European Journal of Operational Research 249 (2016) 498–505

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Models and forecasts of credit card balance

Pak Shun Hon, Tony Bellotti*

Department of Mathematics, Imperial College London, London, UK

ARTICLE INFO

Article history:
Received 22 January 2014
Accepted 9 December 2014
Available online 19 December 2014

Keywords:
Credit cards
Balance estimation
Mixture model
Panel model

ABSTRACT

Credit card balance is an important factor in retail finance. In this article we consider multivariate models of credit card balance and use a real dataset of credit card data to test the forecasting performance of the models. Several models are considered in a cross-sectional regression context: ordinary least squares, two-stage and mixture regression. After that, we take advantage of the time series structure of the data and model credit card balance using a random effects panel model. The most important predictor variable is previous lagged balance, but other application and behavioural variables are also found to be important. Finally, we present an investigation of forecast accuracy on credit card balance 12 months ahead using each of the proposed models. The panel model is found to be the best model for forecasting credit card balance in terms of mean absolute error (MAE) and the two-stage regression model performs best in terms of root mean squared error (RMSE).

1. Introduction

Credit cards are an example of revolving retail exposures. They are the largest component of revolving retail credit in the United States (Qi, 2009). They account for 6.9 percent of bank assets and 16 percent of bank earnings in 2007 and this figure is rising (Terris, 2008). Unlike fixed term repayment loans, the outstanding balance on revolving credit is difficult to determine. However, it is important to be able to forecast credit card balance at an account level for several reasons. First, it allows for an accurate revenue prediction and hence estimates of expected profits. Second, it allows financial institutions to identify customer behaviour. In particular, customers who are building large debt and possibly exceed their credit limits can be identified, enabling management of that risk. Third, it allows for estimation of assets on book at granular level; e.g. by risk groups. Fourth, modelling balance will give insights into how to model Exposure at Default (EAD): one of the factors used in the internal ratings based approach to computing expected loss and capital requirements as part of the Basel II Capital Accord (Taplin, To, & Hee, 2007). This is relevant to our study since EAD is just outstanding balance at the time of default.

The literature on balance estimation is limited. There are several studies of EAD, but mainly for corporate loans. Most published studies use loan equivalent (LEQ), credit conversion factor (CCF) and exposure at default factor (EADF) as a proxy to model EAD (Jimenez, Lopez, & Saurina, 2009; Qi, 2009; Taplin et al., 2007). LEQ has the disadvantage of being undefined when a consumer has fully utilized the credit limit. CCF estimation does not include information on the credit line and EADF does not take information about outstanding balance into account. Taplin et al. (2007) have shown that modelling EAD via CCF would be inappropriate. For instance, the estimate of EAD would be unstable when the card balance is close to the credit limit. Also, for forecasting, modelling EAD directly may be more appropriate. In particular, LEQ and CCF is not well-defined for all observations (Leow & Crook, 2013). Estimation of profitability on retail loan portfolios has been published. However, outstanding balance is considered implicit (Finlay, 2010) or is known for fixed term repayment loans (Ma, Crook, & Ansell, 2010), rather than requiring an explicit statistical model. None of these studies consider modelling account balance, unconditional on default, for revolving credit. An unconditional estimate is required if we want revenue prediction or to optimise customer treatment based on future expectation of card usage and balance. EAD and unconditional balance are different. In particular, for the same account, we would expect EAD to be greater than unconditional balance, as bad debtors are likely to build up a higher balance under financial distress, and we would expect the predictors to be somewhat different. Therefore, it makes sense to have separate models for both factors. However, since unconditional balance is a generalization of EAD, it will provide insights into the modelling of EAD. Modelling approaches for EAD are also surveyed in Yang and Tkachenko (2012), including the use of mixture models and neural networks.

In this article, several alternative models of outstanding balance are proposed and a data set of UK credit cards is used to assess their relative performance. Between an OLS model, two-stage

* Corresponding author. Tel.: +44 20 759 48521.
E-mail address: a.bellotti@imperial.ac.uk (T. Bellotti).

http://dx.doi.org/10.1016/j.ejor.2014.12.014
0377-2217/

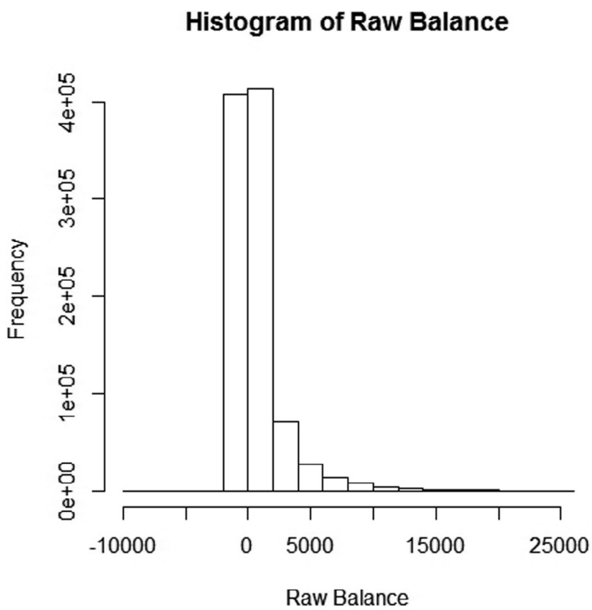

Fig. 1. Histogram of raw balance (GB pounds).

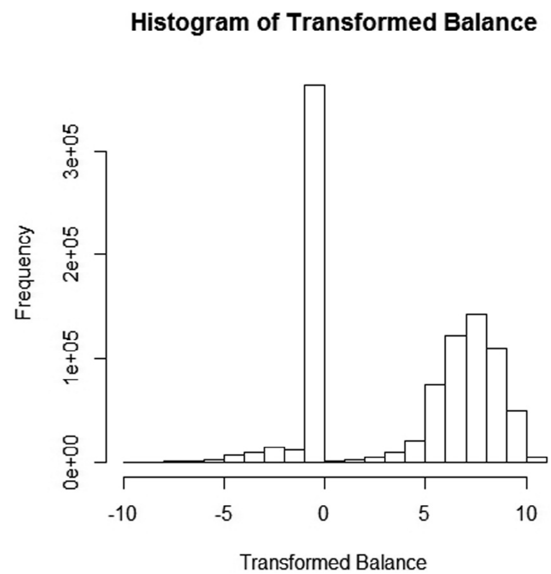

Fig. 2. Histogram of asinh transformed balance ($\theta = 1$).

model, mixture regression model and random effects panel model, the latter model performs best at forecasting balance in terms of mean absolute error (MAE). Section 1.1 discusses the general distribution of balance, Section 2 introduces the statistical models and evaluation method, Section 3 describes the data used in this study along with data processing, Section 4 gives results in terms of model estimates and forecasts and Section 5 gives some conclusions and discussion of possible future work.

1.1. Distribution of credit card balance

For this study, a positive value of credit card balance indicates the credit used and outstanding at the end of each accounting month. Credit card balance typically has a right-skewed distribution with lower frequencies of accounts having larger balances. Also, many values will be exactly zero, representing those individuals who either do not use the credit card for several months or who repay their balance in full. Additionally, it is possible to record a negative balance for account months when the credit card holder overpays or a refund is paid to an account. Fig. 1 shows the distribution of balance for the credit card data used in this study (this data is described in more detail in Section 3) and it illustrates these features. In particular, 35.3 percent of account records have a balance of exactly zero and 7.8 percent record a negative balance.

The right-skew suggests a log transformation. However, the presence of non-positive values makes this impossible. A feasible alternative is to transform the dependent variable using the inverse hyperbolic sine (asinh) transformation (Johnson, 1949). This is an efficient way to transform monetary data involving zero and negative values. It is defined as:

$$f(x) = \frac{\log(\theta x + \sqrt{\theta^2 x^2 + 1})}{\theta} = \frac{\sinh^{-1}(\theta x)}{\theta} \tag{1.1}$$

A nice property of this function is that as $x \to 0$, $f(x) \sim x$. In addition, it scales down large values just as the logarithmic transformation does: as $x \to \pm\infty$, $f(x) \sim \pm \log(2\theta x)/\theta$. Fig. 2 shows the distribution of the asinh transformed balance. The features of the credit card balance are more distinctive after the transformation is applied. In this study, $\theta = 1$ was used. Other values were considered, but did not give substantial improvement to visualise the distribution of balance.

2. Methodology

2.1. Statistical models

We consider alternative explanatory models for credit card balance. In particular, we consider Ordinary Least Squares (OLS) regression, two-stage regression models, mixture regression and panel regression models. Predictor variables available for regression will be static application variables, behavioural variables regarding usage of the credit card and previous balance. The following notation is used: let $y_{i,t}$ be the credit card balance, possibly after transformation, and $x_{i,t}$ be a vector of predictor variables for account i at calendar time t. $x_{i,t}$ will include static application variables, having effectively the same value for all t, behavioural variables and a constant term to allow for an intercept in the models. The constant $l > 0$ is used to represent a lag on the behavioural variables and previous balance. The value l controls how far forward the model can predict ahead. In particular, a lag of l allows forecasts l periods ahead. Hence, a larger value of l implies a more useful model. However, a longer lag is likely to lead to weaker coefficient estimates, since the correlation between current balance and past data will reduce over longer periods. For this study, values of $l = 6$ and $l = 12$ months are used which allow a 6 month and 1 year forecast ahead period respectively. In practice, other values of l will be useful, depending on the business goal and availability of data.

2.1.1. Ordinary least squares

The simplest model to consider is the OLS regression model. The credit card balance of each account is chronologically correlated to previous balance. This correlation structure cannot be ignored in the analysis of the data. Therefore, if we use all the data to perform OLS regression, the estimates will not be reliable since the independence of observations assumption is violated. To circumvent this problem, OLS regression is performed over a cross-section of the data. However, the autocorrelation suggests previous balance as a lagged predictor variable. Therefore, for a fixed t, the model is given by the regression equation:

$$y_{i,t} = x_{i,t-l}^{\mathrm{T}}\beta + y_{i,t-l}\gamma + \varepsilon_{i,t} \tag{2.1}$$

where β is a vector of regression coefficient, γ is the regression coefficient on the lagged balance and $\varepsilon_{i,t} \sim N(0, \sigma^2)$ and i.i.d, for some σ. We found that diagnostic plots of the model residuals suggest the normality assumption on the error terms in the OLS model is