Journal of Empirical Finance 16 (2009) 457–465

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Estimation of default probabilities using incomplete contracts data

J.M.C. Santos Silva [a,c,*], J.M.R. Murteira [b,c,*]

[a] Department of Economics, University of Essex, UK
[b] Faculdade de Economia, Universidade de Coimbra, Portugal
[c] CEMAPRE, Portugal

ARTICLE INFO

Article history:
Received 4 October 2007
Received in revised form 11 August 2008
Accepted 11 November 2008
Available online 24 November 2008

JEL classification code:
C21
C51
G21

Keywords:
Beta-binomial distribution
Credit scoring
Population drift

ABSTRACT

This paper develops a count data model for credit scoring which allows the estimation of default probabilities using incomplete contracts data. The main advantage of the proposed approach is that it permits a more efficient use of the data, including that for the most recent clients. Moreover, because the probability of default is specified as a function of the age of the contract, the model provides some information on the timing of the defaults. The model is based on the beta-binomial distribution, which is found to be particularly adequate for this purpose. A well-known dataset on personal loans is used to illustrate the application of the proposed model.

1. Introduction

Models for credit scoring are widely used in practice and raise a number of interesting and challenging research questions. Therefore, it is not surprising to find that they have been the subject of a considerable literature (see, among many others, Altman et al., 1981; Maddala, 1996; Hand and Henley, 1997; Hand and Jacka, 1998; Thomas, 2000; Thomas et al., 2002, and the references therein).

Consider a lending institution, hereinafter referred to as the bank, that wants to use information on the characteristics and repayment behavior of its clients to estimate the probability of a prospective borrower to default on a loan.[1] The nature of the statistical methods to use in the estimation of this model will depend on the type of loan being considered. Indeed, different types of loans generate data with different characteristics, and that is critical for the choice of the statistical methodology to use.

In this paper we restrict our attention to a particular, yet important, type of loan. Specifically, we develop a model for the probability of default for the case in which the loan is to be repaid in a number of regular installments.[2] Scoring models for this type of loan are typically estimated using only data on contracts that have reached their maturity. However, models constructed in this way waste the information on clients who are currently repaying their loans. This is inappropriate, not only because it is an inefficient use of the data, but also because the resulting model may be affected by *population drift* problems, caused by changes in

* Corresponding authors. Santos Silva is to be contacted at Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. Murteira, Faculdade de Economia, Universidade de Coimbra, Av. Dias da Silva, 165, 3004-512 Coimbra, Portugal.
E-mail addresses: jmcss@essex.ac.uk (J.M.C. Santos Silva), jmurt@fe.uc.pt (J.M.R. Murteira).

[1] In this paper only the probability of default is modelled. Although this is only a part of the optimization problem faced by the lending institution, it is of critical importance both for its profitability and for the households' welfare (see Carling et al., 2001).

[2] Therefore, the model developed here may be inappropriate to describe defaults when the loans take the form of overdrafts, or are granted through credit cards. Thomas (2000) surveys different techniques that are useful in constructing scoring models for these other types of loans.

0927-5398/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jempfin.2008.11.003

the distribution of the characteristics of the clients (Kelly et al., 1999). To mitigate these problems, in practice, current clients are often included in the sample and are classified according to their present status. However, this procedure will inevitably induce a degree of misclassification because some clients currently classified as non-defaulters may eventually default before the end of their contracts.

This paper shows that, using an appropriate count data model for the number of payments missed by the borrowers, it is possible to use the data on all contracts, including the more recent ones, to estimate the conditional probability of a client becoming a defaulter. This is particularly important in the case of long term contracts, e.g., mortgages, because in these cases the characteristics and repayment behavior of clients with completed contracts may have little to do with those of the prospective borrowers. Moreover, for smaller and newer banks, the number of observations on clients with completed long term contracts may be very small.

The proposed model also has some additional advantages. First, by modelling the actual number of missed payments rather than just an indicator of default, the model explores more of the available information. Second, it allows the probability of default to depend on the age of the contract, thereby providing some information on the way the probability of default varies in time. The timing of defaults has been previously addressed in the literature (e.g., Roszbach, 2004; Duffie et al., 2007), but the particular characteristic of the model developed here is that it allows the researcher to estimate the probability of default for different time horizons, without actually using any duration data on the timing of defaults. Finally, the proposed modelling strategy is interesting because it can be adapted to a variety of circumstances and permits the test of a number of interesting hypothesis, like the possible change in repayment behavior after the client is considered a defaulter.

An issue that has to be addressed in the construction of credit scoring models is the potential sample selection problem caused by the fact that the bank only has information on clients to whom it has decided to grant a loan. This situation is problematic if the decision to accept or refuse the credit applications is made using information on the clients that is not available for the construction of the credit scoring model. In this case the sample is endogenously stratified and there is not much that can be done to solve the problem without relying on very strong assumptions (see Hand and Henley, 1993). On the contrary, if all the information used to decide about the credit applications is available for the construction of the credit scoring model, standard inference methods can be used because in this case the sample is exogenously stratified (see Pudney, 1989; Wooldridge, 1999). This more favorable situation is the one considered here.

The remainder of the paper is organized as follows. The next section introduces a count data model that allows the estimation of default probabilities using data on incomplete contracts. Section 3 presents, purely for illustrative purposes, an application of the proposed model using a well-known dataset on personal loans, and Section 4 concludes the paper.

2. An appropriate count data model

Consider a loan that has to be repaid in N regular installments and let n denote the present age of the contract, measured by the number of installments that should have been paid since the contract began. Furthermore, let Y be the number of payments so far missed by a client. Therefore, we have that $0 \leq Y \leq n \leq N$.

The purpose of the scoring model is to estimate the probability that Y will be larger than the maximum number of payments that a client is allowed to miss without being considered a defaulter, denoted by l. Obviously, this probability is zero for any $n \leq l$.

For clients whose loans have reached their maturity date, i.e., $n = N$, it is possible to construct a binary variable indicating whether or not the client defaulted. Credit scoring models are typically estimated by using appropriate statistical methods to model these indicators. The main drawback of this approach is that is wastes information, non only on the current clients whose final status is yet unknown, but also on the actual number of payments missed by the former clients.

Rather than just modelling the default indicator, the alternative approach we follow here is to model the count variable Y, taking into account that this variate is bounded between 0 and n. These bounds on Y have implications for the type of count data model to use. Indeed, the distributions more often used in applied work, e.g., Poisson and negative binomial, are not suitable in this context because they assume that the counts have no upper bound. In order to account for the peculiar characteristics of Y, we use a beta-binomial model, which explicitly imposes that $Y \leq n$. The beta-binomial regression was first used by Heckman and Willis (1977) in a very different context, but has seen little use in practice.

2.1. Beta-binomial regression

Suppose that, besides Y, N and n, the bank observes a set x of characteristics of the contract and of its clients.[3] The objective is then to estimate the probability that Y will cross the threshold above which the client will be classified as a defaulter, given N, n and x.

In order to take explicitly into account the upper bound on the value of Y, the model developed here has as a starting point the binomial distribution characterized by

$$P(Y = y|n) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \tag{1}$$

[3] Depending on the nature of the problem, macroeconomic indicators may also be useful predictors of default (see, e.g., Bellotti and Crook, 2007).