Journal of Banking & Finance 37 (2013) 1040–1045

Contents lists available at SciVerse ScienceDirect

Journal of Banking & Finance

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

Reject inference in consumer credit scoring with nonignorable missing data ☆

Michael Bücker, Maarten van Kampen, Walter Krämer *

Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany

ARTICLE INFO

Article history:
Received 30 June 2011
Accepted 8 November 2012
Available online 17 November 2012

JEL classification:
C25
C58
G21

Keywords:
Credit scoring
Reject inference
Logistic regression

ABSTRACT

We generalize an empirical likelihood approach to deal with missing data to a model of consumer credit scoring. An application to recent consumer credit data shows that our procedure yields parameter estimates which are significantly different (both statistically and economically) from the case where customers who were refused credit are ignored. This has obvious implications for commercial banks as it shows that refused customers should not be ignored when developing scorecards for the retail business. We also show that forecasts of defaults derived from the method proposed in this paper improve upon the standard ones when refused customers do not enter the estimation data set.

1. Introduction

Statistical models for predicting defaults in the consumer credit industry and elsewhere suffer from the non-availability of default information for customers who were denied credit in the first place (Hand and Henley, 1993; Crook and Banasik, 2004, among many others). This is known as the reject-inference-problem; it affects the estimation of the model parameters in the same way as the non-availability of high-probability rainy days would affect the parameter estimates of a meteorological model for predicting rain.

This non-availability does not matter if observations are missing at random (MAR) in the sense of Rubin (1976). Missing at random means that the probability of default, given all the exogenous variables of the model, is the same whether an applicant is granted a credit or not (or in the meteorological example: if the probability of rain, given a set of relevant regressors, is the same for days observed and unobserved). In applications, this can reasonably be assumed if creditors base their decision on the same statistical model (or a preliminary version thereof) which is to be estimated.

However, such procedures are illegal in many countries. In Germany, for instance, the federal data privacy act explicitly forbids banks to grant consumer credit solely on the basis of a statistical model – there must be some human judgement involved as well (for instance to determine whether applicants conceal relevant information, see Fees et al. (2011)). This means that loan officers have both the right and the duty to override a statistical model if they think this is warranted by extra information. Among applicants with otherwise identical sets of explanatory variables, some may therefore be granted a credit and some may not. Or technically speaking, the probability of being granted a credit, given the observed regressors, is not the same as the probability of being granted a credit, given the observed regressors *and* future default information. Whenever human judgement adds any additional information on future defaults, these probabilities will differ. This implies that data are missing not at random (MNAR) in the Rubin (1976) sense.

The present paper adds to previous approaches to take credit decision processes into account when estimating models of default (see e.g. Boyes et al. (1989) or Marshall et al. (2010)) by proposing a new approach to cope with this. It is based on Qin et al. (2002), who show how to reweight observations in the light of missing data, given a parametric model for the missings, using empirical likelihood (Owen, 2001). It compares favorably to other techniques that have been suggested in the literature to mitigate the effects of missing data in the credit scoring business in that we are able to analytically derive the limiting distribution of the resulting estimator. Most prominent among established methods are extrapolation, reweighting, or simultaneous bivariate probit modeling of acceptance and default along the lines of Boyes et al. (1989). Extrapolation means assigning a default status also to the rejects, based on the same model that is fitted to the accepted cases only, and then

☆ Financial support from Deutsche Forschungsgemeinschaft (SFB 823) is gratefully acknowledged. We thank Jing Qin, Matthias Arnold and an anonymous referee for helpful discussions and comments.

* Corresponding author. Tel.: +49 231 755 3125; fax: +49 231 755 5284.
E-mail addresses: buecker@statistik.tu-dortmund.de (M. Bücker), maarten.vankampen@tu-dortmund.de (M. van Kampen), walterk@statistik.tu-dortmund.de (W. Krämer).

0378-4266/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jbankfin.2012.11.002

reestimating the model. Reweighting is based on the preliminary estimation of a model for acceptance, using both accepts and rejects, and a subsequent redistribution of all cases into classes with varying percentages of defaults. All accepts are then reweighted according to the proportion of defaults in their respective class. See Crook and Banasik (2004) for a survey and a discussion of the pros and cons of the various approaches.

We do not want to add to this comparison literature here, as this would greatly expand the scope of our paper. Rather, we would like to introduce a new competitor and derive and illustrate its properties. In particular, in the context of a logistic regression model for defaults, we suggest an alternative reweighting scheme and show analytically that it delivers consistent and asymptotically normal parameter estimates even when credit decisions and defaults are still correlated, given all regressors. We also investigate the relationship between the severity of the missing data problem and the improvement provided by our new estimator and show that there is a monotonous relationship between the two.

When applied to a recent data set of almost 10,000 individuals requesting credit with a major German bank, our approach yields parameter estimates which are significantly different from standard ones both in a statistical and in an economic sense. This shows that ignoring the missing data problem has the potential to mislead credit granting decisions in practice and is therefore also relevant for practitioners: Whenever the credit granting process is a mixture of formal scoring and informal judgment by credit officers, the parameter estimates of the scoring model may be biased and the default predictions derived from them may be inaccurate, with obvious implications for the profit of banks and financial institutions. We also show by Monte Carlo experiments that default forecasts derived from our new estimator indeed improve upon default forecasts obtained from standard Maximum Likelihood estimators of the model parameters.

2. An alternative way of reweighting observations in the presence of rejects

We consider N applicants for a credit, n of whom are granted a credit and $N–n$ are not. Default is coded by a dichotomous variable Y, where $Y_i = 0$ in case of default and $Y_i = 1$ in case of no default. We assume that Y_i depends on a set of k regressors which we collect together in a $(k \times 1)$ vector $\boldsymbol{X}$. Potential errors of measurement concerning $\boldsymbol{X}$ are neglected; see however Fees et al. (2011). We also assume that the dependence of Y on $\boldsymbol{X}$ can be described by a logistic regression model

$$P(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{\beta}) := \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i})} \quad (1)$$

$(i = 1, \ldots, N)$. This is still by far the most popular statistical model entertained in this context, see e.g. Thomas (2000), Jacobson and Roszbach (2003) or Crook and Banasik (2004). (For a rather different approach, see Khandani et al. (2010)). The primary difference to a conventional logistic regression is that not all N outcomes are observed. Let $R_i = 1$ if credit is granted and $R_i = 0$ if credit is denied. Without loss of generality, we assume that $R_i = 1$ for the first n data points and $R_i = 0$ for the remaining ones.

From a statistical point of view, the problem is that ignoring all data beyond n produces inconsistent ML-estimates for the model (1) whenever data are missing not at random in the sense that

$$P(R = 1|\boldsymbol{X}, Y) \neq P(R = 1|\boldsymbol{X}). \quad (2)$$

We now show, following Qin et al. (2002), how this inconsistency can be removed. To that purpose, let $F(y, \boldsymbol{x})$ be the joint distribution function of $(Y, \boldsymbol{X})$ (no parametric model is needed for this), let

$$w(y, \boldsymbol{x}, \boldsymbol{\theta}) := P(R = 1|Y, \boldsymbol{X}, \boldsymbol{\theta})$$

be some parametric model for observability (sometimes also called accept-reject-model; see Crook and Banasik (2004)), let $W{:=}P(R = 1)$, and consider the following semiparametric likelihood for $\boldsymbol{\theta}$, W, and F:

$$L_n(\boldsymbol{\theta}, W, F) = \left[\prod_{i=1}^{n} w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) dF(y_i, \boldsymbol{x}_i)\right] \cdot (1 - W)^{N-n}. \quad (3)$$

This function is maximized under the constraints

$$p_i \geqslant 0, \;\; \sum_{i=1}^{n} p_i = 1, \;\; \sum_{i=1}^{n} p_i [\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}] = 0, \text{and} \;\; \sum_{i=1}^{n} p_i [w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W] = 0, \quad (4)$$

where $p_i = dF(y_i, \boldsymbol{x}_i) = F(y_i, \boldsymbol{x}_i) - F_{-}(y_i, \boldsymbol{x}_i)$, i.e. p_i is the increase in the joint distribution function at $(y_i, \boldsymbol{x}_i)$ and $\boldsymbol{\mu}_{\boldsymbol{X}}$ is either the known expectation or the empirical mean of $\boldsymbol{X}$. By introducing Lagrange multipliers and profiling for all values of p_i, it is seen that

$$p_i = \frac{1}{n\left[1 + \boldsymbol{\lambda}_1^{\top}(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \lambda_2 (w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W)\right]},$$

where $\boldsymbol{\lambda}_1$ and λ_2 are Lagrange multipliers. Substituting p_i into (3) results in a profile likelihood that can be maximized numerically. (Qin et al. (2002), Theorem 1) show that under mild regularity conditions, the resulting empirical likelihood estimates for $\boldsymbol{\theta}$ and W are consistent and asymptotically normal.

Here we are interested in the plug-in estimate $\hat{p}_i$ of p_i in order to reweight the likelihood derived from (1). Doing this, we obtain

$$L_n^{\star}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \hat{p}_i f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}), \quad (5)$$

where $f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = [P(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{\beta})]^{y_i} \cdot [1 - P(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{\beta})]^{1-y_i}$. The full likelihood function is then given by

$$L_n^{\star}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{1}{n\left[1 + \hat{\boldsymbol{\lambda}}_1^{\top}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{X}}) + \hat{\lambda}_2 (w(y_i, \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}) - \widehat{W})\right]} \cdot \left[\frac{\exp(\boldsymbol{x}^{\top}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^{\top}\boldsymbol{\beta})}\right]^{y_i} \cdot \left[1 - \frac{\exp(\boldsymbol{x}^{\top}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^{\top}\boldsymbol{\beta})}\right]^{1-y_i}. \quad (6)$$

The conventional ML-estimator $\hat{\boldsymbol{\beta}}$ which ignores all missings is the solution to (6) without the weights $\hat{p}_i$. Our main theoretical result is that maximizing (6) yields a consistent and asymptotically normal estimator $\tilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ even in the case of (2), i.e. when missingness cannot be ignored.

Theorem 1. *Under mild regularity conditions to be specified in the appendix, the modified ML-estimator $\tilde{\boldsymbol{\beta}}$ is weakly consistent and*

$$\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}),$$

where β_0 denotes the true value of β.

The proof of this theorem and the description of the limiting covariance matrix $\boldsymbol{V}$ are in the appendix.

Table 1 provides some finite sample Monte Carlo evidence for $N = 10{,}000$, a common sample size in consumer credit scoring applications. We consider the case of a single regressor, i.e., $k = 1$, with values $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 4)$ and

$$P(Y_i = 1|X_i = x_i, \boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad (\boldsymbol{i} = 1 \ldots, N).$$

The observability of Y is governed by

$$w(y_i, x_i, \theta) = \frac{\exp(\theta_0 + \theta_1 y_i)}{1 + \exp(\theta_0 + \theta_1 y_i)} \quad (\boldsymbol{i} = 1 \ldots, N). \quad (7)$$