formula for the deleted residual:

\[
y_i - \hat{y}_{i,(i)} = \frac{y_i - \hat{y}_i}{1 - S_{ii}},
\]

where \( \hat{y}_{i,(i)} \) is the fitted value of \( y_i \) if observation \( i \) was omitted during the estimation. In
(3.20) \( S_{ii} \) is the leverage or influence of the \( i^{th} \) case and is simply the \( i^{th} \) diagonal element of \( S \) for \( i = 1, 2, ..., n \). This remarkable, but well-known, result allows us to compute the leave-one-out statistics without refitting the model \( n \) times.

If any model could be written in the form (3.19), the leave-one-out cross-validation is

\[
CV(t) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{h}^{(-i)}(X_i, t))^2,
\]

where \( \hat{h}^{(-i)}(X_i, t) \) is the leave-one-out linear model predicting \( Y_i \) dependent on the turning parameter \( t \). Loader (1999 p. 35) uses Henderson’s theorem to prove that if \( S_{ii} < 1 \)

\[
\hat{h}^{(-i)}(X_i, t) = \frac{\hat{h}(X_i, t) - S_{ii}Y_i}{1 - S_{ii}},
\]

for \( i = 1, 2, ..., n \). Substituting (3.22) into (3.21) we get

\[
CV(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - \hat{h}(X_i, t))^2}{(1 - S_{ii})^2}.
\]

The leave-one-out cross-validation can easily be calculated if the leverage of each variable is known. Should individual leverages be unknown and we only know the sum of the diagonal elements of \( S \), denoted by \( \text{trace}(S) \), we may approximate the individual leverages using the ‘average influence’. In other words, we can set each \( S_{ii} \) equal to average leverage \( \frac{\text{trace}(S)}{n} \). Substituting the average influence into (3.23), we get the GCV, namely

\[
GCV(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - \hat{h}(X_i, t))^2}{(1 - \text{trace}(S)/n)^2} = \frac{SSE(t)}{n(1 - \text{trace}(S)/n)^2},
\]

where the residual sum of squares \( SSE(t) := \sum_{i=1}^{n} (Y_i - \hat{h}(X_i, t))^2 \). We can now estimate \( t \) by \( \hat{t} \) as

\[
\hat{t} = \arg\min_{t} \{GCV(t)\}.
\]

The \( \text{trace}(S) \) can be seen as the effective number of parameters or degrees of freedom of
the model. For simple linear regression it is known that \(\text{trace}(S) = p\), but this equality might not always hold, especially with constrained regression models.

We now turn our attention to the logistic regression case. When applying the lasso to logistic regression, we estimate the parameters by maximizing \(\text{lik}(\beta)\) subject to the constraint that \(\sum_{j=1}^{p} |\beta_j| \leq t\). Maximizing the \(\text{lik}(\beta)\) is approximately equivalent to minimizing the IRLS in (3.12), subject to the same \(L_1\) constraint. We may rewrite the constraint \(\sum_{j=1}^{p} |\beta_j| \leq t\) as \(\sum_{j=1}^{p} \beta_j^2 / |\beta_j| \leq t\); this is done to transform the lasso regression into a ridge-regression scheme. Hence, we estimate \(\beta\) by

\[
\hat{\beta} = \arg\min_{\beta} \left[(z - X\beta)^\top W (z - X\beta)\right]
\]

subject to

\[
\sum_{j=1}^{p} \beta_j^2 / |\beta_j| \leq t,
\]

with \(z = X\beta + W^{-1} (y - \pi)\) and \(W, \pi\) and \(\beta\) evaluated at \(\hat{\beta}^m\), where \(m\) is the \(m^{th}\) iteration in the IRLS methodology. Making use of Lagrangian multipliers we get

\[
\hat{\beta}_{\text{lasso}}^t = \arg\min_{\beta} \left[(z - X\beta)^\top W (z - X\beta) + \lambda(\beta U \beta - t)\right],
\]

where \(U\) is the \((p \times p)\) matrix with \(\text{diag} \left((1/ |\beta_j^m|) I \{|\beta_j^m| > 0\}\right)\). The solution to this optimization problem is obtained by differentiating with respect to both \(\beta\) and \(\lambda\), resulting in respectively,

\[
\frac{\partial}{\partial \beta} \left[(z - X\beta)^\top W (z - X\beta) + \lambda(\beta U \beta - t)\right] = -2X^\top Wz + 2X^\top WX\beta + 2\lambda U^{-1}\beta \quad (3.25)
\]

and

\[
\frac{\partial}{\partial \lambda} \left[(z - X\beta)^\top W (z - X\beta) + \lambda(\beta U \beta - t)\right] = \beta U \beta - t \\
= \beta^\top \text{sign}(\beta) - t,
\]

82
where \( \text{sign}(x) \) is + if \( x \geq 0 \) and − if \( x < 0 \). Setting the partial derivative with respect to \( \lambda \) equal to zero, we obtain

\[
t = \beta^\top \text{sign}(\beta).
\] (3.27)

Setting (3.25) equal to zero we get

\[
\hat{\beta}_t^{\text{lasso}} = (X^\top WX + 2\lambda U^{-1})^{-1} X^\top Wz,
\] (3.28)

with \( \lambda \) still unknown. The estimate of \( \lambda \), denoted by \( \tilde{\lambda} \), can be derived by assuming that \( \frac{\partial}{\partial x} (|x|) = \frac{\partial}{\partial x} [x \text{sign}(x)] = \text{sign}(x) \), hence assuming the derivative of \( |x| \) is zero at the vertex point \( x = 0 \). Now, substituting (3.27) into (3.26) and pre-multiplying with \( \beta^\top \) we get

\[
-2\beta^\top X^\top Wz + 2\beta^\top X^\top WX\beta - \lambda \beta^\top \text{sign}(\beta)
= -2\beta^\top X^\top Wz + 2\beta^\top X^\top WX\beta - \lambda t.
\]

Setting this equation equal to zero, we get the relationship between the estimates \( \tilde{\lambda}, \hat{\beta}_t \) and \( t \)

\[
\tilde{\lambda} = \frac{2}{t} \left( \hat{\beta}_t^\top X^\top WX\beta - \hat{\beta}_t^\top X^\top Wz \right).
\] (3.29)

Substituting \( \tilde{\lambda} \) into (3.28) we get

\[
\hat{\beta}_t = \left( X^\top WX + \tilde{\lambda} U \right)^{-1} X^\top Wz.
\] (3.30a)

Inspecting (3.30a), we have rewritten the logistic problem in a linear form (3.19) with

\[
S = X \left( X^\top WX + \tilde{\lambda} U \right)^{-1} X^\top W
\]

and

\[
Y = z.
\]

We now calculate the complexity of the fit (the effective number of parameters) for this equation. The effective number of parameters can intuitively be interpreted as the number
of parameters used in the regression. Tibshirani suggests using the usual equation:

$$p(t) = \text{trace}(X (X^\top WX + \hat{\lambda}U)^{-1} X^\top W).$$ \hspace{1cm} (3.31)

We found (3.31) behaves erratically, producing a saw tooth wave with jumps at the inclusion of every new covariate. An adjustment should be made in order to counter the possible degenerate nature of $U^{-1}$ caused by its zero-diagonal elements. Note that, ideally, if $t = 0$ all coefficients $\hat{\beta}$ should be zero. Additionally, from (3.29) it follows that $\lim_{t \to \infty}(\lambda) = 0$ and the resulting complexity would be $p(t) = p$, since the trace of the hat matrix is equal to the number of parameters. So, we want a function such that $p(t)$ has properties $\lim_{t \to 0^+} p(t) = 0$ and $\lim_{t \to \infty} p(t) = p$. To accomplish this, we use an idea by Fu (1998 p. 62) to adjust the effective number of parameters. Set

$$p(t) = \text{trace}(X (X^\top WX + \hat{\lambda}U)^{-1} X^\top W) - \sum_{i=1}^{p} 1 \left\{ \hat{\beta}_j^{\text{lasso}} = 0 \right\}. \hspace{1cm} (3.32)$$

To estimate the value of $t$, we prefer to use the deviance function instead of the residual sum of squares. So, replacing the $SSE(t)$ with $-2l(\hat{\beta}_t^{\text{lasso}})$ in (3.24) and using (3.32) as our effective number of parameters, we get the GCV-style statistic for logistic regression

$$GCV(t) = \frac{-2l(\hat{\beta}_t^{\text{lasso}})}{N (1 - p(t)/N)^2}.$$  

Here model complexity, $p(t)$, penalizes the model fit: we want a model that produces a small ‘lack of fit’ and a small number of covariates. We estimate $t$ by

$$\hat{t} = \arg\min_t [GCV(t)].$$
Chapter 4

Modelling dynamic credit events: counting processes

In this chapter we extend the static classification approach of managing credit risk discussed in Chapter 3 to dynamic situations. A natural extension to the classification of cross-sectional data is possible when we have longitudinal data. Longitudinal data (data that is repeatedly collected and random over time) enables us to model credit risk dynamically. This chapter will focus on the use of counting processes to model credit risk. Counting processes are very general processes that are used to model the occurrence of events. Hence, they are very useful when modelling credit risk. Giesecke (2004) states that: "Underlying all credit models is the increasing default (point) process $N(t)$".

We identified counting processes as the preferred credit-risk modelling technique. We will discuss counting processes and their corresponding compensators in Section 4.2. beforehand, we briefly discuss the use of survival analysis. This chapter will show how survival analysis and other techniques developed in this field can be regarded as a specialization of counting process theory (Aalen 1975).

Counting processes are a general framework for most reduced-form model approaches. In counting processes an ‘event’ would usually be a death or failure. Throughout the study we take ‘events’ to be credit events.
4.1 Survival analysis

Survival analysis is concerned with, among other things, predicting the time until an individual incurs a failure. Historically, survival analysis was only concerned with the analysis of the actual survival times or mortality rates of a case study. Today, it is used in any analysis where the timing of any event is of concern. In survival analysis it is important to have unambiguous definitions for the event that occurs and the time origin. The time origin can be any set time point, such as the onset of a disease, a birth date, etc. The event or ‘failure time’ can be a death or the failure of a component. Using survival analysis, we can define the time origin as the date that an account was opened. In our case, this definition is restrictive, since all obligors’ repayment behaviour is influenced by systematic covariates and is not visible at the outset but along the lifetime of the loan. Hence, calendar time would serve us better than event time. In credit-risk management, the event may be any credit event: a failure, recovery, default or write-off. So, instead of referring to the occurrence of a death, we henceforth refer to the occurrence of some credit event.

Survival analysis is characterized by the structure of the observed data: it contains complete and incomplete observations. Incomplete data occurs because of censoring and truncation. The most common type of censoring is known as ‘right censoring’. Right censoring in survival analysis occurs when an individual is removed from the study prematurely or when the credit event has not occurred before the study period has concluded. Hence, we only know that the event will still occur after a certain time. In our case, obligors are censored if, for some reason, they choose to close their accounts early or no credit event occurred during the observation window.

In survival analysis, the assumption of independent censoring is ubiquitous. We discuss this in depth later, but essentially the cause of an obligor’s censorship should not depend on or influence the probability of a credit event. Survival analysis data may also be truncated. Delayed-entry data or staggered-entry data is left as truncated data where individuals are added to the study after the onset of the study. For delayed-entry data, the at-risk set will not strictly decrease, since the number of obligors may increase over time as new obligors enter the study.

Let the non-negative random time of a credit event occurring from a homogeneous
population be denoted by $T$. We assume that $T$ has probability-density function $f(t)$ and corresponding distribution function $F(t)$. As we will see, the central issue in survival analysis is the estimation of the distribution function $F(t)$. Often the probability of an event surviving for more than a certain time is of interest. The survival function (also known as the survivor function) is the probability that a credit event is greater than $t$, namely

$$S(t) = P(T > t) = 1 - F(t),$$

with $t > 0$. Theoretically, no events have failed at or before $t = 0$, so $S(0) = 1$ and all events should fail at the end, so $\lim_{t \to \infty} S(t) = 0$. However, owing to censoring, it is often true that $\lim_{t \to \infty} S(t) = k$ with $k \in [0, 1]$. Hence, a proportion $k$ of the population was censored and did not experience an event. The survival function can easily be estimated by non-parametric methods like the well-known Kaplan Meier and life-table estimators (Kalbfleisch & Prentice 2002 p. 13). Estimating the survival function of incomplete data is greatly simplified by the use of hazard rates and cumulative hazard rates.

The hazard rate is defined as

$$\alpha(t) := \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt},$$

$\alpha(t) \geq 0$. The hazard rate may intuitively, though not entirely correctly, be interpreted as the instantaneous probability or rate of a credit event occurring at time $t$ provided that the obligor has survived up to time $t$. However, $\alpha(t)$ may be larger than one, so a more correct interpretation of $\alpha(t)$ is that it is the expected number of credit events that will occur over a time period of $dt$.

The hazard rate and the survival function are related. If $f(t)$ is known, we may express $\alpha(t)$ in terms of $f(t)$ and $S(t)$ (Kalbfleisch & Prentice 2002 p. 7)

$$\alpha(t) = \frac{f(t)}{S(t)}.$$  

(4.2)

The hazard rate completely specifies the distribution and density functions through

$$S(t) = \exp \left( - \int_0^t \alpha(s) ds \right).$$  

(4.3)
Hazard rates are frequently used, but are not easy to estimate. The *cumulative hazard rate*

\[ A(t) = \int_0^t \alpha(s) \, ds, \]

is often much simpler than estimating \( S(t) \). Hence, once the cumulative hazard rate is estimated, we can derive \( \alpha(t) \), \( S(t) \) and \( f(t) \). Estimation of \( A(t) \) can be done non-parametrically with the well-known Nelson-Aalen estimator. This estimator was originally motivated by using Markov chains, but justifying the derivation is more correctly done using counting processes (Aalen et al. 2009 p. 5).

Several parametric failure-event-time models exist, simplifying matters by admitting a closed-form hazard-rate and survival function (Kalbfleisch & Prentice 2002 p. 31). However, the main concern of studies is mostly to establish a relationship between event times and covariates. Hence, a parametric approach would not suffice. Therefore, we highlight survival analysis techniques that employ covariates through *regression* models.

The hazard rate is essentially any positive function, so models that express the hazard rate as a function of time and certain covariates are often used. It is easier to postulate a regression-type model for the \( \alpha(t) \) than for \( S(t) \). So, all regression models are postulated forms of the hazard-rate function. The most popular regression model is Cox’s proportional hazard regression model (Cox 1972): let \( \alpha(t, X) \) be the hazard rate dependent on time \( t \) and covariate design matrix (without constant term) \( X \), then Cox’s proportional hazard model is

\[ \begin{align*}
\alpha(t, X) &= \alpha_0(t) \exp(X\beta),
\end{align*} \]

where \( \alpha_0(t) \) is some function of \( t \) called the *baseline hazard* rate and \( \beta \) is the coefficient vector. This model is called ‘proportional’ because the ratio of the hazard rate of two obligors will be constant over time. Specifying a closed-functional form for \( \alpha_0(t) \) results in a parametric survival-regression-analysis model (Allison 1982). Here the choice of \( \alpha_0(t) \) is difficult to justify. So, Cox’s model is a semi-parametric model that partially specifies the hazard rate. Cox showed that the coefficient \( \beta \) may be estimated by his so-called partial likelihood without imposing restrictions or assuming some functional form for \( \alpha_0(t) \) (Kalbfleisch & Prentice 2002 p. 101). Cox’s partial likelihood assumes that the baseline hazard rate \( \alpha_0(t) \) can be treated as unobservable and considered a nuisance parameter when estimating \( \beta \). Cox’s model also allows covariates to be time-dependent,
i.e. $X(t)$, a feature that we discuss later. The asymptotic properties of both the Kaplan-Meier estimate and Cox’s proportional hazard model were only formally developed after martingale concepts, which were a by-product of counting processes, were introduced to survival analysis (Aalen et al. 2009).

To summarize, survival analysis techniques were developed to model the timing of events using incomplete data. Estimating the survival function $S(t)$ can be done through postulating and estimating some hazard-rate function. A hazard-rate function can be estimated by models that are non-parametric, semi-parametric or parametric, with or without covariates. In the next section we discuss counting processes that are a more general case than survival analysis.

### 4.2 Counting processes: preliminaries

Counting processes, also known as point processes, form a general framework that enfolds classical survival analysis. Aalen (1975) used martingale theory as a framework to develop a mathematical theory for counting processes. Since then, the theory of counting processes has been applied to survival analysis with the use of martingales (Aalen et al. 2009). Their status as a more general framework allows counting processes to be readily applied to specific modelling problems. Counting processes are especially applicable when normal survival analysis techniques and assumptions are too restrictive. We begin this section by formulating the basic theory of counting processes. Thereafter, we show that survival analysis is a special case in the field of counting processes. We follow the presentations by Kalbfleisch and Prentice (2002), Aalen et al. (2008) and Andersen et al. (1993).

Assume we have a time-dependent univariate counting process $N(t)$ such that $N(t)$ is an integer: $N(t) \geq 0$, $N(0) = 0$ and $N(t) - N(s) \geq 0$ for $t \geq s$. $N(t)$ is assumed to be a cadlag process (continuous from the right with left limits). $N(t)$ can only make jumps of size one and is constant between jumps. The counting process is non-decreasing so it is a sub-martingale that satisfies the property

$$E(N(t)|\mathcal{F}_s) \geq N(s),$$
for all $s < t$, where $\mathcal{F}_s$ is the filtration or ‘history’ until time $s \geq 0$. The filtration $\mathcal{F}_t$ can be interpreted as all the past information available up to and including time $t$. $\mathcal{F}_t$ can represent the simple natural filtration $\mathcal{F}_t = \sigma \{ N(s); s \leq t \}$ or it may contain internal and/or external covariates as well.

The counting process $N(t)$ is said to be driven by an intensity process

$$\lambda(t) = \lim_{dt \to 0} P(dN(t) = 1 | \mathcal{F}_{t-})/dt, \quad (4.5)$$

where $dN(t) = N([t + dt]^-) - N(t^-)$ is the number of jumps in the time interval $[t, t + dt)$ and $t^-$ is the time just before $t$. The intensity process $\lambda(t)$ is random, since it is driven by the random filtration $\mathcal{F}_{t-}$. Rewriting (4.5) as an expectation, we get the equation

$$\lambda(t)dt = E(dN(t) | \mathcal{F}_{t-}), \quad (4.6)$$

so the intensity may be interpreted as the expected number of events occurring within the time period $[t, t + dt)$. The intensity process is said to be a predictable process, implying that $\lambda(t)$ is adapted to $\mathcal{F}_t$ and has left-continuous sample paths. Heuristically, the value of a predictable process at time $t$ is known at time $t$ (and is hence non-random). The cumulative intensity process (also known as the compensator) is

$$\Lambda(t) = \int_0^t \lambda(s) \, ds. \quad (4.7)$$

Rearranging (4.6) we get

$$E(dN(t) - \lambda(t)dt | \mathcal{F}_{t-}) = 0, \quad (4.8)$$

where $\lambda(t)$ can be taken into the expectation, since it is predictable. Using (4.8) and (4.7), the process

$$M(t) = N(t) - \Lambda(t), \quad (4.9)$$

has the property that

$$E(dM(t) | \mathcal{F}_{t-}) = 0, \quad (4.10)$$

which is the definition of a martingale. Intuitively, martingales are a more general concept for the ever-present independent errors assumption in statistics. Martingales have
the advantage of being able to be manipulated or transformed without destroying the martingale property. The same does not hold for the manipulation of the independence property (Aalen et al. 2008 p. 45). Hence, the counting-processes framework (using martingales) is more general than other frameworks. From (4.9) we see that a counting process \( N(t) \) minus its compensator is again a martingale. This useful decomposition (4.9) is known as the Doob-Meier decomposition. The Doob-Meier decomposition can be rewritten (4.9) as

\[
dN(t) = \lambda(t)dt + dM(t),
\]

which corresponds to the usual ‘Observation=Signal+Noise’-type model. Hence, the observed difference in the counting process, \( dN(t) \), can be modelled as the sum of a signal, \( \lambda(t)dt \), plus a noise term, \( dM(t) \). We may interpret the martingale \( M(t) = \int_0^t dM(s) \) as similar to the sum of random errors or cumulative noise in simple regression. Being a martingale, \( M(t) \) enables us to apply a wide variety of theories, such as the central-limit theorem for martingales (Helland 1982). The central-limit theorem for martingales then enables us to derive asymptotic properties of intensity or hazard-rate estimators. These techniques enabled Andersen and Gill (1982) to derive the asymptotic properties of Cox’s proportional hazard rate model rigorously.

### 4.3 Survival analysis via counting processes

We will from now on define all equations for a specific obligor \( i \), with \( i \in \{1, ..., n\} \). In survival analysis, the emphasis is on the random variable \( T_i \), the time of death. In contrast, counting processes emphasize that the whole counting process \( N_i(t) \) is random (timing as well as jumps). Survival analysis is easily shown to be a special case in counting processes. We will prove this by firstly focusing on the uncensored case and then the more general censored case. Survival analysis assumes that only one event takes place. Normally the death at time \( T \) is the only event of interest, and repeating events are not accounted for. Hence, we can rewrite a survival analysis problem by assuming that a counting process is now defined as \( N_i^* (t) = 1 \{T_i \geq t\} \). If we observe complete
data (i.e. no censoring takes place) we get

\[ \lambda_i^*(t) dt = P(dN_i^* (t) = 1| \mathcal{F}_t^-) \]

\[ = P(t \leq T_i < t + dt| \mathcal{F}_t^-) \]

\[ = \begin{cases} \alpha_i(t) dt & \text{if } T \geq t \\ 0 & \text{if } T < t \end{cases} \]

(4.12)

\[ = \mathbf{1}\{T_i \geq t\} \alpha_i(t) dt. \]

The resulting counting process for complete data survival analysis has intensity \( \lambda_i^*(t) = \mathbf{1}\{T_i \geq t\} \alpha_i(t) \). We can write the adapted ‘at-risk’ process as \( Y_i^*(t) = \mathbf{1}\{T_i \geq t\} \) and the hazard rate \( \alpha_i(t) \) is defined in (4.1). This simplifies the intensity of the counting process for complete data to

\[ \lambda_i^*(t) = Y_i^*(t) \alpha_i(t). \]

We now prove that censored survival analysis is also a special case in counting processes. Counting processes are mostly used when modelling censored or incomplete data. Left-censoring in survival analysis occurs when an obligor is removed from the study prematurely or when the event has not occurred within the observation window. We assume independent random censoring; in other words, censored times are independent of event times. This assumption ensures that no relationship exists between credit events and censor events. So, the probability that an event occurs is not effected at time \( t \) if an obligor is censored just before time \( t \). Assume each obligor \( i = 1, 2, ..., n \) has credit event time \( T_i \) and censoring time \( C_i \). Typically we will observe either the event time \( T_i \) or the censoring time \( C_i \) whichever occurs first. So we observe \( \tilde{T}_i = \min(T_i, C_i) \) as well as the event indicator

\[ D_i = \begin{cases} 1 & \text{if } T_i < C_i \\ 0 & \text{otherwise}, \end{cases} \]

for \( i = 1, 2, ..., n \). The probability that obligor \( i \) experiences an event in \([t, t + dt)\), given that the obligor is still at-risk at time \( t \) (i.e. the obligor is neither censored nor has he or
she experienced an event) is

\[
P(t \leq \tilde{T}_i < t + dt, D_i = 1|\tilde{T}_i \geq t, \mathcal{F}_t) = P(t \leq \tilde{T}_i < t + dt, T_i < C_i|T_i \geq t, C_i \geq t, \mathcal{F}_t)
\]

\[
= P(t \leq T_i < t + dt|T_i \geq t, C_i \geq t, \mathcal{F}_t)
\]

\[
= P(t \leq T_i < t + dt|T_i \geq t, \mathcal{F}_t).
\]

The last equality follows if and only if censoring times are independent of default times. Formally the independent censoring or uninformative censoring assumption is

\[
P(t \leq \tilde{T}_i < t + dt, D_i = 1|\tilde{T}_i \geq t, \mathcal{F}_t) = P(t \leq T_i < t + dt|T_i \geq t, \mathcal{F}_t).
\]

Independent censoring exempts us from estimating or concerning ourselves with the distribution of censoring times. For this reason the independent censoring assumption is ambiguous in survival analysis. Censored survival analysis can be written as a counting process by defining the censored counting processes

\[
N^c_i(t) = 1 \left\{ \tilde{T}_i \leq t, D_i = 1 \right\}
\]

for \(i = 1, ..., n\) and \(t > 0\). A jump in \(N^c_i(t)\) will occur only if obligor \(i\) experiences an event at time \(\tilde{T}_i\). 

93
The intensity of $N^c_i(t)$ is

\[
\lambda^c_i(t) = \frac{P(dN^c_i(t) = 1|\mathcal{F}_t)}{dt} = \begin{cases} 
P(t \leq \tilde{T}_i < t + dt, D_i = 1|\mathcal{F}_t)/dt & \text{if } \tilde{T}_i \geq t \\
0 & \text{otherwise}.
\end{cases}
\]

\[
= \text{1}\{\tilde{T}_i \geq t\} P(t \leq T_i < t + dt|T_i \geq t)/dt
\]

\[
= Y^c_i(t)\alpha_i(t),
\]

where we now set the at-risk indicator equal to $Y^c_i(t) = \text{1}\{\tilde{T}_i \geq t\}$ and $\alpha_i(t)$ is the hazard rate as defined in (4.1).

Aggregating over all obligors we obtain the process

\[
N^c(t) = \sum_{i=1}^n N^c_i(t)
\]

with intensity

\[
\lambda^c(t) = \sum_{i=1}^n \lambda^c_i(t) = \sum_{i=1}^n Y^c_i(t)\alpha_i(t).
\]

\[
Y^c(t) = \sum_{i=1}^n Y^c_i(t)
\]

is the number of obligors at-risk at time $t^-$. The intensity $\lambda^c_i(t) = Y^c_i(t)\alpha_i(t)$ is known as the multiplicative intensity model. We are now able to compile a portfolio of obligors, each with their own counting process $N^c_i(t)$ with corresponding intensities

\[
\lambda^c_i(t) = \alpha_i(t) Y^c_i(t).
\]

We conclude by summarizing that survival data, complete or incomplete, can be rewritten in a counting process framework with the appropriate choice of counting process $N^*(t)$ or $N^c(t)$. Hence, survival analysis can be seen as a counting process where only a single jump (event or death) occurs. All survival analysis models are driven by hazard rates, whereas counting processes are driven by intensities. Modelling these rates is usually done by some regression model, such as the Cox proportional hazard model in (4.4). The understanding of regression models like Cox’s proportional hazard model was
greatly enhanced by the counting process approach. Andersen and Gill (1982) were the first to make the basic connection between Cox’s model and counting processes. They used martingale theory to derive asymptotic properties of coefficients and extended the Cox model. These extensions include, among other things, allowing individual processes to experience recurring events. In this study, we will apply some regression model within the counting process framework to model credit events.

In the next section, we will discuss some of the characteristics of the data. These characteristics dictate what type of regression model is applicable.

### 4.4 Counting-process model requirements

In survival analysis, the event that needs to be modelled is a single non-recurring event. Traditional credit-risk models are concerned with modelling/predicting either default or the write-off event. However, in consumer credit risk we encounter recurring events. An obligor can default and then recover by making a payment. This same obligor can default again after a few months. Survival analysis would not be a suitable approach to model/predict defaults in this instance. Our aim is to use counting processes to model/predict the intensity of an obligor’s experience of one or more of the same credit events. We make use of regression-type intensity functions that use systematic and idiosyncratic covariates. In this section we highlight the typical data characteristics of consumer data. Additionally, we highlight some changes that need to be made to counting processes to make them compliant with our data. We start off by explaining the various time measurement scales available. We then discuss how we discretize the intensity model to account for our monthly observed discrete data.

#### 4.4.1 Calendar time and delayed entry

In survival analysis, the time to death or defect is of primary concern. A starting time is defined to be the onset of a disease or the manufacturing date of a product. The duration is measured as the time spent from the starting time to the death or credit event. Not all observations have the same starting time, but this is not problematic as the study only models duration. Hence, all delayed-entry observations are realigned to start at a
predefined time. Calendar time plays no role.

There are at least two time domains: calendar time and the ‘duration’ or age of a loan. Consumer credit risk is a function of both. Survival analysis applied to credit-risk management is not a new idea; Tong et al. (2012) use survival analysis to estimate default probabilities. They do not include any macro-economic factors and only model single default events where the time domain is the age of the loan (duration time). We are interested in the effect of, among other factors, calendar-time-dependent external covariates on the credit-event intensity of obligors. Since macro-economic factors vary over calendar time, we deem it crucial to use this time domain as baseline time. As a result, the approach by Tong et al. (2012) may be a bit too naïve. Furthermore, a time re-alignment of the data changes the natural real-time sequence of events. Such a re-alignment can “destroy the natural martingale structure of the real time counting-process models” (Arjas 1985).

Hazard functions that include two time domains are known as two-way hazard rates. As previously mentioned, we have two time domains: (1) duration and (2) calendar time. A heuristic understanding of the interrelationship between these two time dimensions can be obtained by using a Lexis diagram. Lexis diagrams were developed by Lexis in the 1870s and are described by Keiding (1990). The horizontal axis of the Lexis diagram is calendar time \( t \), and the vertical axis is age or duration, \( t^d \). If obligor \( i \) opens an account on calendar date \( O_i \) and the account experiences a credit event or is censored at calendar time \( \tilde{T}_i \), then the Lexis diagram would plot obligor \( i \) as a straight line that connects the points \((O_i, 0)\) with \((\tilde{T}_i, \tilde{T}_i - O_i)\). All the observations are then plotted on one graph to produce the Lexis diagram. The Lexis diagram also conveys information on whether an obligor had a credit event or was censored at the end of the observation period \((\tilde{T}_i, \tilde{T}_i - O_i)\). This is done by either plotting a solid dot or a circle at \((\tilde{T}_i, \tilde{T}_i - O_i)\). In Figure 4-1 a Lexis diagram is simulated. The solid dots indicate defaults, while the other dots are censored events. In the case of recurrent events, an accurate Lexis diagram would contain solid dots on the line of each obligor at each credit event point.

Several approaches to including both time scales in a regression-type counting-process model exist. The inclusion of the calendar-time measure into a (duration-time) Cox proportional hazard model is often done by treating calendar time as a covariate of the
model. Calendar time can also be used as a baseline time scale; this can easily be done by noting that if we observe some calendar time \( t \in (O_i, T_i) \) then it holds that the duration is \( t^d = t - O_i \).

Efron (2002) uses a Poisson-type counting-process model to compare three different types of hazard models: (1) the two-way hazard-rate model with hazard rate a function of both \( t^d \) and \( t \); (2) the one-way hazard-rate model where the baseline time scale is \( t^d \) and (3) another one-way model where the baseline time scale is \( t \). After estimating the models, he finds that the estimates are closely related, both empirically and theoretically.

Although it would seem that choosing a time scale is purely a matter of preference, we use the calendar time scale for specific reasons. Given the tendency of economic and political environments to change over time, we feel that the age of a loan would not suffice as the primary time domain. We consider including the age of the loan as a covariate. Also, when stress-testing a portfolio that is dependent on macro-economic variables, using calendar time as the base time scale simplifies matters. We follow what is known as the ‘real-time’ approach; see Andersen et al. (1993 Chapter 9).

When using the calendar time as the baseline time scale, we should note that the
traditional interpretation of a hazard rate changes slightly. Now we interpret the hazard rate \( \alpha(t) \) as the intensity of an event occurring at calendar time \( t \). This differs from using duration as the baseline timescale where we interpret \( \alpha(t^d) \) as the hazard rate of an obligor of age \( t^d \). The difference is in the interpretation of the shape of the hazard-rate function drawn over \( t \) or \( t^d \). For duration timeline studies the shape of the plotted hazard-rate function \( \alpha(t^d) \) tempers down to zero if \( t^d \to \infty \) or when censoring is present: \[
\lim_{t^d \to \infty} \alpha(t^d) = k
\]
for some constant \( k \). In contrast, the hazard rate at calendar time \( \alpha(t) \) only conveys information of the hazard rate at a specific calendar date. The function \( \alpha(t) \) does not temper down or exhibit any particular shape that can be interpreted in the same sense as \( \alpha(t^d) \). Also, in survival analysis the objective is to predict a future hazard rate for individuals given their projected age. To achieve this in the calendar-time domain, the time-dependent covariates need to be forecasted forward so that the future hazard rates can then be estimated. Hence, knowing only the age of an observation does not necessarily dictate the occurrence of an event.

Sequential delayed-entry data, which is common in retail credit risk, can easily be incorporated into the counting process framework. We observe the trio \((O_i, \tilde{T}_i, D_i)\), where \( O_i \) is the calendar date when the obligor opens his or her account, \( D_i \) is the default indicator and \( \tilde{T}_i = \min(T_i, C_i) \), where \( T_i \) is the event calendar time and \( C_i \) is the censoring calendar time. The at-risk indicator of this obligor is simply \( Y_i(t) = 1\{O_i < t \leq \tilde{T}_i\} \) (Aalen et al. 2008 p. 32; Arjas 1985). An added benefit of the real-time approach is that it allows obligors to be entered and removed from the at-risk set numerous times.

To summarize, counting processes allow us to model events in duration or calendar time. Additionally, features in the data such as delayed entry, numerous credit events and left censoring are easily incorporated. As a result, we prefer counting processes to survival analysis, which was originally developed for left-censored data (Aalen et al. 2008 p. 5).

4.4.2 Discrete-time regression model

Most survival models are formulated in continuous time; however, in practice, we observe only discrete time. It is irrelevant how small the time intervals are: we still have discrete observations. Although continuous models are a natural way to describe phenomena in
time, discrete time is more practical. Our data is observed monthly, so we need a discrete model. In continuous-time counting processes, a key assumption is that \( dN(t) \leq 1 \). In other words, the aggregated counting process \( N(t) \) in (4.13) may only jump by a size of one, and hence obligors may not experience credit events simultaneously (known as a tied event). Should ties occur, the exact timings of their occurrence are transformed during maximum-likelihood estimation to avoid this limitation. Unfortunately, these transformation techniques for handling tied credit event data are computationally expensive (Kalbfleisch & Prentice 2002 p. 104). If we use discrete-time counting processes, jumps by more than one in \( N(t) \) are perfectly feasible for all \( t \) (Kalbfleisch & Prentice 2002 p. 164).

Assume we split a continuous time scale into evenly spaced time intervals \([t_0, t_1), [t_1, t_2), ..., [t_j, t_{j+1}), ...\). A discrete-time hazard-rate model will have discrete time points \( t = t_0, t_1, ... \) with \( t_0 = 0 \) being the ‘origin’ of time, which is an arbitrarily chosen calendar date. The continuous hazard rate in (4.1) may be written in the discrete form as

\[
P_{t_j} = P(t_j \leq T < t_{j+1} | T \geq t_j).
\] (4.15)

This is the probability of an event occurring at a specific time, given that the event has not yet occurred. Again, we are faced with the estimation of some hazard rate (in this case a probability) that we would like to model as a regression function with idiosyncratic and systematic covariates. Throughout this study we assume that all systematic covariates are external and predictable in the technical sense.

Allison (1982) states that we can approach discrete models in two ways: (1) treating time as discrete, so events may only occur at discrete times and use discrete models; or (2) starting with a continuous-time model (i.e. assuming time is continuous) and then estimating the model parameters by making allowances for discrete data. Allison (1982) notes that regardless of the approach followed, the results are "remarkably similar". To illustrate the two discrete modelling techniques emphasized by Allison we can estimate the probability in (4.15) via a regression model in two ways.

On the one hand we can assume that time is discrete and only make use of discrete-time models. So we may model \( P_{t_j} \) as a GLM with link function such that \( \hat{P}_{t_j} \in [0, 1] \). Assuming that we use the logit link function and that \( P_{t_j} \) is a function of some explanatory
time-dependent variables we get

$$\text{logit} \left( P_{t_j} \right) = \alpha_0 (t_j) + \beta^T X(t_j), \quad (4.16)$$

where $\alpha_0 (t_j)$ is the discrete baseline hazard function, $X(t_j)$ is the time-dependent design matrix without a constant term and $\beta$ is the coefficient vector at time $t_j$, $j = 0, 1, \ldots$.

On the other hand, we can assume that the continuous model in (4.4) holds. Substituting (4.3) and (4.4) into (4.15) we get

$$P_{t_j} = P(t_j \leq T < t_{j+1}|T \geq t_j)$$

$$= \frac{[F(t_{j+1}) - F(t_j)]}{S(t_j)}$$

$$= 1 - \exp \left[ - \int_0^{t_{j+1}} \alpha(s) ds + \int_0^{t_j} \alpha(s) ds \right]$$

$$= 1 - \exp \left[ - \int_{t_j}^{t_{j+1}} \alpha(s) ds \right]$$

$$= 1 - \exp \left[ - \int_{t_j}^{t_{j+1}} \exp \left( \alpha_0 (s) + \beta^T X(s) \right) ds \right]$$

$$= 1 - \exp \left[ - \exp \left( \alpha_0 (t_j) + \beta^T X(t_j) \right) \right]. \quad (4.17)$$

The last equality holds if all values of $t$ and $X(t)$ are assumed to be constant over the interval $[t, t + 1)$. Rearranging (4.17) we get

$$\log \left( - \log \left( P_{t_j} \right) \right) = \alpha_0 (t_j) + \beta^T X(t_j), \quad (4.18)$$

which is the complementary log-log link function. So, the discrete version of the continuous Cox proportional hazard-rate model is a GLM with a complementary log-log link function.

The difference between the complementary log-log link function and the logit link function is small, and they only really differ at the tails of the distributions. Interestingly,
Thompson (1977) finds that if the time intervals are short, the logit model converges to the proportional hazard model. This confirms the remark by Allison (1982) that they were "remarkably similar".

Estimating the unknown $\alpha_0(t_j)$ may be done in several ways. The following, among other things, are possible: (1) assuming a constant baseline hazard rate over time

$$\alpha_0(t_j) = \alpha$$ (4.19)

for $j = 0, 1, 2, \ldots$ (2) using a linear function of time, $\alpha_0(t_j) = \alpha_0 + \alpha_1 t_j$ or (3) using other functions such as $\alpha_0(t_j) = \alpha_0 + \alpha_1 \log(t_j)$ (Allison 1982). Other methods of estimating $\alpha_0(t_j)$ assume that the baseline hazard is constant over every discrete-time interval. Following this approach, Prentice and Gloeckner (1978), among others, assume that if there are $k$ time intervals during the study period, the baseline hazard function is given by

$$\alpha_0(t) = \begin{cases} 
\alpha_0 & \text{for } t_0 \\
\alpha_1 & \text{for } t_1 \\
\vdots \\
\alpha_k & \text{for } t_k.
\end{cases}$$ (4.20)

This $\alpha_0(t)$ in (4.20) adds $k - 1$ additional parameters (coded as dummy variables) to their proportional hazard model. This approach only seems feasible over a short period of time, since the degrees of freedom decrease with each additional time-step. Kalbfleisch and Prentice (2002 p. 136) detail the estimation method of $\alpha_0(t)$ using Newton-Raphson iterations for binary GLM models where various link functions are used. Note that by specifying the baseline hazard-rate model we are effectively postulating a parametric model.

Applying discrete-time counting processes to credit risk is also not new. The approach by Shumway (2001) is to substitute (4.19) into (4.16). Another model akin to (4.18) with (4.20) was proposed by Ding et al. (2012). In the next chapter we will discuss these two models, as well as the real-time approach by Arjas and Haara (1987).
Chapter 5

The models: theory

In this chapter we will explore two approaches to modelling the occurrence of a default event. We begin by exploring the model by Ding, Tian, Yu, and Guo (2012), which will be abbreviated as DTYG. We compare the DTYG model to the model that it evolved from: the Shumway model (Shumway 2001). We then discuss a promising real-time logistic regression approach by Arjas and Haara (1987). In this chapter we discuss the theory, and the next chapter will apply and compare the prediction performance of both of these models on the credit-risk data.

5.1 The Shumway and DTYG models

Shumway (2001) was the first to exploit the natural correspondence between default-risk modelling and survival analysis. Shumway (2001) departs from the traditional static models in that his approach uses time-varying cross-sectional data. He used a longitudinal approach to model the default events of companies. In contrast to our needs, Shumway (2001) considered a default as a ‘terminal’ event, i.e. once an obligor has defaulted, he is written off. He used survival analysis to predict a terminal default using various covariates, which may be time-dependent. This is an approach well suited to the modelling of corporate credit risk. However, when modelling consumer credit risk we should take repeated defaults into account. Hence, the ‘once-off’ approach pioneered by Shumway (2001) (or the terminal default model that is commonplace in corporate credit-risk modelling) will not suffice for consumer data in general.
In essence, the Shumway-type models are discrete versions of Cox’s model (Cox 1972). Malik and Thomas (2012) developed a type of Cox proportional hazard model to estimate default probabilities for retail loans. They use as covariates the behavioural score, the age of the loan and an "economic variable". They find that macro-economic covariates play a significant role in the prediction of default. A survival-analysis-type model was also developed by Belotti and Crook (2009). Their paper exploits the use of time-dependent macro-economic variables to predict credit card defaults. We will also include macro-economic covariates in our models.

Say we observe the triple \((O_i, \tilde{T}_i, D_i)\), where \(O_i\) is the \(i^{th}\) obligor’s opening date, \(D_i\) is the default indicator and \(\tilde{T}_i\) is the observed exit time with \(\tilde{T}_i = \min(T_i, C_i)\), where \(T_i\) is the default time and \(C_i\) is the censoring time. Shumway (2001) uses a multi-period logistic regression and sets the probability of default for obligor \(i\) at time \(t_j\) equal to

\[
\pi_{i,t_j} = P \left( T = t_j | \tilde{T}_i \geq t_j, X_i(t_j) \right) = \frac{1}{1 + \exp(-\alpha + \beta^\top X_i(t_j))} \tag{5.1}
\]

where the time can take on discrete values \(t = t_1, t_2, \ldots\) and \(X_i(t_j)\) is the covariate vector of firm \(i\) at time \(t_j\). The parameter \(\beta\) is a column vector and is used to rank the default risk of companies. Higher values of \(\beta^\top X_i(t_j)\) indicate higher default risk at time \(t_j\) for firm \(i\). Hence, \(\beta^\top X_i(t_j)\) can be considered to be a type of credit score, similar to the ratings published by Moody’s and S&F for corporate institutions. Shumway’s (2001) model can be applied in calendar time and the likelihood in the model is assumed to be

\[
Lik(\theta) = \prod_{i=1}^{n} \prod_{j:O_i < t_j \leq \tilde{T}_i} \pi_{i,t_j}^{D_{i,t_j}} (1 - \pi_{i,t_j})^{1 - D_{i,t_j}}, \tag{5.2}
\]

where \(D_{i,t_j} = D_i \mathbf{1} \left\{ \tilde{T}_i = t_j \right\}\) with the parameter column vector \(\theta = [\alpha, \beta^\top]^\top\). Hence, all observations are assumed to be independent and all observations of an obligor over time are also assumed to be independent, conditional on the realization of \(X(t)\).

The model in (5.1) can be extended by using time-varying intercept terms as in (4.20).
We set the default probabilities equal to

\[ \pi_{i,t_j} = \frac{1}{1 + \exp(-\alpha_{t_j} - \mathbf{\beta}^T \mathbf{X}_i(t_j))}. \]  

(5.3)

As stated in Section 4.4.2, if the observation time span is long, this model contains many additional parameters. This model is simply (4.20) substituted into (5.2).

The DTYG approach takes the Shumway model a step further. They put forward the following logarithmic transformation, dependent on an additional transformation parameter \( c \):

\[ \pi_{i,t_j} = \begin{cases} 
1 - \{1 + c \exp(\alpha_{t_j} + \mathbf{\beta}^T \mathbf{U}_i(t_j - 1))\}^{-1/c} & \text{if } c > 0 \\
1 - \exp\{-\exp(\alpha_{t_j} + \mathbf{\beta}^T \mathbf{U}_i(t_j - 1))\} & \text{if } c = 0.
\end{cases} \]  

(5.4)

This model is also known as the log-Pareto model. In addition, Ding et al. (2012) also propose a Box-Cox type transformation with an additional transformation parameter \( \rho \):

\[ \pi_{i,t_j} = \begin{cases} 
1 - \{1 + 1/\rho (\exp[\rho \exp(\alpha_{t_j} + \mathbf{\beta}^T \mathbf{U}_i(t_j - 1))]) - 1\}^{-1} & \text{if } \rho > 0 \\
1 - \{1 + \exp(\alpha_{t_j} + \mathbf{\beta}^T \mathbf{U}_i(t_j - 1))\}^{-1} & \text{if } \rho = 0.
\end{cases} \]  

(5.5)

Note that when \( c = 1 \) or \( \rho = 0 \) the models (5.4) and (5.5) simplify to the binary-regression model with a logit link function and time varying intercept terms as in (5.3). When \( c = 0 \) or \( \rho = 1 \), the models (5.4) and (5.5) are binary GLMs with their link function the complementary log-log function (4.17). In this case, the model is also known as the ‘grouped relative risk model’. Both of these models incorporate the two different GLM link functions. The log-likelihood for both these models is

\[ \text{lik}(\theta) = \sum_{i=1}^{n} \sum_{j:O_i < t_j \leq T_i} D_{i,t_j} \log(\pi_{i,t_j}) + (1 - D_{i,t_j}) \log(1 - \pi_{i,t_j}), \]  

(5.6)

with \( \pi_{i,t_j} \) given by either (5.4) or (5.5) and \( \theta = [\mathbf{\beta}^T, \alpha_1, ..., \alpha_k, c]^T \) or \( \theta = [\mathbf{\beta}^T, \alpha_1, ..., \alpha_k, \rho]^T \) respectively.
To maximize this log-likelihood, Ding et al. (2012) apply a profile-likelihood method in three steps:

1. For a fixed \( c \) or \( \rho \), initialize the parameter estimates \( \hat{\beta}^{(0)} \) and \( \hat{\alpha}_{ij}^{(0)} \) for \( j = 1, 2, \ldots, k \).
   (They propose as initial values zeros, ones or the parameter estimates from the application of the Shumway model.)

2. Given \( \hat{\beta}^{(i)} \), find \( \hat{\alpha}_{ij}^{(i+1)} \) for \( j = 1, \ldots, k \) by maximizing (5.6).

3. Given \( \hat{\alpha}_{ij}^{(i+1)} \) for \( j = 1, \ldots, k \), estimate \( \hat{\beta}^{(i+1)} \) by maximizing (5.6).

They iterate the last two steps until the log-likelihood function converges. This process is repeated for different values of \( c \) and \( \rho \), and they choose the estimate of \( c \) and \( \rho \) where the likelihood is maximized.

We found that the estimation of the parameters on a personal computer (Intel Core i3, CPU 2.53 GHz with 4 GB RAM) using the prescribed approach takes a long time (typically three to four days). Private correspondence with the authors confirmed that when they estimated the parameters on their portfolio (much smaller than ours) using Matlab®, it was also time-consuming. They mentioned that “it may take up to 4 or 5 days”. The numerical maximization procedure used by Matlab® is the Nelder–Mead simplex method. This method is known as a ‘direct’ method, as it requires no derivatives for function minimization.

We developed an alternative approach to estimate the parameters of (5.6) that brought down the computation time to less than a minute. Setting \( \alpha = (\alpha_1, \ldots, \alpha_k)^\top \), we can rewrite the overall maximum of the likelihood as

\[
\max_{c, \beta, \alpha} \{ \text{lik} (c, \beta, \alpha) \} = \max_{c^*} \left\{ \max_{\beta, \alpha} \text{lik} (c^*, \beta, \alpha) \right\}.
\]

Fixing the value of \( c \) at \( c^* \) we are able to find the maximizers of the likelihood for \( \text{lik} (c^*, \beta, \alpha) \) by means of the Newton-Raphson method. We calculate the likelihood values at different values of \( c^* \) and then maximize over \( c^* \). The same can be done for (5.5) dependent on \( \rho \). The full details of estimating either (5.4) or (5.5) are discussed in the appendix at the end of this chapter. The reason why this approach is so much faster is that the Newton-Raphson method always requires far fewer iterations than ‘direct’ methods (Olsson & Nelson 1975).
In general, the Newton-Raphson optimization procedure does not guarantee that the
maximum is global. Hence, to avoid local maxima, we used various sets of starting values.
We chose the starting values of a vector of ones or zeros, or used the parameter estimates
of the Shumway model in (5.1). In all cases the estimated values were the same, so we
are comfortable that the estimated parameters are indeed at the maximum-likelihood.

Although the DTYG model seems more flexible than the Shumway model, it does so
at the cost of added parameters: each time step needs its own intercept. Furthermore,
the model only allows one default per obligor. Finally, macro-economic covariates cannot
be included in the covariate matrix \( U(t) \), because this leads to a singular design ma-
trix. Macro-economic covariates are fixed at each time step and influence all individuals.
Including two or more covariates (for example the intercept term and another macro-
economic covariate) that are fixed on each time for all obligors and assuming that each
time step is independent of the others, results in a singular design matrix.

Ding et al. (2012) prove the asymptotic normality of the parameter estimates. Assum-
ing that some regularity conditions hold they prove that for a fixed number of time periods
\( k \), the estimates \( \hat{\theta} \) of the parameters \( \theta = [\beta^T, \alpha_1, \alpha_2, ..., \alpha_k, c] \) or \( \theta = [\beta^T, \alpha_1, \alpha_2, ..., \alpha_k, \rho] \)
are consistent and that \( \sqrt{n} (\hat{\theta} - \theta_0) \) is asymptotically normal with a mean of zero and
covariance \( [I(\theta_0)]^{-1} \).

Ding et al. (2012) propose the following data-based estimate of the Fisher information
matrix:
\[
\hat{I}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{j: O_i < t_j \leq T_i} \left( \frac{D_{i,t_j}}{\pi_{i,t_j}} - \frac{1 - D_{i,t_j}}{1 - \pi_{i,t_j}} \right) \frac{\partial}{\partial \theta} \pi_{i,t_j} \right\} \otimes 2, \tag{5.7}
\]
where \( \mathbf{v} \otimes 2 = \mathbf{v} \mathbf{v}^T \). The DTYG approach does not give an expression for \( \frac{\partial}{\partial \theta} [\pi_{i,t_j}] \), but
we derive one in the appendix at the end of this chapter. One of the standard regularity
conditions is that the true parameter value \( \theta_0 \) should be an interior point of the parameter
space. So, should the estimated value of either of the two transformation parameters be
zero (\( c = 0 \) or \( \rho = 0 \)), Ding et al. (2012) propose that the other model transformation
family be used (with \( c = 1 \) or \( \rho = 1 \) respectively). Next we discuss the approach of Arjas
and Haara (1987).
5.2 The Arjas and Haara real-time logistic regression model

All the unique features of the data in this study and its distinct model requirements are addressed by Arjas and Haara’s (1987) ‘real-time’ approach. Arjas (1985, 1986) developed a discrete approach to modelling incomplete failure time data. Arjas and Haara (1987), henceforth abbreviated to AH, developed a discrete-time model that allows for multiple events (in our context, credit events) in calendar time. The model proposed by AH, as well as the estimation of parameters are discussed in this section.

The AH model is fully parametric, and in our case the hazard rate of an obligor is dependent on the current prevailing macro-economic conditions. The baseline time is calendar time. This is in contrast to classical survival analysis, where the baseline hazard rate is specified as a function of some other conveniently chosen measurement of time, such as time since the diagnosis of patient with a disease. Hazard curves cannot be interpreted in the same way as in survival analysis; they should now be interpreted as the hazard rate of an obligor at a specific calendar time. A main feature of this model is the use of logistic regression as the primary model structure. The estimation of parameters will be done with the use of partial likelihood maximization. We use the logistic-regression-type model proposed by AH to model credit events.

We now describe the AH model, which is a logistic-regression-type model. This model can be applied to modelling failures in discrete time for incomplete data. For practical reasons, we make use of a discrete-time model, and as is usual in credit-risk literature, we assume the conditional independence of credit events given the current prevailing economic conditions (Giesecke 2004). Obligors may have delayed entry times and not all obligors are assumed to be present at any calendar time \( t \geq 1 \). Obligors may also fail several times. For obligor \( i = 1, \ldots, n \) and time index \( t = 1, 2, \ldots \) we let the left continuous at-risk indicator of obligor \( i \) be

\[
Y_i(t - 1) = \begin{cases} 
1 & \text{if obligor } i \text{ is at-risk at time } t \\
0 & \text{otherwise.}
\end{cases}
\]
The counting process for the failure of an obligor is

\[ \Delta N_i(t) = \begin{cases} 
1 & \text{if obligor } i \text{ experiences a failure during } (t - 1, t] \\
0 & \text{otherwise},
\end{cases} \]

where \( \Delta N_i(t) = N_i(t) - N_i(t - 1) \). Note that we use the notation \( \Delta N \) instead of \( dN \) to emphasize the discrete nature of this model. Also, \( \Delta N(t) = 1 \) implies a failure occurs at discrete-time \( t \) exactly. Denote the risk set by \( R(t - 1) = \{ i \geq 1 : Y_i(t - 1) = 1 \} \) in \( (t - 1; t] \). The size of the risk set at time \( t - 1 \) is the number of obligors at-risk:

\[ \text{card}[R(t - 1)] = \sum_i Y_i(t - 1), \]

where \( \text{card} \) is the cardinality, or size, of the set and we assume \( R(t - 1) \) is finite for all \( t \geq 1 \). Suppose we know the \((1 \times p)\) covariate vector at time \( t \), \( U_i(t) = (U_{i1}(t), \ldots, U_{ip}(t)) \), which may contain covariates that are constant over time for each obligor \( i \). \( \Delta N_i(t) \) is assumed to be a Bernoulli random variable with the probability of default a function of \( U_i(t) \). Hence, we follow a regression model approach to modelling the credit events.

We follow AH in defining the \( \sigma \)-field inductively. Let \( F_t^{(i)} \) and \( G_{t-1}^{(i)} \), \( t \geq 0 \) be the \( \sigma \)-field for each obligor \( i \), then

\[
\begin{align*}
G_{t-1}^{(i)} &= F_{t-1}^{(i)} \lor \sigma \{ Y_j(t - 1), U_i(t - 1) \}, \\
F_t^{(i)} &= G_{t-1}^{(i)} \lor \sigma \{ \Delta N_i(t) \}.
\end{align*}
\]

We interpret \( G_{t-1}^{(i)} \) as the at-risk set of obligor \( i \) and its covariates at time \( t \) excluding any default at time \( t \). \( F_t^{(i)} \) is the union of \( G_{t-1}^{(i)} \) and the defaults in \( (t - 1, t] \). So, \( F_t^{(i)} \) contains all the information of the at-risk set, covariates and default events on time \( t \).

The filtration is defined inductively to allow for the use of partial likelihood estimation, as will become apparent later. The initial information is contained in the \( \sigma \)-field \( F_0 \). We explain the derivation of the full likelihood and the partial likelihood expressions that will be used for parameter estimation below. Aggregating all obligors \( i = 1, 2, \ldots, n \) the \( \sigma \)-fields are
The likelihood for obligor $i$ with parameter set $\theta$ is

\[
\mathcal{L}_{i}^{(t)}(\theta)
\]

\[
= P[\bigcap_{v \leq t} Y_{j}^{(v - 1)} = y_{i}^{(v - 1)}, U_{i}^{(v - 1)} = u_{i}^{(v - 1)}, \Delta N_{i}^{(v)} = \Delta n_{i}^{(v)}]
\]

\[
= P[Y_{j}^{(0)} = y_{i}^{(0)}, U_{i}^{(0)} = u_{i}^{(0)}]
\]

\[
\times P[\Delta N_{i}^{(1)} = \Delta n_{i}^{(1)}| Y_{j}^{(0)} = y_{i}^{(0)}, U_{i}^{(0)} = u_{i}^{(0)}]
\]

\[
\times P[Y_{j}^{(1)} = y_{i}^{(1)}, U_{i}^{(1)} = u_{i}^{(1)}| \mathcal{G}_{i}^{(1)}, \Delta N_{i}^{(1)} = \Delta n_{i}^{(1)}]
\]

\[
\times P[\Delta N_{i}^{(2)} = \Delta n_{i}^{(2)}| \mathcal{F}_{i}^{(2)}, Y_{j}^{(1)} = y_{i}^{(1)}, U_{i}^{(1)} = u_{i}^{(1)}]
\]

\[
\times P[Y_{j}^{(2)} = y_{i}^{(2)}, U_{i}^{(2)} = u_{i}^{(2)}| \mathcal{G}_{i}^{(2)}]
\]

\[
\times P[\Delta N_{i}^{(3)} = \Delta n_{i}^{(3)}| \mathcal{G}_{i}^{(2)}]
\]

\[
\vdots
\]

\[
\times P[Y_{j}^{(t - 1)} = y_{j}^{(t - 1)}, U_{i}^{(t - 1)} = u_{i}^{(t - 1)}| \mathcal{F}_{i}^{(t - 1)}]
\]

\[
\times P[\Delta N_{i}^{(t)} = \Delta n_{i}^{(t)}| \mathcal{G}_{i}^{(t - 1)}]
\]
Thus, the full likelihood can be decomposed into two factors: the part dependent on $F_i^{(i)}$ and the part dependent on $G_i^{(i)}$. This factorisation leads to the following assumptions by Arjas and Haara (1987).

**Assumption 1** Arjas and Haara (1987)

The parameter set $\theta$ can be decomposed into two $\theta = (\beta^*, \beta)$: (i) the first factor of the full likelihood in (5.8),

$$
\prod_{v \leq t} P \left[ Y_i (v - 1) = y_i (v - 1), U_i (v - 1) = u_i (v - 1) | F_v^{(i)} \right],
$$

which is assumed to only be dependent on $\beta^*$ and not $\beta$; and (ii) the second factor,

$$
\prod_{v \leq t} P \left[ \Delta N_i (v) = \Delta n_i (v) | G_v^{(i)} \right],
$$

is assumed to only be dependent on the parameters $\beta$ and not $\beta^*$. This second factor (5.9) is similar to Cox’s partial likelihood. Cox’s partial likelihood treat the unobserved baseline hazard rate as a nuisance parameter, and here we make a similar assumption. Also, because data is revealed sequentially over time, (5.9) is known as a partial likelihood function. Note that a partial likelihood function should not to be confused with a conditional likelihood function (Slud & Kedem 1994). We are only interested in $\beta$, so $\beta^*$ is treated as a nuisance parameter.

Applying this assumption, the partial likelihood for the parameters $\beta$ is

$$
Lik_i^{(i)} (\beta) = \prod_{v \leq t} P \left[ \Delta N_i (v) = \Delta n_i (v) | G_v^{(i)} \right],
$$

for $v \geq 1$ and $i = 1, ..., n$. We can simplify this assumption further by using the second assumption made by Arjas and Haara.
Assumption 2 Arjas and Haara (1987)

For each $v \geq 1$ and $i = 1, \ldots, n$ the random variables $\Delta N_i(v)$ are conditionally independent given $G_{v-1}^{(i)}$. This conditional independence assumption is made extensively throughout credit-risk modelling literature: it is assumed that the model covariates will account for all dependence between obligors. Assumption two, together with the fact that only at-risk obligors make a contribution to the likelihood imply that we can combine the partial likelihood in (5.10):

$$
\text{Lik}_t(\beta) = \prod_{R(v-1)} \text{Lik}_t^{(i)}(\beta).
$$

So

$$
\text{Lik}_t(\beta) = \prod_{R(v-1)} \prod_{v \leq t} P \left[ \Delta N_i(v) = \Delta n_i(v) | G_{v-1}^{(i)} \right].
$$

(5.11)

We further simplify the partial likelihood by following Arjas and Haara in making the next assumptions.

Assumption 3 Arjas and Haara (1987)

$\Delta N_i(v)$ and $G_{v-1}^{(i)}$ are conditionally independent given $Y_i(v-1)$ and $U_i(v-1)$ for all $v, i \geq 1$. That is, for all $B \in G_{v-1}^{(i)}$

$$
P \left[ \Delta N_i(v) = \Delta n_i(v) | B, Y_i(v-1), U_i(v-1) \right]
= P \left[ \Delta N_i(v) = \Delta n_i(v) | Y_i(v-1), U_i(v-1) \right].
$$

This assumption ensures that all relevant information in $G_{v-1}^{(i)}$, when used as conditioning for the probability of the events $\{ \Delta N_i(v) = \Delta n_i(v) \}$, is contained in $U_i(v-1)$ and $Y_i(v-1)$. Therefore, the partial likelihood function is only dependent on the information contained in the previous time step. Note that we do not impose a Markovian assumption on this process, since the values of covariates at time $t < v - 1$ may be included as covariates in the vector $U_i(v-1)$. This assumption reduces (5.11) further to

$$
\text{Lik}_t(\beta) = \prod_{v \leq t} \prod_{R(v-1)} P \left[ \Delta N_i(v) = \Delta n_i(v) | U_i(v-1), Y_i(v-1) \right],
$$

(5.12)

We now need to specify some statistical model for the probabilities in (5.12). Following Arjas and Haara we assume that the probabilities in (5.12) have a logit structural form. Any other link function could have been chosen.
Assumption 4 Arjas and Haara (1987)

The probabilities of the obligors that are at-risk are given by

\[
\text{logit}(P[\Delta N_i(v) = 1|U_i(v-1), Y_i(v-1) = 1]) = \mathbf{\beta}^\top U_i(v-1)
\]  

(5.13)

for all \(v \geq 0\). Rewriting (5.13) we get

\[
P[\Delta N_i(v) = 1|U_i(v-1), Y_i(v-1) = 1] = (1 + \exp (-\mathbf{\beta}^\top U_i(v-1)))^{-1}.
\]

Simplifying this further by removing the \(Y_i(v-1) = 1\) from the conditioning we get the probability of a failure occurring as

\[
L_i(v) := Y_i(v-1) (1 + \exp (-\mathbf{\beta}^\top U_i(v-1)))^{-1}.
\]  

(5.14)

\(Y_i(v-1) = 0\) implies that \(\Delta N_i(v) = 0\). Formula (5.14) is almost the same as the logistic regression model for obligor \(i\) at time \(v\), bar the at-risk indicator and time-dependent covariates. As we will see, this model turns out to be very similar to a weighted logistic regression model with weights \(Y_i(v-1)\).

If we substitute the probabilities \(L_i(v)\) in the partial likelihood \(L_i(\beta)\) (5.12) we get

\[
L_i(\beta)
\]

\[
= \prod_{v \leq t} \prod_{R(v-1)} P[\Delta N_i(v) = \Delta n_i(v)|U_i(v-1), Y_i(v-1)]
\]

(5.15)

\[
= \prod_{v \leq t} \prod_{i} \left(L_i(v)^{\Delta n_i(v)} (1 - L_i(v))^{1-\Delta n_i(v)} \right)^{Y_i(v-1)}.
\]

Hence, the log-likelihood is

\[
lik_i(\beta)
\]

\[
= \sum_{v \leq t} \sum_{i} Y_i(v-1)[\Delta n_i(v) \log (L_i(v)) + (1 - \Delta n_i(v)) \log (1 - L_i(v))].
\]  

(5.16)
The maximum-likelihood parameter estimate is found by setting the score vector equal to zero and solving for $\beta$. The score vector is

$$
U(\beta, t) = \frac{\partial \text{lik}(\beta, t)}{\partial \beta}
$$

and

$$
= \sum_{v \leq t} \sum_{i} Y_i (v - 1) \left[ \Delta n_i(v) \beta^T U_i(v - 1) - \log \left( 1 + \exp \left[ \beta^T U_i(v - 1) \right] \right) \right]
$$

(5.17)

The second partial derivative of the log-likelihood function with respect to $\beta_j$ and $\beta_k$ is

$$
\frac{\partial^2 \text{lik}(\beta, t)}{\partial \beta_j \partial \beta_k} = -\sum_{v \leq t} \sum_{i} U_{ij}(v - 1)U_{ik}(v - 1)Y_i(v - 1)V(\beta^T U_i(v - 1)),
$$

(5.18)

where $V(X) = L(X) \left[ 1 - L(X) \right]$. For the purposes of numerical calculation it is convenient to write (5.17) and (5.18) in matrix form. Let $U(v - 1)$ be a $(n \times p)$ matrix of $p$ covariates for $n$ obligors at time $v - 1$. Furthermore, let $\Delta n(v)$ and $Y(v - 1)$ be $(n \times 1)$ vectors containing the change in the counting process and the at-risk indicators respectively for all $n$ obligors. In matrix form, the log-likelihood at time $t$ is

$$
\text{lik}(\beta, t) = \sum_{v \leq t} Y^T(v - 1) \left[ \text{diag} (\Delta n(v)) U(v - 1) \beta - \log(1 + \exp(\text{U}(v - 1) \beta)) \right].
$$

The score vector is

$$
U(\beta, t) = \sum_{v \leq t} \text{U}(v - 1)^T \text{diag} [Y(v - 1)] \left[ \Delta n(v) - \{ L(\text{U}(v - 1) \beta) \}^T \right].
$$

The information matrix, or the negative of the second-order partial derivative of the log-likelihood function in (5.18), can be written as

$$
\tau(\beta) = -\frac{\partial^2 \text{lik}(\beta, t)}{\partial \beta \partial \beta^T} = -\sum_{v \leq t} \text{U}(v - 1)^T \text{diag} [Y(v - 1)] \text{diag} [V(\text{U}(v - 1) \beta)] \text{U}(v - 1).
$$

(5.19)

The value of $\beta$ that solves $\partial \text{lik}(\beta, t)/\partial \beta = 0$ is the parameter estimate $\hat{\beta}$ of the coeffi-
cents. Using Newton-Raphson iteration and an appropriate starting vector $\beta_0$ we can find the solution of $U(\beta, t) = 0$ iteratively by the algorithm

$$\hat{\beta}^m = \hat{\beta}^{m-1} + \left( I(\hat{\beta}) \right)^{-1} U(\hat{\beta}, t),$$

where $I(\hat{\beta})$ and $U(\hat{\beta}, t)$ are evaluated at $\hat{\beta}^{m-1}$, $m > 1$. The stopping criterion is the first $\hat{\beta}_m$ where $\left( lik(\hat{\beta}^m, t) - lik(\hat{\beta}^{m-1}, t) \right)^2 < \varepsilon$ for a small $\varepsilon > 0$. Using this IRLS estimation approach has the benefit of producing the standard errors of $\hat{\beta}$, namely

$$s(\hat{\beta}) = \left( \text{diag} \left( I(\hat{\beta}) \right)^{-1} \right)^{\frac{1}{2}}. \quad (5.20)$$

Arjas and Haara show that $\hat{\beta}$ is asymptotically normally distributed (Theorem 1, Arjas & Haara 1987). This property, as well as some others, will be discussed in the next section.

This model is surprisingly similar to a simple logistic regression model. Upon closer inspection, the likelihood in (5.15) with probability (5.13) looks very similar to the likelihood for a logistic regression model. The difference is in the addition of the time dimension, which is not a part of cross-sectional logistic regression. This is akin to the longitudinal approach by Diggle et al. (2009 Section 8). The likelihood in (5.15) leads us to believe that the AH model can be estimated by using standard GLM software. In fact, when we estimated the parameters by using the AH approach on some data and estimated parameters using conventional GLM software on the same data, we found that the parameter estimates and the standard errors of the parameter estimates were exactly the same.

To implement the AH model in GLM software the data needs to be transformed appropriately. To estimate the parameters of the AH model with GLM software we would proceed as follows: we treat each time observation for each obligor as a separate cross-classification observation. This is only done for the individuals that are at-risk in that time interval, i.e. for each pair $(i, t)$ such that $Y_i(t - 1) = 1$. The response variable is binary $\Delta N(t)$ and the covariates $U(t - 1)$ are used as explanatory covariates. For example, if an obligor was at-risk for 20 time intervals, we would create 20 separate observations (one for each time period during which the obligor was at-risk) with corresponding response and explanatory variables. This obligor will have 20 observations in a GLM model. Arjas
(1986) notes that representing the data in a multiple logistic regression model for binary data has some practical problems: if the number of time intervals is large, the logistic regression presentation has data matrices with enormous dimensions. Although this was a problem in 1986, computational power has increased significantly since then. However, we are nevertheless faced with a similar problem, as the data sets in this study are very large to begin with. The representation of the AH model as a GLM is similar to the presentation of Cox’s proportional hazard model as a GLM. The use of GLMs to solve survival-type models is not new: Whitehead (1980) shows that the partial likelihood for the Cox model is equivalent to the likelihood for a Poisson model. Whitehead (1980) goes on to show how to estimate the Cox proportional hazard parameter estimates by using GLIM (Generalized Linear Interactive Modelling) software.

The interpretation of the coefficients of the AH model is similar to that of the coefficients of a simple logistic regression analysis. In other words, a unit increase in $U_{i,p}(t-1)$ would result in an increase of $\exp(\beta_p)$ in the odds ratio.

Ideally, a separate model should be developed to independently predict censoring events (in our case, the early closure of an account). Should censoring occur the obligor must be removed from the at-risk set, i.e. $Y_i(s) = 0$ for $s \geq C_i$ where $C_i$ is the early closure time. We do not concern ourselves with the development of an early closure model, though this can be a useful extension to the model we propose.

During this study we use hypothesis tests quite frequently. Unless specifically stated, when we refer to a $p$-value, we refer to the Wald test $p$-value (Rao 1973 p. 417). Further detail on the various types of hypothesis tests is given in the appendix at the end of the chapter.

### 5.3 Asymptotic results for the AH model: a simulation study

This section provides context to the asymptotic theory developed by Arjas and Haara (1987). This is a practical application of the theory. Throughout this section we will use the symbol $\xrightarrow{D}$ to denote convergence in distribution and $\xrightarrow{P}$ to denote convergence in probability. Arjas and Haara (1987) prove that the parameter estimates are asympt-
totically normally distributed. They argue that if a study has delayed entry and the observation period is large ($t \to \infty$), a larger number of obligors will be included in the study. This is in contrast to the more traditional approach, where the time period is fixed but the number of obligors tends to infinity ($n \to \infty$). In the latter case, all obligors are assumed to be present at $t = 0$ and some alignment of the data points of obligors needs to be made. If $\beta_0$ is the true value of the parameter and $\hat{\beta}_t$ is the estimated parameter at time $t$, Arjas and Haara prove that as $t$ approaches infinity through the values $1, 2, \ldots$ (denoted as $t \to \infty$),

\[
c_t^{1/2} (\hat{\beta}_t - \beta_0) \xrightarrow{D} N(0, \Sigma_0^{-1}),
\]  

(5.21)

where the bounded continuous matrix defined as $\Sigma_0 := \Sigma (\beta_0)$, has property

\[
c_t^{-1} I^{(\beta)} \xrightarrow{P} \Sigma (\beta),
\]

where the information matrix $I^{(\beta)}$ is in (5.19). Here $c_t$ is a stable normalizing sequence such that $\lim_{t \to \infty} c_t = \infty$. Arjas and Haara suggest that when using the ‘total time on test’ an appropriate choice for $c_t$ would be $c_t = t$. They also suggest a random normalizing sequence, $C_t$, defined as

\[
C_t := \sum_{v \leq t} \sum_i Y_i(t),
\]

which they call the "total number of days at risk", as an appropriate stable sequence. Defining $C_t$ in this manner is only possible if there exists some constant $c \in (0, \infty)$ such that

\[
c_t^{-1} C_t \xrightarrow{P} c.
\]  

(5.22)

Furthermore, they derive the corollary

\[
C_t^{1/2} (\hat{\beta}_t - \beta_0) \xrightarrow{D} N(0, c\Sigma_0^{-1}),
\]

when $n \to \infty$.

To verify the above mentioned theory, we conducted a simulation study. We performed the simulation study for two reasons, namely (1) to check whether the estimated parameters look approximately normally distributed in finite sample and (2) to check the appropriateness of the information matrix as an estimate for the real estimated parame-
ter standard errors. This is done in Section 5.3.1 and Section 5.3.2 respectively. We also spend some time on determining the optimum choice for the normalizing term $C_i(t)$ in Section 5.3.3.

We intend on simulating data that adheres to the four assumptions made by Arjas and Haara (discussed in Section 5.2). Note that these are sufficient assumptions, but are not necessary assumptions. So, even if one of the assumptions is violated, the properties of the parameter estimates may still hold true.

In this simulation we are concerned with the modelling of failures (missed payments), $N_i(t)$. The counting process $N_i(t)$ represents the cumulative number of failures that occurred up to time $t$ for obligor $i$. Note that $N_i(t)$ does not necessarily represent the delinquency of an account at time $t$, since repayments could have been made on any time $v \leq t$. We simulated the following covariates at $t = 1$: (1) the starting age of the bond, $Age_i(1)$; (2) the time-fixed bond amount, $Bond_i$; and (3) the cumulative number of failures up to the previous time, $N_i(0)$. These covariates were all simulated from the empirical distribution function obtained from the bank data. We allowed delayed entries to occur. We also generated a random early-closure time, $C_i$, by using empirical distribution functions from our data. On each time step we incremented the age covariate, i.e. $Age_i(t + 1) = Age_i(t) + 1$ and set the at-risk indicator to $Y_i(t) = 1 \{Age_i(t) > 0, t \leq C_i\}$. Hence, we prevent the occurrence of a dependent censoring mechanism. We also used the prime interest rate with a lag of three months, $Prime(t - 3)$, as a covariate. The number of obligors was $n$ and we simulated the failures, $\Delta N_i(t) = 1$, over time $t = 1, 2, \ldots, T$. In the AH context the arbitrarily chosen covariate vector is $U_i(t) = [1, N_i(t - 1), Prime(t - 3), Bond_i, Age_i(t)]$ for obligors $i = 1, 2, \ldots, n$, on times $t = 1, 2, \ldots, T$. We chose the covariate $N_i(t - 1)$ as a covariate, since it seemed to be of practical significance to this model. Later we show how choosing this covariate does not violate the first assumption made by AH. The true parameter, $\beta_0$, was chosen to be $\beta_0 = [-3.5, 1.2, 0.1, 0.04, -0.3]$. The parameter values were chosen to closely reflect the estimated parameters from our data. We calculate the failure probabilities with the model

$$P[\Delta N_i(t) = 1|U_i(t - 1)] = Y_i(t - 1) \left(1 + \exp(-\beta_0^T U_i(t - 1))\right)^{-1}.$$  

These probabilities are used to simulate the failure events. We chose any $n$, set $T = 48$.
and did 400 repetitions. We then estimated the parameters of the model on the simulated data by using the methodology described by AH. The pseudo code of this simulation with \( j = 400 \) repetitions is:

1. Set time \( t = 1 \), generate \( N_i(t - 1), Bond_i, C_i \) and \( Age_i(t) \) from the empirical marginal distributions of the data for all \( i = 1, \ldots, n \).

2. Generate a uniform random number \( X_i(t) \), get \( U_i(t) \) (with known \( \text{Prime}(t - 3) \)) and \( Y_i(t) = 1 \{ Age_i(t) > 0, t \leq C_i \} \) for all \( i = 1, \ldots, n \).

3. Set \( \Delta N_i(t) = 1 \left\{ X_i < Y_i(t - 1) \left( 1 + \exp \left( -\beta_0^T U_i(t - 1) \right) \right)^{-1} \right\} \) for all \( i = 1, \ldots, n \).

4. While \( t \leq 47 \), then set \( Age_i(t + 1) = Age_i(t) + 1, N_i(t) = N_i(t - 1) + \Delta N_i(t) \) for all \( i = 1, \ldots, n \), set \( t = t + 1 \) and repeat steps 2 to 4, otherwise go to step 5.

5. Estimate the parameters, \( \hat{\beta}^{(j)} \).

Repeat this simulation \( j \) times.

The simulated data adheres to the four assumptions of the AH approach. To verify this, we check each assumption individually. Firstly, the parameter of interest is the parameter coefficient \( \beta \), which governs whether a failure has occurred. \( \beta \) does not depend on the nuisance parameter, \( \beta^* \), which governs the distribution of the covariates and the at-risk set. For obligor \( i \) we have the covariate matrix \( U_i(t) = [1, N_i(t - 1), R(t), Bond_i, Age_i(t)] \). The covariates are driven by parameters \( \beta^* \). It follows for some observed \( k \) that

\[
P \left( Y_i(t) = 1, U_i(t) = k | F^{(j)}_t \right)
\]

\[
= P \left( Y_i(t) = 1, U_i(t) = k | Y_i(1), ..., Y_i(t - 1), U_i(1), ..., U_i(t), \Delta N_i(1), ..., \Delta N_i(t) \right)
\]

\[
= P \left( Y_i(t) = 1 | F^{(j)}_t \right)
\]

for \( k = 0, 1, 2, \ldots \). So, the first factor of (5.8) only depends on \( \beta^* \).
Also,

\[
P \left( \Delta N_i(t) = \Delta n_i(t) \mid G_{i-1}^{(i)} \right) = P \left( \Delta N_i(t) = \Delta n_i(t) \mid F_{i-1}^{(i)}, Y_i(t-1), U_i(t-1) \right)
\]

\[
= P (\Delta N_i(t) = \Delta n_i(t) \mid Y_i(1), ..., Y_i(t-1), U_i(1), ..., U_i(t-1), \Delta N_i(1), ..., \Delta N_i(t-1))
\]

implying that the second factor in (5.8) depends only on \( \beta \).

Secondly, the assumption that the obligors \( i = 1, ..., n \) are conditionally independent is satisfied because no dependence structure was postulated during the simulation. Thirdly, when we simulated the response variables we only used the current covariate variables and the current at-risk status. Hence, assumption three is satisfied. Lastly, we assumed the logit structural form for failures.

### 5.3.1 Normality of the estimated parameters

Using \( n = 100 \) we obtain Figure 5-1 for \( \hat{\beta}^{(j)} \). We depict the distribution of the parameter estimates of all five parameters using a kernel estimator on the left-hand side and a corresponding \( qq \)-plot on the right-hand side. The kernel density nearest-neighbour parameter was determined visually (Cook & Weisberg 1997). The distribution of the estimated parameters, \( \hat{\beta}^{(j)} \), seems symmetrically distributed around the true value \( \beta^{(j)} \). This can be seen on the left-hand side, where we plotted the true parameter values as vertical dotted red lines as well as the mean of the estimated parameters on the same graph as a solid vertical blue line. The dashed vertical blue lines are the mean plus or minus one sample standard error. The \( qq \)-plots all show that the estimated distributions have lighter tails than the normal distribution. The tails seem to be thinner than that of the normal distribution, implying that the estimates are close to the real values. We tested the null hypotheses that the parameter estimates are normally distributed with significance \( \alpha = 0.05 \) using a Kolmogorov-Smirnov test and the Anderson-Darling tests. In no case could we reject the null hypothesis. Thus, it would seem as if the asymptotic results start to hold at a rather small \( n \). Even when we decreased the \( n \) to 20, the
estimated parameters still tested as being normally distributed.

5.3.2 The appropriateness of the information matrix

We tested whether the estimated information matrix is a good estimator for the true standard error of the parameter estimates. In Figure 5-2 we plot the distribution of the Fisher-information-estimated standard errors, \( \hat{\sigma}_\beta \), against the real standard errors, \( \hat{\sigma}_{\beta_0} \), \( \hat{\sigma}_{\beta_n} \), obtained by taking the standard deviation of the 400 estimated parameters \( \beta \). The average Fisher information matrix can be obtained by 
\[
\hat{\sigma}_\beta = \left( \sum_{i=1}^{n} s(\tilde{\beta})^2 / n \right)^{1/2},
\]
where \( s(\tilde{\beta}) \) is the standard error estimate in (5.20) derived from the Fisher information matrix.

From Figure 5-2 we plot the kernel-density function of \( s(\tilde{\beta}_i) \) for all \( i = 1, ..., n \). Also superimposed is the real standard deviation estimate \( \hat{\sigma}_{\beta_0} \). We used \( n = 1000 \). Clearly, the real standard errors \( \hat{\sigma}_{\beta_0} \) (the vertical dashed red lines) are well within the estimated distributions.

5.3.3 Determining the normalizing term \( c_t \)

Assume we have a sequence of random variables \( \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_t, ... \) with equal expectation \( \beta_0 \). We estimate the variance of the random sequence by \( \hat{\sigma}^2_{\hat{\beta}_1}, \hat{\sigma}^2_{\hat{\beta}_2}, ..., \hat{\sigma}^2_{\hat{\beta}_t}, ... \). We know that \( \lim_{t \to \infty} \left( \hat{\sigma}^2_{\hat{\beta}_t} \right) = 0 \) and
\[
\left( \hat{\beta}_t - \beta_0 \right) \overset{D}{\to} N(0, \hat{\sigma}^2_{\hat{\beta}_t}).
\]

We need to determine what the normalizing sequence \( c_t \) is such that
\[
c_t^{1/2} \left( \hat{\beta}_t - \beta_0 \right) \overset{D}{\to} N(0, \Sigma)
\]
for some \( \Sigma \). We estimated the \( \hat{\beta}_i \) for a few points \( i = 10, 20, ..., 100, 200, ..., 1000, 2000, ..., 10000 \), where \( i \) corresponds to the number of obligors \( n \). Each \( \hat{\beta}_i \) had corresponding estimated \( \hat{\sigma}^2_{\hat{\beta}_i} \) value derived from the information matrix (5.19). We assume that \( c_t \) would be some function of either \( n \) or the random total time on the test, \( C_t := \sum_{v \leq t} \sum_i Y_i(t) \).
Figure 5-1: The kernel density estimates were fitted on each parameter estimate and a qq-plot was drawn. The blue vertical lines’ left-hand plots are the estimated mean and the dashed lines are the mean plus or minus one standard error. The dashed red line is the theoretical parameter value.
Figure 5-2: The kernel-density estimates for the estimated standard errors compared to the real standard error depicted by the red dotted line.
So, we assume that either
\[ c_t(1) = n^\alpha \]
or
\[ c_t(2) = (C_t)^\alpha. \]

There exists some relationship between the estimated standard error \( \hat{\sigma}_{\beta_t} \) and either \( n \) or \( C_t \). So, we assumed that

\[ \hat{\sigma}_{\beta_t} = \begin{cases} 
  c_t(1) \Sigma_1 & \text{if } c_t = n^\alpha \\
  c_t(2) \Sigma_2 & \text{if } c_t = C_t^\alpha 
\end{cases}, \]

where \( \Sigma_1 \) is the covariance matrix when the normalizing term \( c_t(1) \) is used and \( \Sigma_2 \) is the covariance matrix when \( c_t(2) \) is used. Taking the logarithms we get

\[ \log(\hat{\sigma}_{\beta_t}) = \begin{cases} 
  \gamma_1 + \alpha_1 \log(n) & \text{if } c_t = n^\alpha \\
  \gamma_2 + \alpha_2 \log(C_t) & \text{if } c_t = C_t^\alpha 
\end{cases}, \]

where \( \gamma_1 = \log(\Sigma_1) \) and \( \gamma_2 = \log(\Sigma_2) \) for \( t = 10, 20, ..., 100, 200, ..., 1000, 2000, ..., 10000 \).

We used a scatter plot to visualize the relationship between \( \log(\hat{\sigma}_{\beta_t}) \) and \( \log(c_t(x)) \), \( x = 1, 2 \). Figure 5-3 represents the relationship between \( \log(\hat{\sigma}_{\beta_t}) \) and \( \log(c_t(x)) \) (for \( x = 1 \) and \( x = 2 \)) for the covariate \( Age_i(t) \). It is clear that there is some type of linear relationship. The estimated parameters, \( \hat{\gamma}_1, \hat{\gamma}_2, \hat{\alpha}_1 \) and \( \hat{\alpha}_2 \), are summarized in Table 5.1 for all covariates. Both transformations \( c_t(1) \) and \( c_t(2) \) have similar slopes but different intercepts for all covariates.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>( \hat{\alpha}_1 )</th>
<th>( \hat{\alpha}_2 )</th>
<th>( \hat{\gamma}_1 )</th>
<th>( \hat{\gamma}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.5169</td>
<td>-0.5169</td>
<td>2.0491</td>
<td>3.9945</td>
</tr>
<tr>
<td>( N_t(t - 1) )</td>
<td>-0.5134</td>
<td>-0.5134</td>
<td>-0.6672</td>
<td>1.265</td>
</tr>
<tr>
<td>( Prime_{33}^t )</td>
<td>-0.5113</td>
<td>-0.5113</td>
<td>-1.5257</td>
<td>0.3986</td>
</tr>
<tr>
<td>( Bond_i )</td>
<td>-0.5181</td>
<td>-0.5181</td>
<td>-0.6461</td>
<td>1.3038</td>
</tr>
<tr>
<td>( Age_i(t) )</td>
<td>-0.5163</td>
<td>-0.5163</td>
<td>-0.6116</td>
<td>1.3316</td>
</tr>
</tbody>
</table>
Figure 5-3: The linear relationship of \( \log(\hat{\sigma}_{\beta_i}) \) against either \( \log(c^{(1)}_i) \) or \( \log(c^{(2)}_i) \) for the covariate Age.
Clearly, the value of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ is close to $-0.5$, implying that the normalizing factor is either $n^{-1/2}$ or $C^{-1/2}_t$. We visually represent the findings in Figure ?? This is in line with the normal rate of convergence of $n^{-1/2}$.

5.4 Appendix

5.4.1 Various hypothesis tests

We perform many hypothesis tests in the following sections, so we elaborate on the various types of tests available. Say we have random variables $X_1, X_2, ..., X_n$, which are independent and identically distributed and let $p(x, \theta)$ be the common probability density dependent on some $q$-dimensional parameter $\theta = (\theta_1, \theta_2, ..., \theta_q)^T$. Assume we have a differentiable likelihood function, $L(\theta)$, log-likelihood $\text{lik}(\theta)$ and information matrix $\mathcal{I}$ with elements

$$\mathcal{I}_{r,s} = E \left[ \frac{\partial \text{lik}(\theta)}{\partial \theta_r} \frac{\partial \text{lik}(\theta)}{\partial \theta_s} \right].$$

To test the null hypothesis that $H_0 : \theta \in \Omega_0$, versus $H_a : \theta \in \Omega_a$ we may use likelihood-ratio test

$$\frac{\text{Lik}(\Omega_0)}{\text{Lik}(\Omega)} = \frac{\max_{\theta \in \Omega_0} (\text{Lik}(\theta))}{\max_{\theta \in \Omega} (\text{Lik}(\theta))},$$

where $\Omega = \Omega_0 \cup \Omega_a$. Under some regularity conditions we find that asymptotically

$$-2 \log \left( \frac{\text{Lik}(\Omega_0)}{\text{Lik}(\Omega)} \right)$$

is $\chi^2$ distributed with degrees of freedom specified by the difference between the number of free parameters specified by the null hypothesis and the number of free parameters under the alternative hypothesis.

If we want to test whether a single parameter estimate is equal to a certain value, say zero, applying the likelihood ratio test implies that we need to calculate the likelihood $\text{Lik}(\hat{\theta}_0)$ over and over again for every parameter of interest. This is viable in this day and age, but it was once problematic without modern computers. As a result of this, two approximations to the likelihood-ratio test were developed by Wald and Rao (Rao 1973 p. 417).
The Wald test is given by

\[ W_0 = D_0^\top I(\hat{\theta}) D_0, \]

where \( D_0 = [\sqrt{n} (\hat{\theta} - \theta_0), \ldots, \sqrt{n} (\hat{\theta} - \theta_q)]^\top \) is the vector of deviances and \( I(\hat{\theta}) \) is the information matrix evaluated at \( \theta = \hat{\theta} \). The test proposed by Rao (also known as the Lagrange-multiplier test in econometrics) is an approximation that is based on the scores

\[ S_0 = V_0^\top I^{-1}(\theta_0) V_0, \]

where \( V_0 = [\phi_1(\theta_0), \ldots, \phi_q(\theta_0)]^\top \), where

\[
\phi_i(\theta) = \frac{1}{\sqrt{n}} \left\{ \frac{1}{p(x_1, \theta)} \frac{\partial p(x_1, \theta)}{\partial \theta_i} + \cdots + \frac{1}{p(x_n, \theta)} \frac{\partial p(x_n, \theta)}{\partial \theta_i} \right\}.
\]

Both these test statistics are also \( \chi^2 \) distributed with the same number of degrees of freedom as the likelihood-ratio test. The basic difference between these tests is geometrically depicted in Figure 5-4 for a one-dimensional problem (\( q = 1 \)). The likelihood-ratio test focusses on the difference between the log-likelihoods of \( H_0 \) and \( H_a \) (the vertical axis). The Wald test looks at the difference in the horizontal axis and Rao’s score uses the scores, or first derivatives, at the point \( \theta_0 \) (Rao 1973 p. 417). Rao (1973 p. 418) shows that when the null hypothesis holds true, the three tests are asymptotically equivalent.
Figure 5-4: Different types of hypothesis tests.
Chapter 6

The models: application with terminal default events

In this chapter, we apply and compare the models discussed in the previous chapter to the data. We assume that defaults are once-off or terminal, and that they occur once an account is delinquent for 90 days or more. An account may not return to the portfolio if it becomes less than 90 days delinquent, so the default is seen as a terminal event. We decided on this definition of default to compare the AH model with the DTYG model. In contrast to the AH model, the DTYG model cannot handle multiple defaults per obligor.

6.1 Fitting the AH model to data

We implement the AH model to predict defaults. Here the response variable is \( \Delta N_i(t) = 1 \) if default occurred at time \( t \) for obligor \( i \) and is \( \Delta N_i(t) = 0 \) otherwise. Only the first occurrence of a default is taken, and the obligor is then removed from the at-risk set. So if \( \Delta N_i(t) = 1 \) then \( Y_i(t + 1) = 0 \). The observed default rate, which is the number of defaults divided by the number of accounts at-risk, is depicted in Figure 6-1. The default rate peaked in June 2009, roughly two years after the American sub-prime lending crisis began to ripple through the global economy. We implement the AH model with three covariates: the current FNB property price index, \( PPI(t) \), the delinquency rate in the previous month, \( D(t - 1) \), and the total number of months an account was delinquent during the previous six months, \( TD(t) \).
We chose these covariates in a logical manner: we ran a stepwise logistic regression technique for a training set of 80% of the data and chose the most important covariates. The covariates needed to be representative of the idiosyncratic part as well as the systematic part. Some covariates need to be idiosyncratic, whether static or dynamic, and others need to be systematic. We decided to only use three covariates, since we wanted to keep the model concise. Limiting the number of covariates also limits multicollinearity, ensuring that the estimated parameter signs are in line with expectations. We then fitted a model using only the three covariates that were identified, which is described in Table 6.1.

Table 6.1: Covariates summary.

<table>
<thead>
<tr>
<th>Step entered</th>
<th>Covariate</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$D(t-1)$</td>
<td>The previous months’ delinquency level.</td>
</tr>
<tr>
<td>2</td>
<td>$TD(t)$</td>
<td>The total number of months an account was delinquent over the previous six months.</td>
</tr>
<tr>
<td>3</td>
<td>$PPI(t)$</td>
<td>The FNB house-price index (year on year).</td>
</tr>
</tbody>
</table>

We now investigate each covariate in turn. We estimate the observed default proba-
bilities for the different levels of the discrete covariate $D(t - 1)$ by dividing the number of observed defaults by the total number of observations. See Table 6.2. In Figure 6-2 we graphically depict the logarithm of the odds against different values of $D(t - 1)$ where

$$odds = \frac{P[\text{Observed probability}]}{1 - P[\text{Observed probability}]}.$$  

Clearly we see a positive linear relationship here that is indicative of a positive parameter sign. So, we expect the parameter estimate of $D(t - 1)$, namely $\hat{\beta}_1$, to be positive. The next covariate, $TD(t)$, has seven distinct values. The log of the odds are summarized in Table 6.3. In Figure 6-3 we plot the log(odds) against $TD(t)$. Clearly, the log(odds)
of $TD(t)$ is not as linear as the log($odds$) of $D(t-1)$, so we attempted various types of transformations. None of the transformations attempted (squared, cubed and log transformations) seemed appropriate. The shape of this curve is not unexpected: it shows that, should an obligor miss at least two payments in the previous six months, the obligor is at a higher risk of default. A piecewise regression analysis with a breakpoint at $TD(t) = 2$ seems most appropriate. Creating a new covariate $ID(t) = 1 \{TD(t) < 2\}$, we may fit the following model:

$$
log(odds) = \beta_0 + \beta_2 TD(t) + \beta_3 ID(t) (TD(t) - 2).
$$

(6.1)

The slope of the curve $TD(t) \geq 2$ is given $\beta_2 + \beta_3$. Figure 6-3 shows the graphical fit; we perform a formal hypothesis test to see whether the slope for $TD(t) \geq 2$ is significantly different from zero later. Should the cut-off value (in this case we assumed it to be two) be unknown or unclear, it can be estimated.

![Figure 6-3: The log(odds) of TD(t).](image)

Looking at the combined effect of the two discrete covariates $D(t-1)$ and $TD(t)$, we depict the logarithm of the odds in Figure 6-4. When the $D(t-1)$ increases,
Table 6.3: The observed defaults of $TD(t)$.

<table>
<thead>
<tr>
<th>$TD(t)$</th>
<th>Number of observations</th>
<th>Defaults</th>
<th>Observed Probabilities</th>
<th>log(odds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/6</td>
<td>2,806,733</td>
<td>459</td>
<td>0.00016</td>
<td>-8.71832</td>
</tr>
<tr>
<td>1/6</td>
<td>416,328</td>
<td>2,693</td>
<td>0.00647</td>
<td>-5.03433</td>
</tr>
<tr>
<td>2/6</td>
<td>231,682</td>
<td>13,060</td>
<td>0.05637</td>
<td>-2.81779</td>
</tr>
<tr>
<td>3/6</td>
<td>149,749</td>
<td>8,806</td>
<td>0.05881</td>
<td>-2.77292</td>
</tr>
<tr>
<td>4/6</td>
<td>100,980</td>
<td>7,248</td>
<td>0.07178</td>
<td>-2.55971</td>
</tr>
<tr>
<td>5/6</td>
<td>66,694</td>
<td>5,592</td>
<td>0.08385</td>
<td>-2.39121</td>
</tr>
<tr>
<td>6/6</td>
<td>66,975</td>
<td>6,591</td>
<td>0.09841</td>
<td>-2.21502</td>
</tr>
</tbody>
</table>

log(odds) increases as well. When the $TD(t)$ increases, the level of $D(t-1)$ determines whether the log(odds) will decrease or increase. If $D(t-1) = 0$, there is a clear increase in the log(odds) as $TD(t)$ increases. Otherwise, if $D(t-1) = 1$ or 2 there seems to be a slight increase or decrease respectively in the log(odds) as $TD(t)$ increases. As a result, we decided to include an interaction term $D(t-1) \times TD(t)$ in the model. The parameter estimate of this interaction term should have a positive parameter estimate that is close to zero. A formal hypothesis to test whether the coefficient is significantly different from zero will be done later.

The last covariate is a systematic or macro-economic covariate. The FNB house-price index is discussed in Section 2.2.3. We plot the default rate against the FNB house-price index in Figure 6-5.

Once house prices fall, an obligor may be in a situation where the outstanding loan amount on the house is larger than the value of the house. In such a case, an obligor might be reluctant to repay his loan timeously. Conversely, should house prices increase, a financially stressed obligor might rather sell the house than default on payments. For this reason, we expect the estimated parameter of this covariate $\beta_5$ to be negative.

We fitted the AH model and the resulting parameter estimates on a training sample of 5,000 accounts (roughly 10%), all of which experienced a failure (one month overdue, but not necessarily a 90-day default). Although the sample size of 10% seems small, we had a total of 209,581 obligor months in the sample. We observe the data sample for 48
Figure 6-4: The log(\textit{odds}) of the interaction term $D(t - 1) \times TD(t)$.

Figure 6-5: The year-on-year FNB house-price index ($PPI(t)$) versus the observed default rate.
months and the average at-risk time for each obligor is 42 months. The model we fit was

\[
\text{logit}(P [\Delta N_i(v) = 1|G_{v-1}])
\]

\[
= \beta_0 + \beta_1 D (t - 1) + \beta_2 TD (t)
\]

\[
+ \beta_3 ID (t) (TD (t) - 2)
\]

\[
+ \beta_4 D (t - 1) TD (t) + \beta_5 PPI (t),
\]

where the slope of the piecewise covariate \( TD (t) \) was given by \( \beta_2 + \beta_3 \). Table 6.4 has the parameter estimates, standard errors and \( p \)-value. The \( p \)-value is that of the Wald test, used to determine whether the parameter estimates are significantly different from zero. Of course, this model is conditional on the fact that the cut point for the piecewise linear regression was correctly chosen to be two. All the signs of the parameter estimates are in line with expectations. To test whether the slope of the piecewise regression covariate \( TD (t) \) in (6.1) is zero, we perform a likelihood-ratio test

\[
H_0 : \beta_2 + \beta_3 = 0
\]

\[
H_a : \beta_2 + \beta_3 \neq 0.
\]

<table>
<thead>
<tr>
<th>Parameter estimate</th>
<th>Covariate</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>Intercept</td>
<td>-6.9741</td>
<td>0.0849</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>( D (t - 1) )</td>
<td>3.6463</td>
<td>0.0780</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>( TD (t) )</td>
<td>1.5351</td>
<td>0.1415</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>( ID (t) (TD (t) - 2) )</td>
<td>-1.4439</td>
<td>0.1798</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>( D (t - 1) TD (t) )</td>
<td>-0.3177</td>
<td>0.0556</td>
<td>0.0009</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>( PPI (t) )</td>
<td>-0.1319</td>
<td>0.0177</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>
The likelihood for the saturated model (6.2) is

$$\text{logit}(P[\Delta N_i(v) = 1|\mathcal{G}_{v-1}])$$

$$= \beta_0 + \beta_1 D(t - 1) + \beta_2 TD(t) + \beta_3 ID(t)(TD(t) - 2)$$

$$+ \beta_4 D(t - 1)TD(t) + \beta_5 PPI(t),$$

while the reduced model is

$$\text{logit}(P[\Delta N_i(v) = 1|\mathcal{G}_{v-1}])$$

$$= \beta_0 + \beta_1 D(t - 1) + \beta_2 (TD(t) - ID(t)(TD(t) - 2))$$

$$+ \beta_4 D(t - 1)TD(t) + \beta_5 PPI(t).$$

The ratio-log-likelihood test gives: $X^2 = -2(-9.969.2 - (-9.968.7)) = 1.0192$. The $p$-value is $P(X^2 > 1.1041) = 0.31271$. Accordingly, we cannot reject the null hypothesis and the slope of the piecewise part is not significantly different from zero. Therefore, we choose the reduced-form model over the saturated model. To simplify notation, let $f(TD(t)) = TD(t) - ID(t)(TD(t) - 2)$. The final model and parameter estimates are summarized in Table 6.5.

Table 6.5: The parameter estimates of the final AH model.

<table>
<thead>
<tr>
<th>Parameter estimate</th>
<th>Covariate</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>Intercept</td>
<td>-6.9347</td>
<td>0.07417</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$D(t - 1)$</td>
<td>3.5882</td>
<td>0.05228</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$f(TD(t))$</td>
<td>1.5588</td>
<td>0.13902</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>$D(t - 1)TD(t)$</td>
<td>-0.2633</td>
<td>0.01464</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>$PPI(t)$</td>
<td>-0.13215</td>
<td>0.01774</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>
Rearranging the linear part of the GLM we get the systematic-idiocyncratic split

\[
\text{logit}(P[\Delta N_i(v) = 1|\mathcal{G}_{v-1}]) = \beta_0 + \beta_5 PPI(t) + \beta_1 D(t-1) + \beta_2 f(TD(t)) + \beta_4 D(t-1) TD(t) .
\]

The systematic part is represented by the constant term as well as the \textit{PPI}(t) covariate, and the other covariates are idiosyncratic terms. From Table 6.5 we see that the null hypothesis \( H_0 : \beta_5 = 0 \) is rejected. Hence, we are able to find some significant relationship between the macro-economic covariate \textit{PPI}(t) and the systematic-credit-risk part. Note that since a default is a terminal event, this model is exactly the same as the model in (5.1) proposed by Shumway (2001).

### 6.2 Fitting the DTYG model to data

Before we fit the DTYG model in (5.4) and (5.5), we fit a special case of the DTYG model where \( c = 1 \) or \( \rho = 0 \), namely

\[
P[\Delta N_i(t) = 1|\mathcal{G}_{t-1}, c = 1]) = Y_i(t-1) \left[ 1 - \{1 + \exp(\alpha_t + \beta^T U_i(t-1))\}^{-1} \right] . \tag{6.3}
\]

The only difference between (6.3) and the Shumway (2001) model is a time-varying constant term, \( \alpha_t \) (see (4.20)). We programmed the IRLS methodology in \textit{Matlab} as described in Kalbfleisch and Prentice (2002 p. 139) to estimate the parameters. We already have one time covariate, \( \alpha_t \), that changes over time and is equal for all obligors. Therefore, a macro-economic covariate that is also the same for each individual but that changes over time may not also be a covariate. As mentioned previously, including macro-economic covariates in the DTYG model leads to ambiguity and a singular design matrix. Fitting a model without the time-varying covariate \textit{PPI}(t), we would be able to use the parameter estimate time series \( \hat{\alpha}_t \) for \( t = 1, 2, \ldots \) as an approximation of the
systematic factors. $\hat{\alpha}_t$ is also known as the baseline hazard rate in the Cox-framework, i.e. a common risk factor affecting all obligors. We fit the model

$$P[\Delta N_i(t) = 1|G_{t-1})]$$

$$= Y_i(t-1) \left[1 - \left\{1 + \exp \left(\alpha_t + \beta^T U_i(t-1)\right)\right\}^{-1}\right],$$

with the same design matrix bar the exclusion of the $PPI(t)$ covariate

$$U_i(t-1) = [D_i(t-1), f(TD_i(t)), D_i(t-1) TD_i(t)], \quad (6.4)$$

In Figure 6-6 we plot the estimated value of $\alpha_t$, $\hat{\alpha}_t$, against the linear loess smoother with bandwidth 0.3. The bandwidth parameter was chosen visually. In Figure 6-7 we draw

Figure 6-6: The time varying constant term $\alpha_t$ of the DTYG model with $c = 1$. The smoothing parameter was 0.4.

the smoothed $\hat{\alpha}_t$ against the $PPI(t)$. We want to compare these to see whether the role of the systematic factor $PPI(t)$ is taken over by $\alpha_t$. It is evident that there is a negative relationship between the two, except in the tail values at the far left-hand side. Tails
are known to be volatile when smoothing, so caution should be exercised when drawing conclusions regarding smoothed tails. The estimated covariates for the model (6.3) with covariates in (6.4) are summarized in Table 6.6.

![Figure 6-7: The year on year FNB house-price index versus the smoothed $\hat{\alpha}_t$.](image)

Table 6.6: The parameter estimates for the DTYG model with $c = 1$.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\hat{\beta}$</th>
<th>Standard Error</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D(t - 1)$ ($\beta_1$)</td>
<td>3.5960</td>
<td>0.0524</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$f(TD(t))$ ($\beta_2$)</td>
<td>1.5608</td>
<td>0.1393</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$D(t - 1)TD(t)$ ($\beta_3$)</td>
<td>-0.2665</td>
<td>0.0147</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

The parameter estimates are very similar to the parameter estimates of the AH model given in Table 6.5. The average $\hat{\alpha}_t$ is at about $-6.9629$ compared to the intercept of the AH, $-6.9347$. The other covariates are also of the same order and sign.

We now fit the DTYG models in (5.4) and (5.5). We repeat the formulae here to
We estimate the parameters for different DTYG models dependent on transformation parameter $c$. Figure 6-8 represents the log-likelihood for (6.6) at different levels of $c$.

We estimated the DTYG model with transformation parameter $c$ as well. In Figure 6-9 (a) we estimate the parameters for different DTYG models dependent on transformation parameter $\rho$. Figure 6-8 represents the log-likelihood for (6.6) at different levels of $\rho$. We estimated the DTYG model with transformation parameter $c$ as well. In Figure 6-9
we superimposed the log-likelihood functions of (6.5) and (6.6) as functions of $c$ or $\rho$ respectively. The maximum-likelihood estimate for $c$ is $\hat{c} = 0$ (the dashed black vertical line) and the maximum-likelihood estimate of $\rho$ is $\hat{\rho} = 4$ (the dashed red vertical line). From the figure it can clearly be seen that the two DTYG transformation models have the same log-likelihood value for either $c = 0$ and $\rho = 1$ or $\rho = 0$ and $c = 1$ (the horizontal black lines). The fact that $\hat{c} = 0$ does not imply that the complementary log-log is the appropriate model to use. As mentioned in Section 5.1, the estimated parameter should be an interior point in the parameter space. If this is not the case (as here), then the other transformation family (in this case the DTYG model with transformation parameter $\rho$) should be used. Therefore, we used the DTYG model that is dependent on $\rho$ in (6.6).

The estimated parameters for the DTYG model (6.6) are summarized in Table 6.7.

To test the hypothesis $H_0 : \rho = 0$ we use the likelihood ratio and find that the test statistic is $X^2 = -2 ((-9919.4) - (-9932)) = 12.591$. The $p$-value is $P (\chi^2_1 > 12.591) = 0.0004$. So, the likelihood ratio test rejects the null hypothesis and we can therefore reject the hypothesis that $\rho$ is equal to zero.

We estimated the parameters in Table 6.7 by using various starting values for the
Table 6.7: Covariates summary for the DTYG model with parameter $\rho$.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\hat{\beta}$</th>
<th>Standard Error</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D(t - 1)$ ($\beta_1$)</td>
<td>3.0187</td>
<td>0.0460</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$f(TD(t))$ ($\beta_2$)</td>
<td>1.4599</td>
<td>0.1219</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$D(t - 1)TD(t)$ ($\beta_3$)</td>
<td>-0.1773</td>
<td>0.0094</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

Newton-Raphson procedure. In all cases the parameter estimates were the same. This hints at the estimation of a global maximum-likelihood instead of a local maximum-likelihood. To confirm this, we performed a parametric bootstrap. The parametric-bootstrap estimates for the standard errors of the parameters were in line with the estimation method proposed by DTYG in (5.7), as well as the standard errors obtained through the Newton-Raphson approach. As a result, we believe that we have obtained a global maximum. Also, the parameter estimates found using the parametric bootstrap seemed to be distributed close to normal.

From now on, we only focus on this particular DTYG model ($\hat{\rho} = 4$) and continue to compare it to the AH model.

6.3 Diagnostics

6.3.1 Prediction validation

We validate the prediction accuracy of the AH model versus the DTYG model. Firstly, we determine the optimal cut point for each model on the training data sample. This can be done using either the accuracy or the expected cost, as discussed in Section 3.4.4. The accuracy for various values of the cut point $c^*$ is depicted in Figure 6-10. This figure is not informative, since it shows that the optimal cut point is 1. So, we should assume all obligor monthly observations are non-defaulters and will obtain an accuracy level of about 99%. This is in line with the observed non-default rate over the whole time span, which was a high $3,794,692/3,839,141 = 98.842\%$. This confirms the remark by Hosmer et al. (2013 p. 171) that: "Classification is sensitive to the relative sizes of the two component groups and always favors classification into the larger group, a fact that is also independent of the fit of the model". To counter this uninformative approach, we follow
Figure 6-10: The accuracy of the AH and DTYG (\(\hat{\rho} = 4\)) models.

Bellotti and Crook (2009) by assuming that there is some cost involved in misclassifying an obligor. We use the same weights as they do, namely 
\[\text{Cost}(\Delta N_i(t) = 1|\Delta N_i(t) = 1) = \text{Cost}(\Delta N_i(t) = 0|\Delta N_i(t) = 0) = 0, \quad \text{Cost}(\Delta N_i(t) = 1|\Delta N_i(t) = 0) = 1 \quad \text{and} \quad \text{Cost}(\Delta N_i(t) = 0|\Delta N_i(t) = 1) = 20.\]

Hence, no cost is incurred by correctly predicting either a default or a non-default, but there is a cost of one if a default was predicted but not observed and a cost of 20 if a non-default was predicted but a default was observed. The cost ratio is 20:1, and the resulting expected-cost graph is depicted in Figure 6-11 for various levels of \(c^*\). For robustness, Bellotti and Crook (2009) also use cost ratios of 15:1 and 25:1. We tested these specific cost ratios and drew the expected-cost functions in Figure 6-12 for the DTYG model. The expected-cost graphs for the AH model looked very similar and were omitted. Comparing Figure 6-11 with Figure 6-12, it is clear the cost ratio does not influence the choice of cut point, \(c^*\). From Figure 6-11 it is clear that the optimal cut point should be between about 0.03 and 0.012, so we still choose \(c^* = 0.05\) as our optimal cut point. Having estimated the optimal cut point from the training data sample we now go about predicting the default forward for one month on the validation data set. We sampled a validation sample of 5,000 accounts over all time.
At time $t = 48$ there were only 2,866 that were still at-risk, for which we are able to predict the probability of default. The contingency matrix for both models turned out to be the same and is given in Table 6.8. The corresponding accuracy and expected cost are 97.6971% and 0.0694 respectively.

Table 6.8: The contingency table for the DTYG $\rho=5$ as well as the AH model.

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y = 1$</td>
<td>$Y = 0$</td>
<td></td>
</tr>
<tr>
<td>$Expected$</td>
<td>1</td>
<td>23</td>
<td>59</td>
</tr>
<tr>
<td>0</td>
<td>7</td>
<td>2,777</td>
<td>2,784</td>
</tr>
<tr>
<td>$Total$</td>
<td>30</td>
<td>2,836</td>
<td>2,866</td>
</tr>
</tbody>
</table>

Hence, the prediction accuracy (measured either through accuracy or expected cost) does not indicate which model is better. As a result, we compare the models’ discriminatory power in the next section.
Figure 6-12: The $E(Cost|c^*)$ for cost ratios 25:1, 20:1 and 15:1 for the DTYG model ($\hat{ρ} = 4$).

6.3.2 Model discrimination

To compare the performance of the two models, we now focus on the ability of the models to discriminate. We used the well-known and widely used Receiver Operating Characteristic (ROC) curve on the validation sample. We also computed the area under the ROC curve ($AUC$). The ROC curves of the AH and DTYG ($\hat{ρ} = 4$) models are shown in Figure 6-13. Figure 6-13 shows that the discrimination of both models is high and implies that the discriminatory power of the models is very similar. The $AUC$ of the validation sample is summarized in (6.9). Again, these values are high, with a value of the AUC higher than 0.9 seen as having "outstanding discrimination" (Hosmer et al. 2013 p. 177). We find that the AUC coefficients are close to one another for both models. We tested the null hypothesis $H_0 : AUC_{AH} = AUC_{DTYG}$ versus $H_0 : AUC_{AH} \neq AUC_{DTYG}$ using the non-parametric test developed by DeLong et al. (1988). The $p$-value for this test was 0.69, implying that we cannot reject the null hypothesis in either case. So, the two models have the same discriminatory power.

In summary, we used both prediction and discrimination power to rank the AH and
Table 6.9: The AUC and Gini’s of the hold-out validation sample.

<table>
<thead>
<tr>
<th>Model (Validation sample)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH</td>
<td>0.9836</td>
</tr>
<tr>
<td>DTYG with $\hat{\rho} = 4$</td>
<td>0.9835</td>
</tr>
</tbody>
</table>

DTYG models. In both cases we found the models performed similarly. Therefore, the use of either the AH or DTYG model comes down to personal preference. We prefer using the AH model because it can be extended to incorporate repeated default events, whereas the DTYG model only models a single terminal default event. The AH model does not have the restrictive assumption of a time-dependent intercept term $\alpha_t$. Performing stress-testing would be easier in the AH model than in the DTYG model. Stress-testing the systematic factor $\alpha_t$ in the DTYG model would certainly lose some intuitive interpretation.

We realized that the probability of default is strongly related to the current number of months delinquent. We therefore use the AH model in the next two chapters to develop a delinquency-rate model for obligors. In Chapter 7 we develop a model that incorporates
repeated failures (or missed payments) with a recovery process to form a novel approach to modelling and predicting delinquencies $D(t)$. We compare this approach with a more traditional transition-matrix approach.

6.4 Appendix

6.4.1 Computation of the partial derivative $\frac{\partial}{\partial \theta} \pi_{i,t_j}$

Here we derive the partial derivative $\frac{\partial}{\partial \theta} \pi_{i,t_j}$. The derivation is not shown in the article by Ding et al. (2012). For the DTYG model with transformation parameter $c$, the probability of an event occurring is

$$
\pi_{i,t_j} = \begin{cases} 
1 - \{1 + c \exp(\alpha_t + \beta^T U_i(t-1))\}^{-1/c} & \text{if } c > 0 \\
1 - \exp\{-\exp(\alpha_t + \beta^T U_i(t-1))\} & \text{if } c = 0.
\end{cases}
$$

Let the parameter set be $\theta = (\beta^T, \alpha_1, \alpha_2, \ldots, \alpha_k, c)$ for $c > 0$, hence

$$
\frac{\partial}{\partial \theta} \pi_{i,t_j} = \left[ \frac{\partial \pi_{i,t_j}}{\partial \beta_1}, \ldots, \frac{\partial \pi_{i,t_j}}{\partial \beta_p}, \frac{\partial \pi_{i,t_j}}{\partial \alpha_1}, \ldots, \frac{\partial \pi_{i,t_j}}{\partial \alpha_k}, \frac{\partial \pi_{i,t_j}}{\partial c} \right]^T.
$$

Setting $x = \alpha_t + \beta^T U_i(t-1)$, the partial derivative of $\pi_{i,t_j}$ with respect to $\beta_l$ is

$$
\frac{\partial}{\partial \beta_l} \pi_{i,t_j} = \frac{U_{i,l}(t-1)\exp(x)}{(1 + c \exp(x))^{1/c+1}}
$$

for $l = 1, 2, \ldots, p$. The partial derivative of $\pi_{i,t_j}$ with respect to $\alpha_j$ is

$$
\frac{\partial}{\partial \alpha_j} \pi_{i,t_j} = \frac{\exp(x)}{(1 + c \exp(x))^{1/c+1}}
$$

for a specific $j$ and for the rest

$$
\frac{\partial}{\partial \alpha_h} \pi_{i,t_j} = 0
$$

for $h \neq j, j, h = 1, 2, \ldots, k$. Lastly, partial derivative of $\pi_{i,t_j}$ with respect to $c$ is

$$
\frac{\partial}{\partial c} \pi_{i,t_j} = (1 + ce^x)^{-1/c} \left( \frac{1}{c} \frac{e^x}{1 + ce^x} - \frac{\log(1 + ce^x)}{c^2} \right).
$$
If the DTYG model with transformation parameter $\rho$ is used, the probability of an event occurring is

$$
\pi_{i,t_j} = \begin{cases} 
1 - \left\{ 1 + \frac{1}{\rho} \left( \exp \left[ \rho \exp \left( \alpha_t + \beta^T U_i (t - 1) \right) \right] - 1 \right) \right\}^{-1} & \text{if } \rho > 0 \\
1 - \left\{ 1 + \exp \left( \alpha_t + \beta^T U_i (t - 1) \right) \right\}^{-1} & \text{if } \rho = 0
\end{cases}
$$

and $\theta = (\beta^T, \alpha_1, \alpha_2, ..., \alpha_k, \rho)$ we have for $\rho > 0$ that

$$
\frac{\partial}{\partial \theta} [\pi_{i,t_j}] = \left[ \frac{\partial \pi_{i,t_j}}{\partial \beta_1}, ..., \frac{\partial \pi_{i,t_j}}{\partial \beta_p}, \frac{\partial \pi_{i,t_j}}{\partial \alpha_1}, ..., \frac{\partial \pi_{i,t_j}}{\partial \alpha_k}, \frac{\partial \pi_{i,t_j}}{\partial \rho} \right]^T.
$$

Now, the partial derivative of $\pi_{i,t_j}$ with respect to $\beta_l$ is

$$
\frac{\partial}{\partial \beta_l} \pi_{i,t_j} = \frac{U_{i,l} (t - 1) e^x \exp (\rho e^x)}{[(\exp(\rho e^x) - 1)/\rho + 1]^2}
$$

for $l = 1, 2, ..., p$. The partial derivative of $\pi_{i,t_j}$ with respect to a specific $\alpha_j$ is

$$
\frac{\partial}{\partial \alpha_j} \pi_{i,t_j} = \frac{e^x \exp (\rho e^x)}{[(\exp(\rho e^x) - 1)/\rho + 1]^2}
$$

with the other time instances having partial derivative

$$
\frac{\partial}{\partial \alpha_h} \pi_{i,t_j} = 0
$$

for $h \neq j, h = 1, 2, ..., k$. Lastly we may derive the partial derivative of $\pi_{i,t_j}$ with respect to $\rho$:

$$
\frac{\partial}{\partial \rho} \pi_{i,t_j} = -\frac{[\exp(\rho e^x) - 1]/\rho^2 - (e^x \exp (\rho e^x))/\rho]}{[(\exp(\rho e^x) - 1)/\rho + 1]^2}.
$$

### 6.4.2 Maximizing the likelihood in the DTYG model

We want to find $c, \beta, \alpha$ such that

$$
\max_{c,\beta,\alpha} \{ \text{lik} (c, \beta, \alpha) \} = \max_c \left\{ \max_{\beta,\alpha} \{ \text{lik} (c^*, \beta, \alpha) \} \right\}.
$$
To solve \( \max_{\beta, \alpha} \text{lik}(c^*, \beta, \alpha) \), we follow the approach by Kalbfleisch and Prentice (2002 p. 139). Recall that the probability of a default occurring (for the DTYG model dependent on \( c \)) is

\[
\pi_{i,t_j} = \begin{cases} 
1 - \left\{ 1 + c \exp \left( \alpha_{t_j} + \beta^T U_i (t_j - 1) \right) \right\}^{-1/c} & \text{if } c > 0 \\
1 - \exp \left\{ - \exp \left( \alpha_{t_j} + \beta^T U_i (t_j - 1) \right) \right\} & \text{if } c = 0.
\end{cases} 
\] (6.7)

For now we focus our attention on (6.7) where \( c > 0 \). Suppose that at time \( t \) for obligor \( i \) the default probability is given by

\[
g(\alpha_{t_j} + \beta^T U_i (t_j - 1)) = 1 - \left\{ 1 + c \exp \left( \alpha_{t_j} + \beta^T U_i (t_j - 1) \right) \right\}^{-1/c}. \] (6.8)

Let \( x_{t_j,i} := \alpha_{t_j} + \beta^T U_i (t_j - 1) \) and \( y_{t_j,i} := D_{i,t_j} \). The log-likelihood can be written as

\[
\text{lik}(\beta, \alpha) = \sum_{i=1}^{n} \sum_{j: O_i < t_j \leq \tilde{T}_i} y_{t_j,i} \log \left( g(x_{t_j,i}) \right) + (1 - y_{t_j,i}) \log \left( 1 - g(x_{t_j,i}) \right), \] (6.9)

where \( \alpha = [\alpha_{t_1}, \ldots, \alpha_{t_k}]^T \), \( O_i \) is the opening date of the \( i^{th} \) obligor and \( \tilde{T}_i \) is the observed exit time, \( \tilde{T}_i = \min(T_i, C_i) \), \( T_i \) is the default time and \( C_i \) is the censoring time.

The first-order partial derivative of the log-likelihood with respect to \( x_{t_j,i} \) of (6.9) is

\[
\frac{\partial \text{lik}}{\partial x_{t_j,i}} = \sum_{i=1}^{n} \sum_{j: O_i < t_j \leq \tilde{T}_i} \left[ \frac{(e^{x_{t_j,i}}(y_{t_j,i} - 1)(c e^{x_{t_j,i}} + 1)^{(1/c^*)}}{(c e^{x_{t_j,i}} + 1)^{(1/c^*) + 1)}) - \frac{(y_{t_j,i} e^{x_{t_j,i}})}{(1/c^* + 1)(c e^{x_{t_j,i}} + 1)^{(1/c^*) - 1}(c e^{x_{t_j,i}} + 1)^{(1/c^*) + 1})} \right]
\]

\[
: = \sum_{i=1}^{n} \sum_{j: O_i < t_j \leq \tilde{T}_i} \gamma_{t_j,i}^c.
\]

So, the \( m^{th} \) element of the vector \( \partial \text{lik} / \partial \alpha \) is

\[
\frac{\partial \text{lik}}{\partial \alpha_{t_m}} = \sum_{i=1}^{n} \sum_{j: O_i < t_j \leq \tilde{T}_i} 1 \{ t_j = t_m \} \gamma_{t_j,i}^c
\]
for \( m = 1, \ldots, k \) since the partial derivative for \( j \neq m \) is
\[
\sum_{i=1}^{n} \sum_{j: O_i < t_j \leq \hat{T}_i, j \neq m} y_{t_j,i} \log(g(x_{t_j,i})) + (1 - y_{t_j,i}) \log(1 - g(x_{t_j,i})) = 0.
\]

The \( h \)th element of \( \partial \text{lik} / \partial \beta \) is
\[
\frac{\partial \text{lik}}{\partial \beta_h} = \sum_{i=1}^{n} \sum_{j: O_i < t_j \leq \hat{T}_i} U_{i,h}(t_j - 1) \gamma_{t_j,i}^c
\]
for \( h = 1, \ldots, p \), where \( U_{i,h}(s) \) is the \( h \)th covariate of obligor \( i \) at time \( s \).

We now need to set the score vectors equal to zero and solve the equations simultaneously. So, we use the Newton-Raphson procedure. We need the Fisher information matrix
\[
H = \begin{pmatrix}
H_{11} & H_{12} \\
H_{21} & H_{22}
\end{pmatrix}
:= \begin{pmatrix}
-\frac{\partial^2 \text{lik}}{\partial \alpha \partial \alpha} & -\frac{\partial^2 \text{lik}}{\partial \alpha \partial \beta} \\
-\frac{\partial^2 \text{lik}}{\partial \beta \partial \alpha} & -\frac{\partial^2 \text{lik}}{\partial \beta \partial \beta}
\end{pmatrix}.
\tag{6.10}
\]

To derive the second derivatives in (6.10) we need the following identity
\[
\frac{\partial^2 \text{lik}}{\partial x_{t_j,i}^2} = \sum_{i=1}^{n} \sum_{j: O_i < t_j \leq \hat{T}_i} \left[ \frac{e^{x_{t_j,i}} (y_{t_j,i} - 1)(c^* e^{x_{t_j,i}} + 1)^{(1/c^*)} - y_{t_j,i} e^{x_{t_j,i}}}{(c^* e^{x_{t_j,i}} + 1)^{(1/c^*) - 1} (c^* e^{x_{t_j,i}} + 1)^{(2/c^*) + 1}} \right.
\]
\[
- \frac{e^{2x} (y_{t_j,i} - 1)(c^* e^{x_{t_j,i}} + 1)^{(1/c^*) - 1}}{(c^* e^{x_{t_j,i}} + 1)^{(1/c^*) + 1}}
\]
\[
+ \frac{y_{t_j,i} e^{x_{t_j,i}}}{(1/(c^* e^{x_{t_j,i}} + 1)^{(1/c^*)}) - 1) (c^* e^{x_{t_j,i}} + 1)^{(1/c^*) + 1}}
\]
\[
+ \frac{c^* y_{t_j,i} e^{x_{t_j,i}} (1/c^* + 1}{(1/(c^* e^{x_{t_j,i}} + 1)^{(1/c^*)}) - 1) (c^* e^{x_{t_j,i}} + 1)^{(1/c^*) + 2}}
\]
\[
- \frac{c^* e^{x_{t_j,i}} (1/c^* + 1) (y_{t_j,i} - 1)(c^* e^{x_{t_j,i}} + 1)^{(1/c^*)}}{(c^* e^{x_{t_j,i}} + 1)^{(1/c^*) + 2}} \right]
\]
\[
: = \sum_{i=1}^{n} \sum_{j: O_i < t_j \leq \hat{T}_i} \xi_{t_j,i}^c.
\]
\( \partial^2 \text{lik} / \partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha} \) is a diagonal matrix with diagonal elements

\[
\frac{\partial^2 \text{lik}}{\partial \alpha_{tm} \partial \alpha_{tm}} = \sum_{i=1}^{n} \sum_{j: \xi_i < t_j < \tilde{T}_i} 1 \{ t_j = t_m \} \xi_{t_j,i}^c,
\]

Element \((h^*, h)\) of \( \partial^2 \text{lik} / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta} \) is given by

\[
\frac{\partial^2 \text{lik}}{\partial \beta_{hm} \partial \beta_{hm}} = \sum_{i=1}^{n} \sum_{j: \xi_i < t_j < \tilde{T}_i} U_{i,h^*(t_j - 1)} U_{i,h}(t_j - 1) \xi_{t_j,i}^c
\]

for \( h^*, h = 1, ..., p \). The \((m, h)\)th element of \( \partial^2 \text{lik} / \partial \boldsymbol{\alpha} \partial \boldsymbol{\beta} \) is

\[
\frac{\partial^2 \text{lik}}{\partial \alpha_{tm} \partial \beta_{hm}} = \sum_{i=1}^{n} \sum_{j: \xi_i < t_j < \tilde{T}_i} 1 \{ t_j = t_m \} U_{i,h}(t_j - 1) \xi_{t_j,i}^c
\]

for \( h = 1, ..., p \) and \( m = 1, ..., k \).

The Newton-Raphson iteration to compute \((\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})^\top\) involves updating the current values \((\hat{\alpha}^v, \hat{\beta}^v)^\top\) to \((\hat{\alpha}^{v+1}, \hat{\beta}^{v+1})^\top\) by using the following equation:

\[
\begin{pmatrix}
\hat{\alpha}^{v+1} \\
\hat{\beta}^{v+1}
\end{pmatrix} = \begin{pmatrix}
\hat{\alpha}^v \\
\hat{\beta}^v
\end{pmatrix} + (\mathbf{H}^v)^{-1} \begin{pmatrix}
\partial \text{lik} \\
\partial \text{lik}
\end{pmatrix},
\tag{6.11}
\]

where \( \mathbf{H}^v \) is (6.10) evaluated at \( \hat{\alpha}^v \) and \( \hat{\beta}^v \). Typically no guidelines are provided in the literature on the starting values taken by \((\hat{\alpha}^0, \hat{\beta}^0)^\top\), though an arbitrary choice of values, such as all zeros or ones, seems to work. The iteration in (6.11) is performed until the parameter estimates converge. Once we have obtained \( \text{argmax}_{\beta, \alpha} \text{lik}(c^*, \beta, \alpha) \) for a specific \( c^* \) by using this methodology, we may choose a grid of \( c^* \) to find \( \text{max}_{c^*} \{ \text{max}_{\beta, \alpha} \text{lik}(c^*, \beta, \alpha) \} \).

The same methodology of estimating \((c^*, \beta, \alpha)\) may also be applied to (6.7) when \( c = 0 \). Set

\[
g(x_{tj,i}) = 1 - \exp \left\{ - \exp \left( x_{tj,i} \right) \right\}.
\]

The rest of the procedure follows in a similar fashion, bar the different \( \gamma_{tj,i}^c \) and \( \xi_{tj,i}^c \) values:

\[
\gamma_{tj,i}^c = \frac{e^{x_{tj,i}}(y_{tj,i} - 1) - y_{tj,i}e^{x_{tj,i}}}{\exp(e^{x_{tj,i}})(1/\exp(e^{x_{tj,i}}) - 1)},
\]

150
\[
\xi_{t_j,i}^c = e^{x_{t_j,i}}(y_{t_j,i} - 1) + \frac{y_{t_j,i}e^{2x_{t_j,i}}}{\exp(e^{x_{t_j,i}})(1/\exp(e^{x_{t_j,i}}) - 1)}
\]
\[
- \frac{y_{t_j,i}e^{2x_{t_j,i}}}{\exp(2e^{x_{t_j,i}})(1/\exp(e^{x_{t_j,i}}) - 1)^2} \frac{(y_{t_j,i}e^{x_{t_j,i}})}{\exp(e^{x_{t_j,i}})(1/\exp(e^{x_{t_j,i}}) - 1)}.
\]

For the other transformation where the DTYG model is a function of \( \rho \), we also follow the same basic procedure. Here we set

\[
g(\alpha_{t_j} + \beta^T U_i (t_j - 1)) = \begin{cases} 
1 - \left\{ 1 + 1/\rho^* \left( \exp \left[ \rho^* \exp \left( \alpha_{t_j} + \beta^T U_i (t_j - 1) \right) \right] - 1 \right) \right\}^{-1} & \text{if } \rho^* > 0 \\
1 - \left\{ 1 + \exp \left( \alpha_{t_j} + \beta^T U_i (t_j - 1) \right) \right\}^{-1} & \text{if } \rho^* = 0.
\end{cases}
\]

We need to estimate \( \rho, \beta \) and \( \alpha \) such that

\[
\max_{\rho,\beta,\alpha} \{ \text{lik} (\rho, \beta, \alpha) \} = \max_{\rho^*} \left\{ \max_{\beta,\alpha} \{ \text{lik} (\rho^*, \beta, \alpha) \} \right\}.
\]

Similar arguments lead to the same Newton-Raphson procedure. The only difference is that there is now another new \( \gamma_{t_j,i}^c \) (call this \( \gamma_{t_j,i}^c \)) and another \( \xi_{t_j,i}^c \) (call this \( \xi_{t_j,i}^c \)). These are given by

\[
\gamma_{t_j,i}^c = \begin{cases} 
\frac{e^{x_{t_j,i}} \exp(\rho^* e^{x_{t_j,i}})(y_{t_j,i} - 1)}{(\exp(\rho^* e^{x_{t_j,i}}) - 1)/\rho^* + 1} - \frac{y_{t_j,i}e^{x_{t_j,i}} \exp(\rho^* e^{x_{t_j,i}})}{(\exp(\rho^* e^{x_{t_j,i}}) - 1)/\rho^* + 1)^2(1/(\exp(\rho^* e^{x_{t_j,i}}) - 1)/\rho^* + 1) - 1) & \text{if } \rho^* > 0 \\
\frac{e^{x_{t_j,i}}(y_{t_j,i} - 1)}{e^{x_{t_j,i}} + 1} - \frac{y_{t_j,i}e^{x_{t_j,i}}}{(e^{x_{t_j,i}} + 1)^2(1/(e^{x_{t_j,i}} + 1) - 1)} & \text{if } \rho^* = 0
\end{cases}
\]
and

\[ \xi_{t,j,i}^\rho = \frac{e^{x_{t,j,i}} \exp(\rho^* e^{x_{t,j,i}}) (y_{t,j,i} - 1)}{(\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1} - \frac{e^{2x_{t,j,i}} \exp(2\rho^* e^{x_{t,j,i}}) (y_{t,j,i} - 1)}{((\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1)^2} \]

\[ - \frac{y_{t,j,i} e^{x_{t,j,i}} \exp(\rho^* e^{x_{t,j,i}})}{((\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1)^2(1/((\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1) - 1)} \]

\[ + \frac{\rho^* e^{2x_{t,j,i}} \exp(\rho^* e^{x_{t,j,i}}) (y_{t,j,i} - 1)}{(\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1} \]

\[ + \frac{2y_{t,j,i} e^{2x_{t,j,i}} \exp(2\rho^* e^{x_{t,j,i}})}{((\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1)^3(1/((\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1) - 1)} \]

\[ - \frac{y_{t,j,i} e^{2x_{t,j,i}} \exp(2\rho^* e^{x_{t,j,i}})}{((\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1)^4(1/((\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1) - 1)} \]

\[ - \frac{\rho^* y_{t,j,i} e^{2x_{t,j,i}} \exp(\rho^* e^{x_{t,j,i}})}{((\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1)^2(1/((\exp(\rho^* e^{x_{t,j,i}}) - 1)/\rho^* + 1) - 1)} \]

if \( \rho^* > 0 \) and

\[ \xi_{t,j,i}^\rho = \frac{e^{x_{t,j,i}} (y_{t,j,i} - 1)}{e^{x_{t,j,i}} + 1} \frac{e^{2x_{t,j,i}} (y_{t,j,i} - 1)}{(e^{x_{t,j,i}} + 1)^2} \]

\[ + \frac{2y_{t,j,i} e^{2x_{t,j,i}}}{(e^{x_{t,j,i}} + 1)^3(1/(e^{x_{t,j,i}} + 1) - 1)} \]

\[ - \frac{y_{t,j,i} e^{2x_{t,j,i}}}{(e^{x_{t,j,i}} + 1)^4(1/(e^{x_{t,j,i}} + 1) - 1)^2} \]

\[ - \frac{y_{t,j,i} e^{x_{t,j,i}}}{(e^{x_{t,j,i}} + 1)^2(1/(e^{x_{t,j,i}} + 1) - 1)} \]

if \( \rho^* = 0 \).
Chapter 7

A failure model: the AH model applied to our data

In this chapter we fit the AH model to our data to predict failures (increases in the delinquency level) by obligors. We assume that there exists a point process \( N_i(t) \) for each obligor \( i \) at times \( t = 1, 2, ..., T \) that represents the cumulative number of failures experienced by obligor \( i \) up to time \( t \). Note the change of notation: previously we let \( \Delta N_i(t) = 1 \) denote the default of obligor \( i \). Since no confusion can occur, we now let \( \Delta N_i(t) = 1 \) denote the failure of obligor \( i \) at time \( t \). \( N_i(t) \) is the cumulative number of failures that obligor \( i \) experienced up to time \( t \). \( N_i(t) \) does not necessarily represent the delinquency state at time \( t \), denoted by \( D(t) \), as repayments may have occurred at any time \( v < t \). To estimate the parameters \( \beta \) we maximize the partial likelihood in (5.16).

7.1 The first-order model

We applied the AH model by using a forward stepwise regression to identify the most significant covariates from all the covariates described in Section 2.2. The resulting first-order covariates, \( U_i(t) \), are summarized in Table 7.1. Initially, we only consider first-order covariates; second-order and interaction covariates will be added later. This model will henceforth be referred to as the initial model. After maximizing the likelihood in (5.16) and calculating the standard error by using the information matrix in (5.19), the resulting parameter estimates, \( \hat{\beta} \), and corresponding standard errors, \( SE(\hat{\beta}) \), are summarized in
Table 7.1: The covariates of the failure model.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TD(t)$</td>
<td>The total number of months an account was in arrears in the previous six months.</td>
</tr>
<tr>
<td>$D(t-1)$</td>
<td>The delinquency level at time $t$.</td>
</tr>
<tr>
<td>$Age(t)$</td>
<td>The current age of the loan since origination.</td>
</tr>
<tr>
<td>$TP(t)$</td>
<td>The total number of additional repayments made during the previous six months.</td>
</tr>
<tr>
<td>$TR(t)$</td>
<td>The time since the last additional repayment was made.</td>
</tr>
<tr>
<td>$N(t-1)$</td>
<td>The cumulative number missed payments over the life-span of the obligor on the previous month.</td>
</tr>
<tr>
<td>$Prime(t-3)$</td>
<td>The prime interest rate with a lag of three months.</td>
</tr>
<tr>
<td>$Bond$</td>
<td>The initial bond price.</td>
</tr>
</tbody>
</table>

Table 7.2: We now check the model’s adequacy and perform some diagnostic tests.

Table 7.2: The parameter estimates for the failure model.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\hat{\beta}$</th>
<th>SE($\hat{\beta}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$-3.06$</td>
<td>$0.115$</td>
</tr>
<tr>
<td>$TD(t)$</td>
<td>$0.266$</td>
<td>$0.009$</td>
</tr>
<tr>
<td>$D(t-1)$</td>
<td>$0.222$</td>
<td>$0.007$</td>
</tr>
<tr>
<td>$Age(t)$</td>
<td>$-0.244$</td>
<td>$0.008$</td>
</tr>
<tr>
<td>$TP(t)$</td>
<td>$0.325$</td>
<td>$0.009$</td>
</tr>
<tr>
<td>$TR(t)$</td>
<td>$0.346$</td>
<td>$0.011$</td>
</tr>
<tr>
<td>$N(t-1)$</td>
<td>$0.477$</td>
<td>$0.015$</td>
</tr>
<tr>
<td>$Prime(t-3)$</td>
<td>$0.037$</td>
<td>$0.003$</td>
</tr>
<tr>
<td>$Bond$</td>
<td>$0.051$</td>
<td>$0.008$</td>
</tr>
</tbody>
</table>

### 7.2 Diagnostic tests on the initial model

The small number of literature studies available on diagnostic tests attests to how low the priority of these tests is. Some work has been done to test the proportionality assumption in the Cox proportional hazard model (Schönfeld 1982 and Grambsch & Therneau 1994). These authors mostly concern themselves with whether the proportional hazard model’s assumption is applicable and whether there is a need to transform covariates. This is done by determining the behaviour of the martingale residuals over time. The Cox
model is mostly applied in the medical field, where data is sparse. We find that there is abundant data in the analysis of credit risk, making the analysis of martingale residuals less necessary and valuable than in the medical field. Instead, we make use of a training sample of the data (usually about 80%) to estimate the model parameters. We then test the prediction accuracy by applying the estimated model on a validation data set (the other 20%). This intuitive and well-known methodology can be used to compare the prediction accuracy of several models, from the minimal model to more complicated postulated models.

In addition to determining the prediction accuracy, we use marginal-model plots to perform diagnostic tests. We discussed and extended the methodology underlying these plots in Section 3.4.3. Essentially, marginal-model plots will assist in determining the right functional-model form for the model covariates. They can also be used as a heuristic goodness-of-fit measure.

We now go about plotting marginal-model plots for the initial model. For a specific time $t$ the observed response is $\Delta N_i(t)$ and the fitted response is

$$\Delta \hat{N}_i(t) = E \left[ \Delta N_i(t) | G_{i-1} \right] = Y_i(t-1) \left( 1 + \exp \left( -\hat{\beta}^T U_i(t-1) \right) \right)^{-1}.$$ 

Recall that in marginal-model plots various values of the direction vector $\mathbf{a}$ need to be chosen to obtain the linear combination of predictors. We write these as $U(t)\mathbf{a}$ (see Section 3.4.3). The following potential $\mathbf{a}$-vectors were chosen: (1) the covariates separately, (2) the parameter estimate $\hat{\beta}$ and (3) a randomly chosen $\mathbf{a}$.

To simplify notation, the dates January 2007 to December 2010 for which data was obtained will now be numbered month 1 to month 48. By nature, a marginal-model plot is a cross-classification diagnostic tool, and hence we cannot apply the marginal-model plots to this longitudinal data problem. However, if we choose a specific time and perform the diagnostic test at this time, the marginal-model plots can be used. We chose the end time $t = 48$ to test the adequacy of the initial model. The marginal-model plots with 100(1 - $\alpha$)% confidence bands (we suggest using the conventional $\alpha = 0.05$) are plotted in Figure 7-1. We used 1000 bootstrap samples to estimate the confidence bands. Marginal-model plots are dependent on smoothers. We used lowess smoothers and followed Cook and Weisberg (1997) by determining the nearest-neighbour fraction visually. Choices of $\mathbf{a}$
were \( a = \hat{\beta} \) and \( a \) equal to all zeros except the covariant of interest. The nearest-neighbour fraction, also referred to as the smoothing parameter, was 0.3 for \( D(t-1) \), \( TP(t) \) and \( TD(t) \); 0.5 for \( N(t-1) \) and \( TP(t) \); 0.6 for \( U(t)\hat{\beta} \); and 0.7 for \( Age(t) \) and \( Bond \). From Figure 7-1 it is clear that the covariates \( D(t-1) \) and \( TD(t) \) need second-order terms added. The covariates \( Age(t) \) and \( Bond \) seem to have the right sign for the slope term, but may also need a second-order term included. The covariates \( N(t-1) \) and \( TR(t) \) seem to have a large deviation from the fit, suggesting that a piecewise linear model may be appropriate for these two covariates. Summarizing all these individual covariate effects can be done by using the linear combination \( U(t)^{\hat{\beta}} \). This graph is the top-left graph in Figure 7-1. Ideally, this linear combination (the estimated model represented by the red dotted line) should exhibit closeness to the free model (the black dashed line) on the marginal-model plot. Here too the slope sign of the plot seems to be correct, but the curvature is wrong.

A problem associated with the use of marginal-model plots in this context is that the evolution of covariates over time is not captured by a marginal-model plot. As mentioned before, a marginal-model plot is essentially a cross-sectional data-analysis methodology and is not applicable to longitudinal data. To bring these deficiencies to the fore, we drew the same pairs of marginal-model plots for an arbitrarily chosen \( t = 24 \) in Figure 7-2. We used the same smoothing parameters as the ones used in Figure 7-1. Now, the marginal-model plot for \( N(t-1) \) suggests a linear relationship between \( \Delta N(t) \) and \( N(t-1) \) and not a piecewise linear relationship as at \( t = 48 \). This implies that at some time \( 0 \leq t^* \leq 48 \) the relationship between \( \Delta N(t) \) and \( N(t-1) \) has changed from a linear relationship to a piecewise linear relationship.

Obtaining conflicting results from marginal-model plots constructed at different times is problematic. Hence, we need a methodology to check the adequacy of the results over all time \( t = 1, \ldots, 48 \).

We developed a three-dimensional marginal-model plot, henceforth referred to as a composite marginal-model (CMM) plot. The first two dimensions are still the linear combination \( U(t)a \) and \( \Delta N(t) \). We add a third dimension: time. The CMM plot is constructed by using a quadratic three-dimensional loess smoother to estimate the smoothed surface for all obligors over all time for all values \( U(t)a \). The smoothing
Figure 7-1: Marginal-model plots with different choices of $\alpha$ at time $t = 48$. 

157
Figure 7-2: Marginal-model plots with different choices of $a$ at time $t = 24$. 

158
parameter used is typically small (about 0.1 or 0.15), and again we determine it visually. We standardize the data before applying the loess smoother, i.e. we ensure that neither of the two original axes \((t \text{ or } U(t)a)\) dominates the other. Omitting this step would have resulted in a surface plot that is not smooth. The response variable is not standardized. Again we should construct a CMM for any linear combination \(U(t)a\) and the true non-parametric estimated surface, \(\tilde{E}_F [\Delta N(t) \mid U(t)a]\), should exhibit closeness to the fitted model surface, \(\tilde{E}_M [\Delta N(t) \mid U(t)a]\). To keep the analysis concise, we decided to only focus on the linear combination \(U(t)\tilde{\beta}\), omitting the scatter plot of observed response variables. To make the CMM more interpretable, we plot two pairs of surfaces. We plot the estimated non-parametric true-model surface \(\tilde{E}_F [\Delta N(t) \mid U(t)\tilde{\beta}]\) in Figure 7-3 and the fitted-model surface \(\tilde{E}_M [\Delta N(t) \mid U(t)\tilde{\beta}]\) adjacent to each other in Figure 7-4. The red surface on the left-hand side of Figure 7-3 is \(\tilde{E}_F [\Delta N(t) \mid U(t)\tilde{\beta}]\) and the blue surface on the right-hand side is \(\tilde{E}_M [\Delta N(t) \mid U(t)\tilde{\beta}]\). In Figure 7-4 we superimpose the two CMM plots and display these at different angles. Here it is clear that the fitted model is somewhat different from the non-parametric model. We will determine the closeness of these two surfaces later.

The CMM approach is better suited to visually representing longitudinal data than the marginal-model plots used earlier. We now add the second-order and interaction terms, and this model will henceforth be referred to as the ‘modified model’.

### 7.3 The modified model

We improved the initial model by adding second-order and interaction terms, as indicated by the diagnostic tests in the previous section. We included the second-order terms of \(TD(t)\) and \(D(t-1)\) and interaction terms that were entered during stepwise regression. The maximum-likelihood estimates of the modified model are summarized in Table 7.3. All covariates were significantly different from zero at a 0.005 significance level. Note that the covariates \(TP(t)\) and \(TR(t)\) were no longer used, as they were not significantly different from zero.
Figure 7-3: The CMM for the linear combination $U(t)\hat{\beta}$ of the initial model. The $\hat{E}_F \left[ \Delta N(t) \mid U(t)\hat{\beta} \right]$ in red on the left and the $\hat{E}_M \left[ \Delta N(t) \mid U(t)\hat{\beta} \right]$ in blue on the right.

Table 7.3: The parameter estimates for the failure model second order term and interaction terms added.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\hat{\beta}$</th>
<th>SE($\hat{\beta}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.756</td>
<td>0.709</td>
</tr>
<tr>
<td>TD(t)</td>
<td>1.018</td>
<td>0.028</td>
</tr>
<tr>
<td>D(t - 1)</td>
<td>0.257</td>
<td>0.066</td>
</tr>
<tr>
<td>Age(t)</td>
<td>-0.649</td>
<td>0.03</td>
</tr>
<tr>
<td>Prime(t - 3)</td>
<td>-0.521</td>
<td>0.054</td>
</tr>
<tr>
<td>Prime$^2$(t - 3)</td>
<td>1.085</td>
<td>0.106</td>
</tr>
<tr>
<td>N$^2$(t)</td>
<td>0.16</td>
<td>0.007</td>
</tr>
<tr>
<td>TD$^2$(t)</td>
<td>-0.695</td>
<td>0.028</td>
</tr>
<tr>
<td>D$^2$(t - 1)</td>
<td>-0.305</td>
<td>0.017</td>
</tr>
<tr>
<td>Age$^2$(t)</td>
<td>0.493</td>
<td>0.036</td>
</tr>
<tr>
<td>TD(t) $\times$ D(t - 1)</td>
<td>-0.157</td>
<td>0.011</td>
</tr>
<tr>
<td>Bond $\times$ D(t - 1)</td>
<td>0.053</td>
<td>0.005</td>
</tr>
</tbody>
</table>
7.4 Diagnostic tests on the modified model and model comparison

Again we calculated the marginal-model plots for the cross-sectional data at times $t = 48$ and $t = 24$. Inspecting Figure 7-5 and Figure 7-6 for $t = 48$ and $t = 24$ respectively it is clear that the covariates are mostly within the confidence bands. We included some random directions $a^{(1)}$ and $a^{(2)}$ (with smoothing parameters 0.7) as part of the marginal-model plot ensemble. The addition of the second-order terms for $D(t - 1)$ and $TD(t)$ seems to have improved the fit significantly. The $N(t)$ still exhibits the same piecewise linear relationship at $t = 48$.

Similarly to the initial model, we now constructed the CMM plot for the modified model in Figures 7-7 and 7-8. It is clear that the two surfaces cross each other much more than in the initial model. This points to a better fit.

We now derive an objective statistic to measure the distance between the two surfaces. In a similar way to the cross-sectional data in Section 3.4.3, we can calculate the mean-
Figure 7-5: The marginal-model plots for different choices of $a$ at time $t = 48$. 
Figure 7-6: The marginal-model plots for different choices of $a$ at time $t = 24$. 
Figure 7-7: The CMM for the modified model with linear combination $U(t)\hat{\beta}$. The $\hat{E}_F \left[ \Delta N(t) \mid U(t)\hat{\beta} \right]$ in red on the left and the $\hat{E}_M \left[ \Delta N(t) \mid U(t)\hat{\beta} \right]$ in blue on the right.

squared error ($MSE$) as a measure of the distance between the two surfaces. To simplify notation, let the loess surface be surface

$$\hat{\theta}_{F,t} \left( U_i(t)\hat{\beta} \right) = \hat{E}_{F,t} \left[ \Delta N_i(t) \mid U_i(t)\hat{\beta} \right]$$

and the postulated model surface be

$$\hat{\theta}_{M,t} \left( U_i(t)\hat{\beta} \right) = \hat{E}_{M,t} \left[ \Delta N_i(t) \mid U_i(t)\hat{\beta} \right].$$

We define the distance between the non-parametric loess surface and the AH model as

$$MSE_0 = \sum_{t=1}^{T} \sum_{i=1}^{n_t} \left( \hat{\theta}_{F,t} \left( U_i(t)\hat{\beta} \right) - \hat{\theta}_{M,t} \left( U_i(t)\hat{\beta} \right) \right)^2 / n,$$

where $n_t$ is the number of obligors at risk at time $t$ and $n = \sum_{t=1}^{T} n_t$. This measure sums over all time, and hence it is not dependent on the current time $t$. For the initial model
we found \(MSE_{\text{initial}} = 0.00072475\). The modified model improved the \(MSE_0\) by about 22% to \(MSE_{\text{modified}} = 0.00012426\). Knowing the absolute level of the \(MSE_0\) does not give us an indication of the model fit. Hence, similarly to the cross-sectional data case in Section 3.4.3, we can bootstrap a sampling density for the \(MSE\), say \(\hat{f}_{X^\beta}(MSE)\) and compare the value of \(MSE_0\) to the estimated density. The mean-squared error for the bootstrap sample \(U_i^*(t)\) is

\[
MSE^* = \sum_{t=1}^{T} \sum_{i=1}^{n} \left( \hat{\theta}_{F,t} \left( U_i(t)\hat{\beta} \right) - \hat{\theta}_{F,t} \left( U_i^*(t)\hat{\beta} \right) \right)^2 / n \tag{7.1}
\]

for \(b = 1, \ldots, B\). To estimate the density function \(f_{X^\beta}(MSE)\) we perform a semi-parametric bootstrap by assuming that the non-parametric estimated surface \(\hat{\theta}_{F,t} \left( U_i(t)\hat{\beta} \right)\) is the true probability of default given the linear combination of covariates \(U_i(t)\hat{\beta}\) at time \(t\). We denote the collection of covariates \(U_1, \ldots, U_n\) by \(U\) and sample with replacement \(n\) values \(U_1^*, \ldots, U_n^*\) from \(U\). Note that once an obligor \(i\) is sampled, we take all the time observations of that specific obligor. i.e. \(U_i(1), \ldots, U_i(T)\). Now compute
\( p_i(t) = \hat{\theta}_{F,t} \left( \mathbf{U}_i(t) \hat{\beta} \right) \) and generate binary random values

\[
Y_i^*(t) = \begin{cases} 
1 & \text{with probability } p_i(t) \\
0 & \text{with probability } 1 - p_i(t) 
\end{cases}
\]

for \( i = 1, \ldots, n \) and \( t = 1, \ldots, T \). By smoothing the surface of \( Y_i^*(t) \) against \( \mathbf{U}_i(t) \hat{\beta} \) and \( t \) we have a bootstrap estimate of \( \hat{\theta}_{F,t} \left( \mathbf{U}_i(t) \hat{\beta} \right) \). We can now estimate the bootstrapped sampling distribution of the surface \( \hat{\theta}_{F,t} \left( \mathbf{U}_i(t) \hat{\beta} \right) \), denoted by \( \hat{\theta}_{F,t}^* \left( \mathbf{U}_i(t) \hat{\beta} \right) \), by using a three-dimensional smoothing technique. We use a three-dimensional kernel smoothing technique. Moreover, for each bootstrap surface estimate of \( \hat{\theta}_{F,t} \left( \mathbf{U}_i(t) \hat{\beta} \right) \) we can compute the \( MSE \) as in (7.1).

The pseudo code for the estimation \( f_{\mathbf{X}\hat{\beta}}(MSE) \) is:

1. Using all data calculate \( \hat{\theta}_{F,t} \left( \mathbf{U}_i(t) \hat{\beta} \right), \hat{\theta}_{M,t} \left( \mathbf{U}_i(t) \hat{\beta} \right) \) and \( MSE_0 \).

2. Take a sample \( \mathbf{U}_1^*, \ldots, \mathbf{U}_n^* \) from \( \mathcal{U} \) with replacement and calculate \( p_i(t) = \hat{\theta}_{F,t} \left( \mathbf{U}_i^*(t) \hat{\beta} \right) \) for all obligors \( i \) over all time \( t = 1, \ldots, T \).

3. Generate binary random variables \( Y_i^*(t) \) with probability \( p_i(t) \).

4. Smooth the scatter surface of \( Y_i^*(t) \) against \( \mathbf{U}_i^*(t) \hat{\beta} \) and \( t \) to find bootstrap surface estimate \( \hat{\theta}_{F,t}^* \left( \mathbf{U}_i(t) \hat{\beta} \right) \).

5. Calculate the \( MSE^* \) in (7.1).

6. Repeat steps two to five \( B \) times.

7. Estimate the sampling distribution of \( MSE^* \) by using either a kernel-density or empirical density estimation method.

Our training set is large (about 300,000 observations), so a single non parametric surface smooth took about 2.5 hours on a personal computer (Intel Core i7 CPU 3.4 GHz with 24 GB RAM). Hence, we simulated only \( B = 50 \) bootstrap samples on both the initial and modified models and fitted a kernel-density function through the bootstrapped \( MSE^* \) values. The estimated-density function and \( MSE_{0 initial} \) are plotted in Figure 7-9. Notice that since the values are very small, we used a log scale. Clearly, the \( MSE_{0 initial} \)
Figure 7-9: The initial model has a $MSE_{0}^{\text{initial}}$ that is not close to the density of $f_{X_{\beta}}(MSE)$. It is far from $f_{X_{\beta}}^{\text{initial}}(MSE)$. We performed a similar analysis on the modified model and plotted the results in Figure 7-10. Here it is clear that $MSE_{0}^{\text{modified}}$ is not within the range of the density distribution, $f_{X_{\beta}}^{\text{modified}}(MSE)$. However, the distance is smaller than the distance in the initial model. This leads us to the conclusion that even though the modified model is not perfect, it is a large improvement over the initial model. We can also apply the heuristic goodness-of-fit model as discussed in Chapter 3 on page 71 for all data over all time. Using the same notation as in that section, both the initial model (in blue) and the modified model (in red) are depicted in Figure 7-11. It is evident that the modified model describes the data much better. We now use the failure model that we developed in this chapter combined with another counting process to derive a model for the prediction of delinquency.
Figure 7-10: The modified model has a $MSE_{\text{modified}}$ that is now closer to the density of $f_{X\beta_{\text{modified}}}(MSE)$, but is still far off.

Figure 7-11: The heuristic goodness-of-fit model. The blue is the initial model and the red is the modified model.
Chapter 8

Modelling delinquencies

In Chapter 6 we found that when predicting defaults the current delinquency level, $D(t - 1)$, as well as functions of the current and past delinquency levels, such as $TD(t)$, were of high statistical significance. Hence, in order to predict defaults further than one month forward we need to develop a model that predicts the delinquency levels for any number of months into the future. Ideally, this model should be dependent on macro-economic covariates to facilitate stress-testing. We are concerned with the evolution of $D_i(t - 1)$ over time, and in this chapter we propose two models to predict the delinquency level, $D_i(t)$ at some future time $t$ for obligor $i = 1, ..., n$.

To grasp the model-related complexities faced when modelling delinquency we make use of the following examples. In Figure 8-1 we represent examples of realizations of plausible delinquency rates for two obligors over a period of four calendar years. Obligor $A$ (the solid blue line) was part of the portfolio of the bank at $t = 0$. His delinquency started off at zero and gradually increased (with a few decreases) until the delinquency level peaked at six months overdue at $t = 29$. His account exited the portfolio of the bank at $t = 42$. We are not interested in the reason for the exit, whether it was early closure or a write-off. We treat both these events as censoring events. Incorporating censoring would be a trivial exercise and can be included in future models. Obligor $B$ (the dashed red line) opened his account at time $t = 11$ and was still part of the portfolio of the bank at $t = 48$. His delinquency also increased and then decreased, eventually increasing to two months in arrears at $t = 48$. He did not exit the portfolio of the bank, but our observation window was only up to time $t = 48$. We would like to develop a model that
can be used to predict the type of movements illustrated by obligor A and B, so not only failures should be modelled, as is done in Chapter 7, but repayments too.

We make the reasonable assumption that $D_i(t) \geq 0$ (an obligor is not permitted to pay his/her loan installments beforehand) and we assume that $D_i(t)$ may only increase by unit size in consecutive months, so $D_i(t) \in \{0, 1, ..., D_i(t - 1) + 1\}$. In other words, an obligor may not increase his delinquency level by more than two months in a one-month period. Modelling delinquency levels may be done with the use of transition matrices or by combining various counting processes. We first discuss the use of combined counting processes in Section 8.1 and then modelling delinquency with Markov chains in Section 8.2. Lastly, we compare the various approaches by measuring the prediction accuracy in Section 8.3.

### 8.1 Combined counting processes to model delinquency

A way of approaching the modelling of delinquency is by addressing two components separately: (1) a failure- (or missed-payment) part that results in an increase of the
delinquency level by one, i.e. the event $D_i(t + 1) = D_i(t) + 1$; and (2) a repayment part that models repayments made at any time while the obligor is delinquent. These delinquency repayments may be larger than size one, but not larger than the current delinquency level. So, $D_i(t) \in \{0, 1, \ldots, D_i(t - 1) + 1\}$. We need a model that governs processes like these and is dependent on systematic and idiosyncratic covariates.

If $D_i(t) = 0$ an obligor is zero months delinquent, implying that his loan is fully paid to date. The change in delinquency $\Delta D_i(t) = D_i(t) - D_i(t - 1)$ can be modelled by the difference between two counting processes. Firstly, we count the cumulative number of failures, $N_i(t)$, that obligor $i$ had up to time $t$. We can model the failures by using the change in the process $\Delta N_i(t) = N_i(t) - N_i(t - 1)$. Since $N_i(t)$ is increasing, we need to subtract a ‘repayment’ counting process from it. So, secondly, a repayment process, $B_i(t)$, should also be modelled. Again we can model this process by only using the change in the process $\Delta B_i(t) = B_i(t) - B_i(t - 1)$. The difference between the processes $N_i(t)$ and $B_i(t)$ is that $N_i(t)$ can only make unit-size jumps, whereas $B_i(t)$ can jump by any number so long as it is smaller than the current level of delinquency. Hence, the jump size of $B_i(t)$ is dependent on the current delinquency level $D_i(t)$. The repayment process is a marked-point process (with marks being the size of the repayments), whereas the failure part is a normal counting process. The failures can be modelled by an AH process $N_i(t)$ that has a unit value if the installment due was not paid and is zero if some payment (not necessarily only the installment due) was made. So

$$\Delta N_i(t) = \begin{cases} 
0 & \text{if some payment was made} \\
1 & \text{if no payment was made.}
\end{cases}$$

This is precisely the failure process developed in Chapter 7.

We stress the fact that we make the implicit assumption that if an obligor pays his normal installment, there is no change in delinquency. Hence, if only one payment was made in the month (the normal payment due), it implies that the delinquency level remains unchanged. If double the payment were to be made, the one part of the payment is seen as the normal installment that is due and the other as a repayment of some previously missed installment. To simplify matters we assume than only whole payments can be made, i.e. a delinquent obligor cannot pay say, 1.5 installments in a month.
The repayment process can be modelled by the process $B_i(t)$ such that the change in repayment is

$$
\Delta B_i(t) = \begin{cases} 
0 & \text{o if only the installment due was paid.} \\
1 & \text{o if the installment due and an additional} \\
\vdots & \text{delinquent months’ payment was made.} \\
D_i(t-1) - 1 & \text{o if the installment due and $D_i(t-1) - 1$ additional} \\
D_i(t-1) & \text{delinquent months’ payments were made.}
\end{cases}
$$

As mentioned earlier, the reason why $\Delta_i B(t)$ may only take on a maximum value of $D_i(t-1)$ is that the repayment may not exceed the current delinquency level.

We propose that the change in delinquency be modelled by

$$
\Delta D_i(t) = \Delta N_i(t) - \mathbf{1}\{D_i(t-1) > 0, \Delta N_i(t) = 0\} \Delta B_i(t).
$$

Interpreting (8.1), we see that if an obligor misses a payment, $\Delta N_i(t) = 1$, the delinquency level increases by one, i.e. $\Delta D_i(t+1) = 1$. If a payment was not missed, $\Delta N_i(t) = 0$, and the current delinquency is zero, $D_i(t-1) = 0$, then the current delinquency stays zero and the change in delinquency is $\Delta D_i(t+1) = 0$. If a payment was not missed, $\Delta N_i(t) = 0$, and the current delinquency is greater than zero, $D_i(t-1) > 0$, then the obligor is ‘eligible’ for an additional repayment. We assume that if an obligor is currently not delinquent and does not miss a payment, his or her delinquency level will stay at zero.

### 8.1.1 A truncated distribution approach

As mentioned earlier, we propose that $P(\Delta N_i(t) = 1)$ be estimated by the AH process discussed in Chapter 7. Therefore, we only need a model for the probability

$$
P(\Delta B_i(t) = k | \Delta N_i(t) = 0, D_i(t-1) > 0)
$$
for \( k = 0, 1, \ldots, D_i(t - 1) \). The repayment process \( \mathbf{1} \{D_i(t - 1) > 0, N_i(t) = 0\} \Delta B_i(t) \) may not exceed \( D_i(t - 1) \), and hence is truncated at \( D_i(t - 1) \). So we will use a truncated distribution to prevent the repayment process from exceeding \( D_i(t - 1) \). To simplify notation, set

\[
\Delta B_i^*(t) := \mathbf{1} \{D_i(t - 1) > 0, N_i(t) = 0\} \Delta B_i(t).
\]

From our training data set, we drew the relative histogram of \( \Delta B_i^*(t) \) in Figure 8-2. We did this for all observations at an arbitrary chosen time \( t = 24 \). Clearly, the most overdue obligors (about 68%) only made the installment due, and hence no additional repayments were made. The second-largest group (25%) made the installment due as well as one additional repayment, and so on. The shape of the distribution in Figure 8-2 is similar to a geometric distribution. So, we propose the use of a truncated geometric distribution to model \( \Delta B_i^*(t) \). We discuss the properties of the geometric truncated distribution in the appendix at the end of the chapter. Other distributions, like the zero-inflated Poisson distribution (Lambert 1992) could also have been truncated (Fokianos 2001).
8.1.2 Modelling repayments: a real-time AH-type truncated geometric model

We now go about modelling $\Delta B_i^*(t)$. Since this model is conditional on the fact that no failure occurred, we call the ‘eligibility’ process

$$Y_{i}^*(t) := 1 \{ D_i(t - 1) > 0, Y_i(t) = 1, \Delta N_i(t) = 0 \}. $$

Hence, an obligor is eligible for a repayment event if he did not miss a payment and is currently delinquent. The process $Y_{i}^*(t)$ is predictable; therefore, in order to predict $\Delta B_i^*(t)$ we only need a model to predict $Y_{i}^*(t)$. We follow similar arguments as AH and make four assumptions, using similar notation as in the previous section, Section 5.2, except that we replace the at risk indicator $Y_i(t)$ with an eligibility indicator, $Y_{i}^*(t) = 1 \{ D_i(t - 1) > 0, Y_i(t) = 1, \Delta N_i(t) = 0 \}$. Additionally, we denote the eligible risk set by

$$R^*(t-1) = \{ i \geq 1 : Y_{i}^*(t - 1) = 1 \} \text{ in } (t-1; t].$$

Hence, obligor $i$ is eligible for a repayment when his account is currently all of the following: (1) in arrears, (2) at-risk (or still active) and (3) we know that he did not miss a payment in the current month.

We follow AH in defining the $\sigma$-field inductively. Let $V_i(t-1)$ be the time-dependent design matrix of covariates, let $N_{t}^{(i)}$ and $M_{t-1}^{(i)}$ be the $\sigma$-field for each obligor $i$, $t \geq 0$ then

$$M_{t-1}^{(i)} = N_{t-1}^{(i)} \cup \sigma \{ Y_j^* (t-1), V_i(t-1) \},$$

$$N_{t}^{(i)} = M_{t-1}^{(i)} \cup \sigma \{ \Delta B_i(t) \}.$$ 

We interpret $M_{t-1}^{(i)}$ as the at-risk set of obligor $i$ and its covariates at time $t$, excluding any repayments at time $t$. $N_{t}^{(i)}$ is the union of $M_{t-1}^{(i)}$ and the repayments in $(t-1, t]$. So $N_{t}^{(i)}$ contains all the information of the eligible set, covariates and repayments on time $t$. We explain the derivation of the full likelihood and the partial likelihood expressions that will be used for parameter estimation below.
For all obligors \( i = 1, 2, \ldots, n \) the \( \sigma \)-fields are

\[
\mathcal{M}_{t-1} = \bigcap_j \mathcal{M}^{(i)}_{t-1} \\
\mathcal{N}_{t-1} = \bigcap_j \mathcal{N}^{(i)}_{t-1}.
\]

The likelihood for obligor \( i \) with parameter set \( \theta \) is

\[
Lik^{(i)}_t(\theta) = P \left[ \bigcap_{v \leq t} Y^*_j (v - 1) = y^*_i (v - 1), \mathbf{V}_i(v - 1) = \mathbf{v}_i(v - 1), \Delta B_i(v) = \Delta b_i(v), D_i(v - 1) = d_i(v - 1) \right]
\]

\[
= \prod_{v \leq t} P \left[ Y^*_j (t - 1) = y^*_i (t - 1), \mathbf{V}_i(v - 1) = \mathbf{v}_i(v - 1) | \mathcal{N}^{(i)}_{v-1} \right] \times \prod_{v \leq t} P \left[ \Delta B_i(v) = \Delta b_i(v) | \mathcal{M}^{(i)}_{v-1} \right].
\] (8.2)

**Assumption 1.**

The parameter set \( \theta \) can be decomposed into two \( \theta = (\alpha^*, \alpha) \): (i) The first factor of the full likelihood in (8.2),

\[
\prod_{v \leq t} P \left[ Y^*_i (v - 1) = y^*_i (v - 1), \mathbf{V}_i(v - 1) = \mathbf{v}_i(v - 1) | \mathcal{N}^{(i)}_{v-1} \right],
\]

is assumed to only be dependent on \( \alpha^* \) and not \( \alpha \). (ii) The second factor,

\[
\prod_{v \leq t} P \left[ \Delta B_i(v) = \Delta b_i(v) | \mathcal{M}^{(i)}_{v-1} \right],
\]

is assumed to only be dependent on the parameters \( \alpha \) and not \( \alpha^* \). We assume that we are only interested in \( \alpha \), so \( \alpha^* \) is a nuisance parameter.

Applying this assumption, the partial likelihood for the parameter \( \alpha \) is

\[
Lik^{(i)}_t(\alpha) = \prod_{v \leq t} P \left[ \Delta B_i(v) = \Delta b_i(v) | \mathcal{M}^{(i)}_{v-1} \right],
\] (8.3)

for \( v \geq 1 \).
Assumption 2.
For each \( v \geq 1 \) and \( i = 1, \ldots, n \) the random variables \( \Delta B_i(v) \) are conditionally independent given \( M^{(i)}_{v-1} \). Assumption 2, together with the fact that only eligible obligors make a contribution to the likelihood, implies that we can combine the partial likelihood in (8.3) by

\[
Lik_t(\alpha) = \prod_{i \in R^*(v)} Lik^{(i)}_t(\alpha).
\]

So

\[
Lik_t(\alpha) = \prod_{i \in R^*(v)} \prod_{v \leq t} P \left[ \Delta B_i^*(v) = \Delta b_i^*(v) | M^{(i)}_{v-1} \right].
\] (8.4)

We further simplify the partial likelihood by following AH in making the next assumptions.

Assumption 3.
\( \Delta B_i(v) \) and \( M^{(i)}_{v-1} \) are conditionally independent given \( Y_i^*(v - 1) \) and \( V_i(v - 1) \) for all \( v, i \geq 1 \). That is, for all \( A \in M^{(i)}_{v-1} \)

\[
P[\Delta B_i(v) = \Delta b_i(v) | A, Y_i^*(v - 1), V_i(v - 1)]
\]

\[= P[\Delta B_i(v) = \Delta b_i(v) | Y_i^*(v - 1), V_i(v - 1)]].
\]

This assumption ensures that all relevant information in \( M^{(i)}_{v-1} \), when used as conditioning for the probability of the events \( \{\Delta N_i(v) = \Delta n_i(v)\} \), is contained in \( V_i(v - 1) \) and \( Y_i^*(v - 1) \). Therefore, the partial likelihood function is only dependent on the information contained in the previous time step. Note that we do not specify a Markovian assumption in this process, since the values of covariates at time \( t < v - 1 \) may be included as covariates in the vector \( V_i(v - 1) \). This assumption reduces (8.4) further to

\[
Lik_t(\alpha) = \prod_{v \leq t} \prod_{i \in R^*(v)} P[\Delta B_i(v) = \Delta b_i(v) | V_i(v - 1), Y_i^*(v - 1)].
\] (8.5)

We now need to specify some statistical model for the probabilities in (8.5). We assume that the probabilities in (8.5) are a GLM with a truncated geometric distribution function.
**Assumption 4.**

The probability of obligor $i$ being eligible for a repayment is given by

$$P[\Delta B_i(v) = \Delta b_i(v)|V_i(v-1), Y_i^*(v-1) = 1] = \frac{\tau_{i,v} (1 - \tau_{i,v}) \Delta b_i(v)}{1 - (1 - \tau_{i,v}) D_i(v-1) + 1}$$

for all $v \geq 0$ and $i = 1, ..., n$. Simplifying this equation further by removing the $Y_i^*(v-1) = 1$ from the conditioning we get the probability of a repayment occurring:

$$P[\Delta B_i(v) = \Delta b_i(v)|V_i(v-1)] = Y_i^*(v-1) \frac{\tau_{i,v} (1 - \tau_{i,v}) \Delta b_i(v)}{1 - (1 - \tau_{i,v}) D_i(v-1) + 1}.$$

If we substitute these probabilities into the partial likelihood $Lik_i(\alpha)$ in (8.5) we get

$$Lik_i(\alpha)$$

$$= \prod_{v \leq t} \prod_{i \in R^v(v)} P[\Delta B_i(v) = \Delta b_i(v)|V_i(v-1), Y_i^*(v-1) = 1]$$

$$= \prod_{v \leq t} \prod_{i} \left( \frac{\tau_{i,v} (1 - \tau_{i,v}) \Delta b_i(v)}{1 - (1 - \tau_{i,v}) D_i(v-1) + 1} \right) Y_i^*(v-1).$$

We further assume that the parameter $\tau_{i,v}$ has the form

$$\text{logit}(\tau_{i,v}|\mathcal{M}^{i}_{v-1}) = V_i(v-1) \alpha,$$

for all $i$ and $v$ such that $Y_i^*(v-1) = 1$ and $v \geq 0$, where $V(v-1)$ has dimension $(n \times n_V)$. Note that the link function used for $\tau_v$ is not the canonical link, but rather a more sensible link function that bounds $\tau$ between zero and one.
8.1.3 Parameter estimation

Making the aforementioned four assumptions, we can estimate the parameter $\alpha$ by maximizing the partial log-likelihood

\[
lik_i(\alpha | M_{t-1}) = \sum_{v=1}^{t} \sum_{i=1}^{n} Y_i^* (v - 1) \log \left( \frac{\tau_{i,v} (1 - \tau_{i,v})^{b_i(v)}}{1 - (1 - \tau_{i,v})^{b_i(v) + 1}} \right). \tag{8.6}
\]

This can be done with a Newton-Raphson procedure, the details of which are given in the appendix at the end of the chapter. We estimated the parameters and used the Wald test to determine which covariates are significantly different from zero. We added first- and second-order terms, and all two-tailed $p$-values for all covariates were smaller than 0.0001. The resulting parameter estimates are summarized in Table 8.1.

Table 8.1: The parameter estimates for the geometric truncated distribution for modeling recoveries.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\hat{\alpha}$</th>
<th>SE($\hat{\alpha}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-6.121</td>
<td>0.54</td>
</tr>
<tr>
<td>$t$</td>
<td>0.018</td>
<td>0.002</td>
</tr>
<tr>
<td>TD($t$)</td>
<td>2.366</td>
<td>0.134</td>
</tr>
<tr>
<td>TD$^2$(t)</td>
<td>-1.154</td>
<td>0.082</td>
</tr>
<tr>
<td>D$^2$(t - 1)</td>
<td>-0.057</td>
<td>0.006</td>
</tr>
<tr>
<td>Age($t$)</td>
<td>0.455</td>
<td>0.13</td>
</tr>
<tr>
<td>Age$^2$($t$)</td>
<td>-0.48</td>
<td>0.137</td>
</tr>
<tr>
<td>TP($t$)</td>
<td>-0.123</td>
<td>0.02</td>
</tr>
<tr>
<td>TR($t$)</td>
<td>0.513</td>
<td>0.119</td>
</tr>
<tr>
<td>TR$^2$($t$)</td>
<td>-0.409</td>
<td>0.112</td>
</tr>
<tr>
<td>Prime($t - 3$)</td>
<td>0.118</td>
<td>0.012</td>
</tr>
<tr>
<td>Rs($t - 3$)</td>
<td>0.07</td>
<td>0.023</td>
</tr>
<tr>
<td>Lead($t - 3$)</td>
<td>0.099</td>
<td>0.019</td>
</tr>
</tbody>
</table>

Having estimated the parameters, we can now compare this model with the matrix approaches developed in the next section.
8.2 Markov chains to model delinquency

A common approach to predict delinquency levels is by using Markov chains. The elements of a delinquency transition-matrix are commonly used for accounting purposes and are often referred to as ‘roll rates’ as they denote the probability of moving from one delinquency state to the next in consecutive months (Grimshaw & Alexander 2010). Let $X_t$ denote the state of the Markov chain at time $t$ with state space $\{0, 1, ..., K\}$ and $t = 1, ..., T$. Then

$$P(X_t = k | X_{t-1} = j)$$

is the probability of an obligor moving from delinquency state $j$ to delinquency state $k$ from time $t - 1$ to time $t$ with $j, k \in \{0, 1, ..., K\}$ and $t = 1, 2, ..., T$. Should the transition-matrix prediction model be able to predict one month forward, then theoretically the model can predict the delinquency level for any number of months. The probability model should (like all the other models developed) ideally be dependent on idiosyncratic and systematic covariates. Typically, a natural progression through the various levels of delinquency is observed, e.g. an obligor may only go from being one month delinquent to being two months delinquent in sequential months. An obligor may not jump from being one month delinquent to being, say, five months delinquent in one month’s time.

This is in line with the assumption made earlier in Section 8.1 that $0 \leq \Delta N(t) \leq 1$. To simplify this presentation we only used seven delinquency states ($K = 6$) for the transition-matrix, namely: $0, 1, ..., 6$. This model implies that 34 transition probabilities need to be estimated. This makes the transition-matrix approach less flexible than the combined counting process approach, where the number of months delinquent is theoretically unbounded. We assumed that all obligors with a delinquency level higher or equal to seven is written off and only estimated the parameters on historical data where accounts were open, i.e. ‘at-risk’. Hence, we do not have any absorbing states and we are only interested in the modelling of delinquency. However, should we be interested in modelling the absorbing states like write-off or early closure (censoring events in the context of Section 8.1), it would be straightforward to include these absorbing states in the transition-matrix. We now look at two approaches to estimate the probabilities in (8.7).
### 8.2.1 Stationary transition matrices

One way to estimate the transition probabilities $P(X_t = k|X_{t-1} = j)$ for $j, k \in \{0, 1, ..., K\}$ is simply dividing the number of obligor transitions from state $j$ to state $k$ from time $t-1$ to $t$ (denoted by $n^t_{j,k}$) by the total number of obligors in state $j$ at time $t-1$ (denoted by $n^t_j = \sum_{k=1}^{K} n^t_{j,k}$). So, we may estimate $P(X_t = k|X_{t-1} = j)$ by

$$
\hat{p}^t_{j,k} = \frac{n^t_{j,k}}{n^t_j},
$$

for $j, k \in \{0, 1, ..., K\}$ at time $t = 1, 2, ..., T$. After estimating these probabilities in the data at each $t$ we found that the transition probabilities $\hat{p}^t_{j,k}$ exhibit trends over time. This can be seen in Figure 8-3 where we plot $\hat{p}^t_{1,0}$, $\hat{p}^t_{1,1}$ and $\hat{p}^t_{1,2}$ for $t = 1, ..., 48$. Clearly, trends are present that imply non-homogeneity (Mattsson & Thorburn 1989). Hence, the ubiquitous assumption of stationarity often found when applying Markov chains is not applicable to the data. Nevertheless, we use the stationarity assumption as a naïve comparison model. The stationarity assumption is

$$
\hat{p}_{j,k} = \hat{p}^t_{j,k}
$$

for all $t = 1, 2, ..., T$. Therefore, we estimate the stationary transition probabilities by

$$
\hat{p}_{j,k} = \frac{\sum_{t=1}^{T} n^t_{j,k}}{\sum_{t=1}^{T} n^t_j}.
$$

We estimated these transition probabilities on a training set of the data and summarized the transition probabilities in Table 8.2. The diagonal values are very large, implying that

<table>
<thead>
<tr>
<th>From/To</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.934</td>
<td>0.066</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.320</td>
<td>0.446</td>
<td>0.234</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.123</td>
<td>0.136</td>
<td>0.413</td>
<td>0.327</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.085</td>
<td>0.038</td>
<td>0.115</td>
<td>0.377</td>
<td>0.384</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.066</td>
<td>0.020</td>
<td>0.035</td>
<td>0.101</td>
<td>0.363</td>
<td>0.416</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.058</td>
<td>0.012</td>
<td>0.014</td>
<td>0.035</td>
<td>0.084</td>
<td>0.357</td>
<td>0.441</td>
</tr>
<tr>
<td>6</td>
<td>0.098</td>
<td>0.013</td>
<td>0.010</td>
<td>0.031</td>
<td>0.042</td>
<td>0.131</td>
<td>0.674</td>
</tr>
</tbody>
</table>
Figure 8-3: The transition probabilities from a delinquency of one to either zero or two or staying at one. The red horizontal dotted line is the estimated stationary transition probability.
most obligors stay at their current delinquency level. Grimshaw and Alexander (2010) extend this naïve model to be dependent on covariates.

8.2.2 The Grimshaw and Alexander approach

An approach to solve the non-stationarity problem is to apply the approach of Grimshaw and Alexander (2010), henceforth abbreviated as GA. They estimate the transition matrices by using either of (or a combination of) two procedures:

1. Identify some factor that will segment the transition matrices into various segments. This can be based on the "characteristics of the financial product". The segmentation is done such that the transition matrices are homogeneous, which simplifies the estimation and prediction later on.

2. Incorporating covariates into the transition matrices and estimating the transition probabilities for each obligor \( i \). Here, the transition probabilities are functions of the obligors’ idiosyncratic covariates as well as systematic covariates.

GA approach the estimation of the first procedure by using Bayesian estimation and empirical Bayes estimation. The choice of segmentation characteristic is subjective and no clear indication or guideline is given on how to select the segmentation characteristic. Hence, we are of opinion that the second procedure (driven by covariates) should, to a large extent, account for the segmentation as advocated by their first procedure. Should the segmentation be done on account of some characteristic (which can normally be coded as a dummy covariate), we feel that this covariate should be included in the second procedure. Therefore, we will be implementing the second procedure with a large number of potential covariates.

A simple way to model the transition-matrix is by modelling each row as a separate multinomial model (also known as polychotomous logistic regression). According to Grimshaw and Alexander, this approach has some problems associated with it: (1) not all probabilities need to be modelled. In our case, the movement from a delinquency level of state \( j \) to a level of \( j + q \) with \( q \geq 2 \) is not allowed, as is clear from the transition-matrix in Table 8.2. (2) Not all covariates driving the probabilities \( P (X_t = k | X_{t-1} = j) \) are the same, so numerous different covariates need to be taken into account for all different
transitions. For example, moving from a delinquency state 1 to 2 may be driven by a covariate $U_1$, while moving the opposite way might not be a function of $U_1$ at all. (3) It is overly complicated to model a transition-matrix with many states.

Instead of using polychotomous logistic regression Grimshaw and Alexander follow an approach developed by Begg and Gray (1984) where "a series of separate simple logistic regression analysis are performed as a replacement for polychotomous logistic regression".

This approach addresses the three problems associated with multinominal regression and creates extra flexibility and simplicity regarding variable selection. Begg and Gray (1984) show that the coefficients obtained in this manner are consistent and have high efficiency. Hosmer and Lemeshow (2000 p. 279) compare this approach, which they call the independent logistic regression (ILR), to standard multinominal regression. They find the ILR approach "provides a good approximation to both the estimates of the coefficients and estimates of the standard errors".

The ILR approach vests in the collection of equations

$$P(X_t = k | X_{t-1} = j) = \exp \left( \mathbf{U}(t)_{j,k} \beta_{t,j,k} \right)$$

(8.8)

for $k \in \{0, 1, \ldots, K\} \setminus \{j\}$, where $\mathbf{U}(t)_{j,k}$ is the covariate matrix associated with all obligors moving from state $j$ to state $k$ and $\beta_{t,j,k}$ is the parameter at time $t$. Hence, a referent or baseline state is chosen (in this case we choose the diagonal state $j$) and all other states are compared to this state via the equations in (8.8).
Using (8.8) we can obtain a logistic regression form

\[
\frac{\exp \left( \mathbf{U}(t)_{j,k} \hat{\alpha}_{t,j,k} \right)}{1 + \exp \left( \mathbf{U}(t)_{j,k} \hat{\alpha}_{t,j,k} \right)}
\]

\[
= \frac{P(X_t = k|X_{t-1} = j)}{1 + P(X_t = k|X_{t-1} = j)(P(X_t = j|X_{t-1} = j))^{-1}}
\]

\[
= \frac{P(X_t = k|X_{t-1} = j)}{P(X_t = j|X_{t-1} = j) + P(X_t = k|X_{t-1} = j)}
\]

\[
= \frac{P(X_t = k|X_{t-1} = j)}{P((X_t = j) \cup (X_t = k)|X_{t-1} = j)}
\]

\[
= P(X_t = k| (X_{t-1} = j) \cap ((X_t = j) \cup (X_t = k)))
\]

\[
= P(X_t = k| ((X_{t-1} = j) \cap (X_t = j)) \cup ((X_{t-1} = j) \cap (X_t = k))). \quad (8.9)
\]

Therefore, to model the individual movements from state \( j \in \{0, 1, ..., K\} \) at \( t - 1 \) to state \( k \in \{0, 1, ..., K\} \setminus \{j\} \) at \( t \) as a logistic regression, we compare all the obligors who moved to either state \( j \) or state \( k \) at time \( t \) from state \( j \) at time \( t - 1 \) and no other states. A ‘success’ is given by the occurrence of a movement to state \( k \in \{0, 1, ..., K\} \setminus \{j\} \) in time \( t \). Herein lies the key difference between the ILR approach and multinomial regression. In multinomial regression, we would have modelled the occurrence of a movement form \( j \) at \( t - 1 \) to state \( k \in \{0, 1, ..., K\} \setminus \{j\} \) at \( t \) (also seen as a ‘success’, with baseline state \( j \)) compared to the movement to any other state in \( t \). In ILR, only the state of interest \( k \in \{0, 1, ..., K\} \setminus \{j\} \) at \( t \) and the base state \( j \) at \( t - 1 \) are compared. Therefore, an obligor only makes a contribution to the ILR likelihood if the obligor is in state \( j \) at time \( t - 1 \) and stays in state \( j \) or moves to state \( k \in \{0, 1, ..., K\} \setminus \{j\} \) at time \( t \). Using IRLS we estimate the parameters of (8.9), \( \hat{\beta}_{t,j,k} \) for a specific time \( t \). We can now calculate the transition probabilities for each obligor \( i \) by using the normalizing constant

\[
C_j = 1 + \sum_{k \in \{0, 1, ..., K\} \setminus \{j\}} \exp \left( \mathbf{U}(t)_{j,k} \hat{\beta}_{t,j,k} \right)
\]
for $j = 0, 1, ..., K$. The estimated transition probabilities ($\tilde{p}_{j,k}^t$) are given by

$$\tilde{p}_{j,k}^t = \begin{cases} \frac{\exp(U_{j,k}^t)}{C_j} & \text{for } k \in \{0, 1, ..., K\} \setminus \{j\} \\ \frac{1}{C_j} & \text{for } k = j. \end{cases}$$

A problem associated with the GA approach is that including macro-economic covariates leads to ambiguity in the model. The value of macro-economic covariates changes over time, but they are fixed at each time period for all obligors. So, estimating the parameters with time-dependent macro-economic covariates included in the design matrix results in a singular design matrix at each time step. An identical problem was encountered in the DTYG model as discussed in Section 5.1. As seen in comparing the AH model to the DTYG model in Chapter 5, it is simpler to predict or stress-test the response by manipulating the covariates over time, but leaving the parameter estimates constant. We overcome this problem by realizing that we have a special application of the AH model here. Making similar assumptions as in the AH approach (such as the use of partial likelihood, conditional independence and logit transformation), we may treat all observations over all time as a single logistic regression model (see Section 5.2). As we have seen with the AH model, this comes down to applying the logistic regression model to all at-risk data irrespective of the time of occurrence. In applying this methodology it is now perfectly viable to include macro-economic covariates in the design matrix. We estimated the transition probabilities for all obligors that are at-risk by maximizing the likelihood of the following probability models for each obligor $i$:

$$P(\{X_t = k\} \mid \{X_{t-1} = j\} \cup \{X_t = j\})$$

$$= \frac{\exp(U_{j,k}^t \beta^*_j)}{1 + \exp(U_{j,k}^t \beta^*_j)} \tag{8.10}$$

where

$$U_{j,k}^t = \begin{bmatrix} U_{j,k}(1) \\ U_{j,k}(2) \\ \vdots \\ U_{j,k}(T) \end{bmatrix}$$
for \( j \in \{0, 1, ..., K\} \) and \( k \in \{0, 1, ..., K\} \setminus \{j\} \). The parameter estimates \( \beta_{j,k}^* \) can now be used to compute the normalizing constant

\[
C_j^* = 1 + \sum_{k \in \{0,1,...,K\} \setminus \{j\}} \exp\left( U_{j,k}^* \beta_{j,k}^* \right) \tag{8.11}
\]

and the transition probabilities can be estimated by

\[
\tilde{p}_{j,k} = \begin{cases} 
\frac{\exp( U_{j,k}^* \beta_{j,k}^*)}{C_j^*} & \text{for } k \in \{0,1,...,K\} \setminus \{j\} \\
\frac{1}{C_j^*} & \text{for } k = j.
\end{cases} \tag{8.12}
\]

This model is particularly well suited to the application of various variable-selection methods. Note that for each combination of \( j = 0, 1, ..., K \) and \( k \in \{0, 1, ..., K\} \setminus \{j\} \) in (8.10) other covariates can be used. Hence, the dimensions of \( \textbf{U}^*_{j,k} \) may differ for every \( j \) and \( k \). In fact, we started with \( \textbf{U}^* \) having large dimensions and all possible covariates. Thereafter, we applied a forward stepwise regression technique to identify the most significant covariates for each possible transition probability. Any other variable-selection method could have been used. We only selected ten covariates at most to keep the model concise, but more could have been used. The simplest model, which has only an intercept term and no covariates (for all transition probabilities), would correspond to the stationary approach discussed in Section 8.2.1. We estimated the parameters by using the following pseudo code:

1. Choose row \( j = 1 \).

2. Estimate the individual logistic regression parameters (using a forward stepwise selection criterion) \( \hat{\beta}_{j,k}^* \) over all times for all \( k \leq j + 1, k \in \{0,1,...,K\} \setminus \{j\} \) using (8.10).

3. Calculate \( C_j^* \) in (8.11).


5. While \( j < K \) set \( j = j + 1 \) otherwise stop.
We applied this procedure to the data and used a forward stepwise regression to estimate the significant parameters. As mentioned previously, we intend on keeping the model concise, so we set the maximum number of significant parameters in each transition probability equal to 10. The parameter estimates of all 34 transition probabilities are cumbersome to summarize, as is apparent from Tables 8.3 to 8.4. The superscript depicts the order of their inclusion in the stepwise model.

### Table 8.3: The parameter estimates for the GA model.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>( \beta_{0,1} )</th>
<th>( \beta_{1,0} )</th>
<th>( \beta_{1,2} )</th>
<th>( \beta_{2,0} )</th>
<th>( \beta_{2,1} )</th>
<th>( \beta_{2,2} )</th>
<th>( \beta_{3,0} )</th>
<th>( \beta_{3,1} )</th>
<th>( \beta_{3,2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.25(^{\text{a}})</td>
<td>2.23(^{\text{b}})</td>
<td>-12.13(^{\text{b}})</td>
<td>0.56(^{\text{b}})</td>
<td>-5.24(^{\text{b}})</td>
<td>-8.52(^{\text{b}})</td>
<td>3.01(^{\text{b}})</td>
<td>-1.07(^{\text{b}})</td>
<td>-11.45(^{\text{b}})</td>
</tr>
<tr>
<td>TD(t)</td>
<td>-0.89(^{\text{b}})</td>
<td>0.13(^{\text{b}})</td>
<td>-1.99(^{\text{b}})</td>
<td>-6.11(^{\text{b}})</td>
<td>-1.24(^{\text{b}})</td>
<td>3.4(^{\text{b}})</td>
<td>-0.59(^{\text{b}})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TD(^2)(t)</td>
<td>-3.01(^{\text{a}})</td>
<td>-0.31(^{\text{a}})</td>
<td>0.07(^{\text{a}})</td>
<td>-0.25(^{\text{a}})</td>
<td>-0.12(^{\text{a}})</td>
<td>-9.38(^{\text{a}})</td>
<td>0.16(^{\text{a}})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>0.06(^{\text{b}})</td>
<td>-0.16(^{\text{b}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>1.33(^{\text{b}})</td>
<td>-0.06(^{\text{b}})</td>
<td>-0.65(^{\text{b}})</td>
<td>-0.29(^{\text{b}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age(^2)</td>
<td>0.07(^{\text{b}})</td>
<td>-0.05(^{\text{b}})</td>
<td>-16.84(^{\text{b}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N(t)</td>
<td>0.51(^{\text{a}})</td>
<td>-0.4(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N(^2)(t)</td>
<td>1.86(^{\text{a}})</td>
<td>-0.4(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>0.17(^{\text{b}})</td>
<td>-0.54(^{\text{b}})</td>
<td>-0.68(^{\text{b}})</td>
<td>-0.11(^{\text{b}})</td>
<td>-0.22(^{\text{b}})</td>
<td>-15.41(^{\text{b}})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prime(_{t-3})</td>
<td>1.86(^{\text{a}})</td>
<td>0.17(^{\text{a}})</td>
<td>0.21(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TR</td>
<td>-0.2(^{\text{b}})</td>
<td>0.42(^{\text{a}})</td>
<td>0.45(^{\text{a}})</td>
<td>0.05(^{\text{a}})</td>
<td>0.29(^{\text{a}})</td>
<td>0.46(^{\text{a}})</td>
<td>0.89(^{\text{a}})</td>
<td>-0.4(^{\text{a}})</td>
<td></td>
</tr>
<tr>
<td>TR(^2)</td>
<td>1.02(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond</td>
<td>-0.59(^{\text{b}})</td>
<td>-1.66(^{\text{b}})</td>
<td>-0.15(^{\text{b}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td>0.49(^{\text{a}})</td>
<td>1.53(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FR(^2)</td>
<td>0.09(^{\text{a}})</td>
<td>-0.02(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR</td>
<td>0.57(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR(^2)</td>
<td>0.17(^{\text{a}})</td>
<td>0.04(^{\text{a}})</td>
<td>0.11(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIns(_{t-6})</td>
<td>0.32(^{\text{a}})</td>
<td>-0.05(^{\text{a}})</td>
<td>0.42(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(_{t-3})$</td>
<td>0.15(^{\text{a}})</td>
<td>-1.8(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lead(_{t-3})</td>
<td>-0.11(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPI</td>
<td>0.12(^{\text{a}})</td>
<td>-0.19(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td>0.26(^{\text{a}})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 8.3 Model prediction comparison

In this section we determine which of the following three models is better suited to predict delinquencies: (1) the combined counting process model (CCPM) in Section 8.1, (2) the stationary transition-matrix or (3) the modified GA approach. We decided to use the
Table 8.4: The parameter estimates for the GA model

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\hat{\beta}_{3,4}$</th>
<th>$\hat{\beta}_{4,0}$</th>
<th>$\hat{\beta}_{4,1}$</th>
<th>$\hat{\beta}_{4,2}$</th>
<th>$\hat{\beta}_{4,3}$</th>
<th>$\hat{\beta}_{4,4}$</th>
<th>$\hat{\beta}_{4,5}$</th>
<th>$\hat{\beta}_{4,6}$</th>
<th>$\hat{\beta}_{4,7}$</th>
<th>$\hat{\beta}_{4,8}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$-12.58^3$</td>
<td>$10.44^0$</td>
<td>$-1.5^0$</td>
<td>$-0.99^0$</td>
<td>$-15.47^0$</td>
<td>$-18.3^0$</td>
<td>$13.43^0$</td>
<td>$-3.64^0$</td>
<td>$-20.56^0$</td>
<td></td>
</tr>
<tr>
<td>TD(t)</td>
<td>$-0.51^3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TD$^2$(t)</td>
<td>$-0.42^1$</td>
<td>$0.2^6$</td>
<td>$-0.52^1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$2^1$</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td>$0.32^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age$^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$-0.19^1$</td>
</tr>
<tr>
<td>N(t)</td>
<td>$-10.59^4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$-0.49^4$</td>
</tr>
<tr>
<td>N$^2$(t)</td>
<td>$0.1^0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>$3.64^2$</td>
<td>$-0.57^3$</td>
<td>$-18.11^1$</td>
<td>$-1.52^2$</td>
<td>$-0.7^1$</td>
<td></td>
<td></td>
<td>$0.83^3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prime$_{t-3}^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$-2.66^1$</td>
</tr>
<tr>
<td>TR</td>
<td>$-1.25^6$</td>
<td>$11.52^4$</td>
<td>$0.31^2$</td>
<td>$0.26^3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$0.44^1$</td>
</tr>
<tr>
<td>TR$^2$</td>
<td></td>
<td>$-0.45^1$</td>
<td>$0.26^3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td>$-0.51^8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FR$^2$</td>
<td>$0.05^9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR</td>
<td>$1.12^7$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$-1.54^1$</td>
</tr>
<tr>
<td>CR$^2$</td>
<td>$-0.07^5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R$_{t-3}$</td>
<td>$-0.98^3$</td>
<td></td>
<td>$-0.2^3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$-25.45^2$</td>
</tr>
<tr>
<td>PPI</td>
<td>$-0.14^1$</td>
<td></td>
<td>$0.26^4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$0.11^2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\hat{\beta}_{5,3}$</th>
<th>$\hat{\beta}_{5,4}$</th>
<th>$\hat{\beta}_{5,6}$</th>
<th>$\hat{\beta}_{6,0}$</th>
<th>$\hat{\beta}_{6,1}$</th>
<th>$\hat{\beta}_{6,2}$</th>
<th>$\hat{\beta}_{6,3}$</th>
<th>$\hat{\beta}_{6,4}$</th>
<th>$\hat{\beta}_{6,5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$-1.44^0$</td>
<td>$-16.4^0$</td>
<td>$-21.17^0$</td>
<td>$27.07^0$</td>
<td>$-72.38^0$</td>
<td>$-0.42^0$</td>
<td>$-3.32^0$</td>
<td>$1.78^0$</td>
<td>$-15.29^0$</td>
</tr>
<tr>
<td>TD(t)</td>
<td>$-0.22^8$</td>
<td>$31.6^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TD$^2$(t)</td>
<td>$-0.34^1$</td>
<td>$-2.47^4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>$-21.23^9$</td>
<td>$1.75^5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age$^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$0.4^3$</td>
</tr>
<tr>
<td>N(t)</td>
<td>$0.44^5$</td>
<td>$-2.06^4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$1.02^4$</td>
</tr>
<tr>
<td>N$^2$(t)</td>
<td>$-0.59^9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>$-0.26^1$</td>
<td>$3.69^1$</td>
<td>$-1.56^1$</td>
<td>$-86.2^1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$-0.36^3$</td>
</tr>
<tr>
<td>TR</td>
<td>$-0.93^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$-0.93^1$</td>
</tr>
<tr>
<td>TR$^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$0.48^1$</td>
</tr>
<tr>
<td>Bond</td>
<td>$0.13^7$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FR$^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR</td>
<td>$0.13^4$</td>
<td>$-0.21^3$</td>
<td>$0.46^2$</td>
<td>$0.37^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CR$^2$</td>
<td>$0.16^2$</td>
<td>$-23.88^6$</td>
<td></td>
<td></td>
<td>$-2.61^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIns$_{t-6}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R$_{t-3}^3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lead$_{t-3}$</td>
<td>$-0.16^3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

188
mean-squared errors (MSE) as criteria on a validation data set. The MSE is defined as

$$MSE = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \left( D_i(t) - E[D_i(t)] \right)^2 / n_t, \quad (8.13)$$

where $n_t$ is the total number of obligors at risk at time $t$, i.e. $n_t = \sum_{k=1}^{K} \sum_{j=1}^{K} n_{j,k}$.

We compare the predicted delinquency one month forward with the realized delinquency. The models we compare here are only suitable for predicting the delinquency one month ahead in time. Should we want to predict more than one step ahead and the prediction model is dependent on time-dependent exogenous covariates, some form of prediction or expert opinion should be used to obtain the predicted covariate values.

In order to compute the MSE in (8.13), we need to know the expected value $E[D_i(t)]$ for all three models. The expected value of the CCPM model is

$$E[D_i(t)]$$

$$= E[\Delta N_i(t)] - 1 \{ D_i(t-1) > 0, \Delta N_i(t) = 0 \} \Delta B_i(t)$$

$$= E[\Delta N_i(t)] - 1 \{ D_i(t-1) > 0 \} E[1 \{ \Delta N_i(t) = 0 \} \Delta B_i(t)]$$

$$= E[\Delta N_i(t)] - 1 \{ D_i(t-1) > 0 \} P(\Delta N_i(t) = 0) E[\Delta B_i(t)]$$

$$= p^* - 1 \{ D_i(t-1) > 0 \} (1 - p^*) \left( \frac{(D_i(t-1) + 1) (1 - \tau^*)^{D_i(t-1) + 1}}{(1 - \tau^*)^{D_i(t-1) + 1} - 1} - \frac{\tau^* - 1}{\tau^*} \right),$$

where

$$p^* = \left( 1 + e^{-U(t)\beta} \right)^{-1}$$

and

$$\tau^* = \left( 1 + e^{-V(t)\alpha} \right)^{-1}.$$
The expected values of the stationary transition-matrix approach is

$$E[D_i(t) | D_i(t-1)] = \sum_{k=0}^{K} k \hat{p}_{D_i(t-1),k}$$

with the estimated transition probabilities $\hat{p}_{j,k}$ summarized in Table 8.2.

Lastly the modified GA approach has expected value

$$E[D_i(t) | D_i(t-1)] = \sum_{k=0}^{K} k \tilde{p}_{D_i(t-1),k},$$

with the estimated transition probabilities $\tilde{p}_{D_i(t-1),k}$ given by equation (8.12) and the parameter estimates summarized in Table 8.3 and Table 8.4.

We also added a model that we call the naïve prediction model. Here the predicted delinquency rate in the next month is simply the current delinquency rate. The mean-squared errors for these four models are summarized in Table 8.5.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve prediction</td>
<td>0.22924</td>
</tr>
<tr>
<td>Stationary Transition Matrix</td>
<td>0.21923</td>
</tr>
<tr>
<td>CCPM</td>
<td>0.21405</td>
</tr>
<tr>
<td>GA dynamic transition-matrix</td>
<td>0.21218</td>
</tr>
</tbody>
</table>

We see that the GA model has the smallest MSE (and so the best prediction accuracy). As expected, the naïve model fares the worst. Interestingly, the CCPM model does not do better than the GA model, but it does better than the stationary approach. To conclude we recommend that the modified GA approach be used to predict delinquencies as its prediction power seems to be the best. The CCPM model is better suited when the dimension of the transition-matrix is very high and delinquency levels are not capped at seven. Then the CCPM model would have fewer parameters and will be easier to estimate. The GA dynamic-transition-matrix approach is more intuitive than the CCPM approach. We recommend that further research be done on deriving a method of comparing the prediction power of different models.
8.4 Appendix

8.4.1 The truncated geometric distribution

We explore a discrete truncated distribution that is used to model repayments. Recall that a single payment can be seen as the payment of the normal installment that is due, but additional payments can be seen as an attempt to ‘recover’ or lower the current positive delinquency level. If the distribution of $\Delta B_i(t)$ were to be a right-truncated geometric distribution with support on the set $\{0, 1, 2, ..., D_i(t-1)\}$ and parameter $0 < \tau \leq 1$ then $\Delta B_i(t)$ would have probability mass function

$$P(\Delta B_i(t) = k|D_i(t-1)) = \frac{\tau (1 - \tau)^k}{1 - (1 - \tau)^D_i(t-1)+1}, \quad D_i(t-1) > 0.$$ 

The right-truncated geometric distribution is a member of the exponential family of distributions. The canonical link is $\theta = \log(1 - \tau)$, but since this link function does not transform $\tau$ to be between zero and one, we use the logit link function. Other link functions such as the log-log, probit, and complimentary log-log functions may also be used. The truncated geometric distribution has expected value

$$E[\Delta B_i(t)] = \frac{(D_i(t-1) + 1)(1 - \tau)^D_i(t-1)+1}{(1 - \tau)^D_i(t-1)+1 - 1} - \frac{\tau - 1}{\tau}$$

and variance

$$Var[\Delta B_i(t)] = \frac{(\tau - 1)^2}{\tau^2} - \frac{(\tau - 1)}{\tau} - \frac{(D_i(t-1) + 1)(\tau - 1)^D_i(t-1)+1}{(1 - \tau)^D_i(t-1)+1 - 1}$$

$$\quad - \frac{(D_i(t-1) + 1)^2(\tau - 1)^2(1 - \tau)^{2k}}{(1 - \tau)^D_i(t-1)+1 - 1}^2$$

$$\quad + \frac{D_i(t-1)(D_i(t-1) + 1)(\tau - 1)^2(1 - \tau)^D_i(t-1)+1}{(1 - \tau)^D_i(t-1)+1 - 1}.$$
8.4.2 Parameter estimation of the truncated geometric distribution

The score vector has dimension \((n_V \times 1)\) and has \(h^{th}\) element

\[
\frac{\partial \log \ell_i(t)}{\partial \alpha_h} = \frac{1}{1} \sum_{v=1}^{t} \sum_{i=1}^{n} Y_v^*(v-1) V_{i,h}(v-1)
\]

\[
\cdot \left[ 1 + e^{\alpha^T V_i(v-1) (D_i(t-1) - \Delta b_i(v) + 1)} + e^{\alpha^T V_i(v-1)} \right]^{-1}
\]

\[
- \left( 1 + e^{\alpha^T V_i(v-1) (D_i(t-1) - \Delta b_i(v) + 1)} \right) \left( e^{\alpha^T V_i(v-1)} + 1 - e^{\alpha^T V_i(v-1) + 1} - D_i(t-1) \right)
\]

\[
+ \left( 1 + e^{\alpha^T V_i(v-1) (D_i(t-1) - \Delta b_i(v) + 1)} \right) \left( e^{\alpha^T V_i(v-1)} + 1 - e^{\alpha^T V_i(v-1) + 1} - D_i(t-1) \right)
\]

\[
+ \left( 1 + e^{\alpha^T V_i(v-1) (D_i(t-1) - \Delta b_i(v) + 1)} \right) \left( e^{\alpha^T V_i(v-1)} + 1 - e^{\alpha^T V_i(v-1) + 1} - D_i(t-1) \right)
\]

for \(h = 1, ..., n_V\) with \(Y_v^*(v-1) = 1\) \(\{D_i(t-1) > 0, Y_i(t) = 1, N_i(t) = 0\}\). The Fisher information matrix is

\[
H = \left( -\frac{\partial^2 \log \ell_i(t)}{\partial \alpha \partial \alpha} \right)
\]

with the second derivative of the log-likelihood with respect to \(\alpha\) element \((j, h)\)

\[
\frac{\partial^2 \log \ell_i(t)}{\partial \alpha_j \partial \alpha_h} = \sum_{v=1}^{t} \sum_{i=1}^{n} Y_v^*(v-1) \left( V_{i,j}(v-1) V_{i,h}(v-1) \left[ x_{v,i} + y_{v,i} + z_{v,i} \right] \right)
\]

\[
\sum_{v=1}^{t} \sum_{i=1}^{n} Y_v^*(v-1) \left( V_{i,j}(v-1) V_{i,h}(v-1) \left[ x_{v,i} + y_{v,i} + z_{v,i} \right] \right)
\]

192
where

\[ x_{v,i} = -e^{\alpha^T V_i(v-1)} (x^a_{v,i} - x^b_{v,i}) , \quad \text{where} \]

\[ x^a_{v,i} = \Delta b_i(v) + e^{2\alpha^T V_i(v-1)} + \Delta b_i(v)(e^{\alpha^T V_i(v-1)} + 1)^{-2D_i(t-1)} \quad \text{and} \]

\[ x^b_{v,i} = D_i(t-1)(e^{\alpha^T V_i(v-1)} + 1)^{-2D_i(t-1)} + \Delta b_i(v)e^{2\alpha^T V_i(v-1)} + 1, \]

\[ y_{v,i} = (e^{\alpha^T V_i(v-1)} y^a_{v,i} + y^b_{v,i})(e^{\alpha^T V_i(v-1)} + 1)^{-D_i(t-1)} , \quad \text{where} \]

\[ y^a_{v,i} = 2\Delta b_i(v) - D_i(t - 1) + e^{2\alpha^T V_i(v-1)} + 2D_i(t-1)e^{2\alpha^T V_i(v-1)} \]

\[ + D_i(t - 1)^2 e^{2\alpha^T V_i(v-1)} + 1 \quad \text{and} \]

\[ y^b_{v,i} = V^2 e^{2\alpha^T V_i(v-1)}(D_i(t-1)^2 + D_i(t+1) + 2\Delta b_i(v) + 2), \]

\[ z_{v,i} = -e^{2\alpha^T V_i(v-1)}(2\Delta b_i(v) + 2)), \]

\[ k_{v,i} = ((e^{\alpha^T V_i(v-1)} + 1)^2(e^{\alpha^T V_i(v-1)} - (e^{\alpha^T V_i(v-1)} + 1)^{-D_i(t-1)} + 1)^2). \]

for \( h, j = 1, ..., n_V \). Having found the Fisher information matrix and the score vector, the Newton-Raphson procedure can be applied. The standard errors of the parameter estimates are given by

\[ SE(\hat{a}) = \sqrt{\text{diag}(H^{-1})}. \]
Chapter 9

Summary and suggestions for further research

An important component of credit-risk management is the prediction of a default event. A well-known and widely used approach to predict defaults is the structural approach of Merton (1974). This approach imposed some structure on the default event, namely when asset levels are less than liability levels. The structural approach is mostly used for corporate obligors and forms the foundation of the Basel accord’s capital requirements. The usage of this Basel-accord model is widely accepted in consumer credit risk, even though it is known that the asset and liability levels of consumers are not observable. More recently, Artzner and Delbaen (1995) as well as Jarrow and Turnbull (1995) developed models that impose no structure on the default event. These reduced-form models are all counting processes and seem to fit the consumer-risk framework better than the structural approach. Here no assumptions regarding the observability of the asset or liability levels need to be made. Counting processes are continuous in nature and we need to discretize the counting process techniques to apply to our regularly observed discrete data. Traditional survival analysis is not well suited to the analysis of credit-risk data, since all observations need to be aligned to some fixed starting time. Survival analysis is concerned with the time until the once-off occurrence of an event. In consumer credit risk we are more concerned with multiple defaults which must also be catered for. Survival analysis is also concerned with event time, but since our data is dependent on macro-economic conditions we prefer to use calendar time. Most credit-risk models only account
for a single event, namely default, after which it is assumed that the obligor is no longer on the bank’s books. We found a model that is well suited to modelling the occurrence of multiple events in discrete time, with staggered entry. Although we only focussed on the modelling and prediction of default events, further research can include other so-called competing risk events such as early closure and write-off.

We implemented the discrete calendar-time staggered-entry reduced-form model of Arjas and Haara (1987) to model both a default event as well as a multiple-failure event for a mortgage loan portfolio. We found that the current delinquency exhibits high causality for default. Hence, we developed a marked (size of repayment) point-process model to predict the repayment(s) made by obligors. Combining this marked-point process with the failure counting process we are able to predict delinquency levels. We also apply the insights gained from Arjas and Haara (1987) to a transition-matrix approach by Grimshaw and Alexander (2010) to predict delinquency levels.

From the outset we ensured that all models are dependent on idiosyncratic as well as systematic covariates. The dependency on systematic covariates enables stress-testing and we have shown how to predict default and delinquency levels one month forward. Further research will be to use this model for stress-testing. This model can be used to implement conditional stress-testing or simulation based stress-testing. Conditional stress-testing is when macro-economic scenarios are given by some expert. Stress-testing by simulation is used when there is some model to predict the relationship between macro-economic covariates and the individual simulated paths are used as inputs for the model. This thesis did not explore either of these stress-testing techniques, but instead provided a model that we hope is useful in predicting credit events.

We tested the goodness of fit for the models by using marginal-model plots that we modified to have a confidence band, as well as a relative measure of the goodness of fit, namely the bootstrapped mean-squared error. We also developed composite marginal-model plots to aid in the heuristic diagnostic testing of longitudinal data. We used a training and validation data sample to estimate the prediction accuracy. Our delinquency prediction model seems to have good predictability. Should these approaches be used on another consumer portfolio with the addition of latent, behavioural and application scorecard data, the fit could be even more favourable. We had limited computing power,
so drawing a large training sample may also lead to better results.

We did not include latent covariates and this option should be considered as an extension of the work done here. McNeil and Wendin (2006) state that: "Unfortunately observed variables as proxies for systematic risk are seldom completely satisfactory". We feel that latent covariates, although reducing interpretability, might improve the prediction accuracy of the model. Further extension or research can be done on, to name a few: (1) the prediction of the number of newly opened accounts; (2) the inclusion of absorbing states (write-off and closure) in the transition matrix approach (equivalently the inclusion of censored events in the AH model); (3) the derivation of alternative statistics for measuring the goodness of fit for longitudinal data; (4) alternative methods of comparing the predictive efficiency of various models; and (5) alternative methods of selecting covariates.

It is our hope that this thesis makes a significant contribution to the literature on consumer credit-risk management by advocating, among other things, the use of Arjas and Haara-type discrete real-time models to predict credit events.
Chapter 10

List of abbreviations, covariates, notation and definitions

Abbreviations:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH</td>
<td>Arjas and Haara (1987)</td>
</tr>
<tr>
<td>CCPM</td>
<td>Combined counting-process model</td>
</tr>
<tr>
<td>CMM</td>
<td>Composite marginal-model plot</td>
</tr>
<tr>
<td>DTYG</td>
<td>Ding, Tian, Yu, and Guo (2012)</td>
</tr>
<tr>
<td>ES</td>
<td>Expected shortfall</td>
</tr>
<tr>
<td>GA</td>
<td>Grimshaw and Alexander (2010)</td>
</tr>
<tr>
<td>GLM</td>
<td>Generalized linear models</td>
</tr>
<tr>
<td>GLMM</td>
<td>Generalized linear mixed models</td>
</tr>
<tr>
<td>ILR</td>
<td>Independent logistic regression</td>
</tr>
<tr>
<td>IRLS</td>
<td>Iteratively reweighted least squares</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean-squared error</td>
</tr>
<tr>
<td>SME</td>
<td>Small and medium enterprises</td>
</tr>
<tr>
<td>VaR</td>
<td>Value at risk</td>
</tr>
</tbody>
</table>
Covariates:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLTV</td>
<td>The loan-to-value ratio at origination of the loan</td>
</tr>
<tr>
<td>AD(t)</td>
<td>Average long-term delinquency rate</td>
</tr>
<tr>
<td>Age(t)</td>
<td>The age of the loan</td>
</tr>
<tr>
<td>Bond</td>
<td>The average of the total loan granted, purchase price, security value and the total bond amount</td>
</tr>
<tr>
<td>CF(t)</td>
<td>Cumulative number of failure events in the lifetime of the loan</td>
</tr>
<tr>
<td>CLTV(t)</td>
<td>Current loan-to-value</td>
</tr>
<tr>
<td>CR(t)</td>
<td>Cumulative number of recovery events in the lifetime of the loan</td>
</tr>
<tr>
<td>D(t)</td>
<td>The current delinquency level</td>
</tr>
<tr>
<td>GDP(t)</td>
<td>The gross domestic product</td>
</tr>
<tr>
<td>Lead(t)</td>
<td>Leading business-cycle indicator</td>
</tr>
<tr>
<td>Liq(t)</td>
<td>Liquidations</td>
</tr>
<tr>
<td>M3(t)</td>
<td>M3 money supply</td>
</tr>
<tr>
<td>MD(t)</td>
<td>Maximum delinquency rate during the lifetime of the loan</td>
</tr>
<tr>
<td>PDisInc(t)</td>
<td>Personal disposable income</td>
</tr>
<tr>
<td>PPI(t)</td>
<td>Property price index</td>
</tr>
<tr>
<td>Prime(t)</td>
<td>Interest rates</td>
</tr>
<tr>
<td>PSav(t)</td>
<td>Personal savings</td>
</tr>
<tr>
<td>R$(t)</td>
<td>Rand-dollar exchange rate</td>
</tr>
<tr>
<td>REAM</td>
<td>The REAM value: An ordinal classification of the quality of the neighbourhood</td>
</tr>
<tr>
<td>TD(t)</td>
<td>The total months that the account was in arrears in the last six months</td>
</tr>
<tr>
<td>Term</td>
<td>The full period over which the loan is to be repaid</td>
</tr>
<tr>
<td>TF(t)</td>
<td>Time since last failure</td>
</tr>
<tr>
<td>TIns(t)</td>
<td>Total number of insolvencies</td>
</tr>
<tr>
<td>TP(t)</td>
<td>The total number of repayments made during the last six months</td>
</tr>
<tr>
<td>TR(t)</td>
<td>Time since last recovery</td>
</tr>
</tbody>
</table>
Notation:

$1 \{ A \}$ The indicator function

:= Defined as

$\lfloor x \rfloor$ The largest integer less than or equal to $x$

$\text{trace}(X)$ The trace of matrix $X$

$\rightarrow$ Tends to

$N(a, b)$ Normally distributed with mean $a$ and variance $b$

$\chi^2(a)$ Chi-squared distributed with degrees of freedom $a$

$X^\top$ Transpose of matrix $X$

$X^{-1}$ Inverse of matrix $X$

$\sim$ Distributed as

$I$ Identity matrix

$\text{diag}(X_i)$ The $i^{th}$ diagonal element of matrix $X$

$\lim$ The limit

$\text{card}[X]$ The cardinality of the set $X$

$n!$ The factorial

$J_n$ A column one vector with $n$ elements

$\partial l / \partial \beta$ The partial derivative of $l$ with respect to $\beta$
List of definitions:

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delinquency</td>
<td>Delinquency is the state that exists when an obligor is in arrears.</td>
</tr>
<tr>
<td>Failure</td>
<td>A failure occurs when there is an increase in an obligor’s level of delinquency.</td>
</tr>
<tr>
<td>Default</td>
<td>A default occurs when an obligor is three months delinquent.</td>
</tr>
<tr>
<td>Recovery</td>
<td>A recovery occurs when there is a decrease in the delinquency level of an obligor.</td>
</tr>
<tr>
<td>Write-off</td>
<td>A write-off occurs when an obligor’s loan account is written off the bank’s balance sheet.</td>
</tr>
<tr>
<td>Credit event</td>
<td>A credit event is either a failure, recovery or a write-off.</td>
</tr>
</tbody>
</table>

Adapted process

A stochastic process \( X(t), t \geq 0 \) is said to be *adapted* to a filtration \( \mathcal{F}_t \) if \( X(t) \) is a function of \( \mathcal{F}_t \) for every \( t \). Measure theoretically, \( X(t) \) is adapted to \( \mathcal{F}_t \) if \( X(t) \) is \( \mathcal{F}_t \)-measurable for each \( t \).

Predictable process

A stochastic process \( X(t), t \geq 0 \) is said to be *predictable* with respect to a filtration \( \mathcal{F}_t \) if \( X(t) \) is a function of \( \mathcal{F}_{t-} \) for every \( t \). Measure theoretically, \( X(t) \) is predictable if \( X(t) \) is \( \mathcal{F}_{t-} \)-measurable for each \( t \).
Bibliography


203


