Models and forecasts of credit card balance

Pak Shun HON and Tony BELLOTTI,

Department of Mathematics, Imperial College London, UK

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 [11]. They account for 6.9 percent of bank assets and 16 percent of bank earnings in 2007 and this figure is rising [14]. 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. Firstly, it allows for an accurate revenue prediction and hence estimates of expected profits. Secondly, 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. Thirdly, it allows for estimation of assets on book at granular level; eg by risk groups. Fourthly, 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 [13]. 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 [13, 5, 11]. 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 [13] have shown that modeling 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 [9]. Estimation of profitability on retail loan portfolios have been published. However, outstanding balance is considered implicit [4] or is known for fixed term repayment loans [10], 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 optimize 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 [16], 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 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. Figure 1.1 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% of account records have a balance of exactly zero and 7.8% record a negative balance.

Figure 1.1: Histogram of raw balance (GB pounds)

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 [6]. This is an efficient way to transform monetary data involving zero and negative values. It is defined as:

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

A nice property of this function is that as $x \rightarrow 0$, $f(x) \sim x$. In addition, it scales down large values just as the logarithmic transformation does: as $x \rightarrow \pm\infty$, $f(x) \sim \pm \log(2\theta x)/\theta$. Figure 1.2 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.

Figure 1.2: Histogram of asinh transformed balance ($\theta = 1$).

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:

$$y_{i,t} = x_{i,t-l}^T \beta + y_{i,t-l} \gamma + \varepsilon_{i,t} \quad (2.1)$$

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 violated. This means the t-tests on the regression estimates are in doubt. Bootstrapping is useful when the validity of the statistical test is questionable. In this case, we compute confidence intervals on the regression coefficients by bootstrapping. The procedure is described as follows:

1. Let α be a significance level and m be the number of bootstrap samples.
2. Construct a bootstrap sample by random sampling n accounts with replacement from our training sample of n accounts.
3. Build the regression model using the bootstrap sample and record the estimate for each coefficient.
4. Repeat steps 2 and 3, m times to construct empirical distributions of the m estimates for each coefficient.
5. Construct $1 - \alpha$ confidence intervals for each coefficient as the range from the $\alpha/2$ and $1 - \alpha/2$ quantiles of the empirical distributions.
6. For each coefficient, let the null hypothesis be that it is zero. Therefore, if zero is not inside the confidence interval for the coefficient, then the null is rejected at the significance level α .

For our analysis, $\alpha = 0.05$ with $m = 1000$, since this gives a manageable computation time, whilst giving a reasonable significance level.

2.1.2 Two-stage regression model

There are significant numbers of zero balances in the data and this suggests an alternative model built in two stages:

1. A classification model to model whether balance is zero;
2. Conditional on balance being non-zero, build a regression model of balance.

Modelling in stages has already been suggested in the context of LGD modelling [1] as a decision tree approach. There are two potential benefits. Firstly, there may be distinct drivers for an individual having zero balance (ie potentially not using the card) to those drivers of an active balance. Secondly, the distribution for the dependent variable in the second regression step will be closer to normal, following an asinh transform, since it is the distribution excluding the stack of zeroes in Fig 1.2. As with OLS regression, two-stage regression models need to be built cross-sectionally. Logistic regression is used for the first stage classification and OLS regression for the second stage, although other approaches could be taken at each modelling stage.

Two-stage regression has been proposed for EAD using a mixture of survival models for the cases when EAD is above or below credit limit [9]. For balance (unconditional on default) this might suggest a three-stage model, where stage two given above is split into two separate models depending on the value of credit limit.

2.1.3 Mixture regression

The multimodal nature of credit card balance, as shown in Fig 1.2 suggests that mixture regression would be a better candidate to model credit card balance than standard OLS regression. Indeed, univariate analysis of the balance using mixture models suggests a five component structure:

$$X_i \sim \left\{ \begin{array}{lll} Deg & \text{with probability } (w.p) & \pi_1 \\ -Weibull(k_1, \lambda_1) & w.p & \pi_2 \\ -Gamma(\alpha_1, \beta_1) & w.p & \pi_3 \\ Weibull(k_2, \lambda_2) & w.p & \pi_4 \\ Gamma(\alpha_2, \beta_2) & w.p & \pi_5 \end{array} \right. \quad (2.2)$$

where *Deg* denotes the degenerate distribution concentrated at 0. The choice of this mixture of distributions was based on visualization of the distribution shown in Figure 1.2 which indicates a degenerate term at 0 and two Weibull distributions to the right and the left. The two Gamma distributions were introduced since further analysis showed that the two Weibull terms were unable to completely model the distributions around 0. This structure was fitted to the

balance data cross-sectionally using the EM algorithm. Overall, this proposed distribution fits the empirical distribution well, except for some overestimates at the lower and upper range of balance.

These univariate results suggest the use of mixture regression. Mixtures of linear regression have been proposed in the form of “switching regression” [12]. A mixture of linear regressions can be used in the event that the membership of the data points assigned is unknown. The following model structure is proposed:

$$y_i = \begin{cases} x_{i1}^T \beta_1 + \varepsilon_{i1} & w.p \quad \pi_1 \\ \vdots & \vdots \\ x_{iJ}^T \beta_J + \varepsilon_{iJ} & w.p \quad \pi_J \end{cases} \quad (2.3)$$

where $\varepsilon_{ik} \sim N(0, \sigma_k^2)$ i.i.d. for $k \in \{1, \dots, J\}$ and $\sum_{k=1}^J \pi_k = 1$. In the context of modelling balance, there is a cluster of zeros in the response. Therefore, the model is modified to suit this problem. Although the univariate analysis showed that the best fit to the distribution is provided by a degenerate plus four sub-distributions, for reasons of computational efficiency, a degenerate term plus a normal distribution will suffice. In particular, we propose

$$y_i = \begin{cases} x_i^T \beta + \varepsilon_i & w.p \quad \pi_1 \\ 0 & w.p \quad 1 - \pi_1 \end{cases} \quad (2.4)$$

Each observation is either related by a regression framework or a degenerate component. In the event that the observation y_i belongs to the regression mixture, we have $y_i \sim N(x_i^T \beta, \sigma^2)$ where x_i is the vector of covariates and β is the vector of regression coefficients. Otherwise, it belongs to the degenerate distribution concentrated at zero. An important feature of this framework is that the membership of each observation is purely based on the proposed probability π_j . Therefore, this implies x_i does not affect the membership of y_i .

The EM Algorithm can be used to fit the model for a finite number of normal mixtures [3]. We modify this approach to allow for the degenerate component. Given a set of independent observations $y_1, \dots, y_n$ and the corresponding sets of covariates $x_1, \dots, x_n$. The sets of parameters

$\theta = (\pi_1, \beta, \sigma^2)$ can be estimated by maximizing the log-likelihood.

$$L(\theta|x_1, \dots, x_n, y_1, \dots, y_n) = \sum_{i=1}^n [\log(\pi_1 \phi(y_i|x_i^T \beta, \sigma^2) + (1 - \pi_1)I(y_i = 0))] \quad (2.5)$$

where $\phi(y_i|x_i^T \beta, \sigma^2)$ is the univariate normal distribution with mean $x_i^T \beta$ and variance σ^2 .

The EM-Algorithm in this context is given as follows:

E-step: If we treat $\theta^{(k)}$ be the estimate of the parameters in the k th iteration.

$$Q(\theta|\theta^{(k)}) = \sum_{i=1}^n (\tau_{i2}^{(k)} \log(I(y_i = 0)) + \tau_{i1}^{(k)} \log(\phi(y_i|x_i^T \beta, \sigma^2))) \quad (2.6)$$

where

$$\tau_{ij}^{(k)} = \frac{\pi_j^{(k-1)} p_j(y_i|\theta)}{\pi_1^{(k-1)} p_1(y_i|\theta) + \pi_2^{(k-1)} p_2(y_i|\theta)}$$

is the posterior probability that the i th observation belongs to the j th mixture after the k th iteration and ϕ is the Gaussian density.

M-step: In the k th iteration,

$$\begin{aligned} \pi_j^{(k)} &= \frac{\sum_{i=1}^n \tau_{ij}^{(k)}}{n}, \\ \beta_j^{(k)} &= (X^T W_j^{(k)} X)^{-1} X^T W_j^{(k)} Y, \\ \sigma_j^{2(k)} &= \frac{\sum_{i=1}^n \tau_{ij}^{(k)} (y_i - x_i^T \beta_j^{(k)})^2}{\sum_{i=1}^n \tau_{ij}^{(k)}} \text{ for each } j \end{aligned}$$

where X is the design matrix for regression and $W_j^{(k)} = \text{diag}(\tau_{1j}^{(k)}, \dots, \tau_{nj}^{(k)})$.

The E-step and M-step are performed alternately until the convergence criterion is achieved.

This mixture model can be used to model balance using a structure generalizing that given in Section 2.1.1, fixed for cross-section t :

$$y_{i,t} = \begin{cases} x_{i,t-l}^T \beta + y_{i,t-l} \gamma + \varepsilon_{i,t} & w.p \quad \pi_1 \\ 0 & w.p \quad 1 - \pi_1 \end{cases} \quad (2.7)$$

where $\varepsilon_{i,t} \sim N(0, \sigma^2)$ and i.i.d.

We have developed our own computer program to implement the proposed mixture model in R using the EM algorithm. This outputs point estimates for model coefficients. Statistical significance of the coefficient estimates is computed using the same bootstrap method described for OLS regression in Section 2.1.1.

2.1.4 Random effect panel model

The three methods discussed so far have the limitation that they only allow models to be developed on cross-sections of data. With sufficient data, this may be reasonable. However, a great deal of useful data is not being used and this would become an acute problem if the amount of data in each cross-section is limited. It is therefore natural to consider using panel model regression to model balance. In particular, balance across the lifetime of each account is modelled with a random effect at each account level. Suppose there are q accounts and n observations in total. The regression equation is given in general by

$$Y = X\beta + Zb + \varepsilon \quad (2.8)$$

where Y denotes the vector of balances for a single account over time, X and Z denote matrices of known values (of possibly lagged variables), β is a vector of unknown coefficients having *fixed* values, b is a vector of q *random* variables (random effects) and ε is a vector of n error terms. In this model, we assume that

$$Y|b \sim N(X\beta + Zb, \sigma^2 I) \text{ and } b \sim N(0, \sigma^2 D).$$

for some variance σ^2 and matrix D . Note that

$$\text{var}[Y] = \text{var}[Zb] + \text{var}[\varepsilon] = \sigma^2 ZDZ^T + \sigma^2 I = \sigma^2 V$$

by letting $V = ZDZ^T + I$ which implies $Y \sim N(X\beta, \sigma^2 V)$ The parameters β , σ^2 and D are then found by maximizing the log likelihood of the data:

$$l(\beta, \sigma^2, D|y) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2 V) - \frac{1}{2\sigma^2} (Y - X\beta)^T V^{-1} (Y - X\beta). \quad (2.9)$$

The problem with using maximum likelihood estimators for this problem is that they are biased. A way to circumvent this is to use restricted maximum likelihood estimators (REML) [8]. In this study, we adopted REML. However, we found that the choice of estimator, ML or REML, actually made little difference to the coefficient estimates.

We require a special case of this model structure with just a random intercept for each account. Then for each account i ,

$$Y_i = \alpha_i + X_i\beta + \varepsilon_i \quad (2.10)$$

where $\alpha_i \sim N(0, \sigma_\alpha^2)$ is the random intercept for each account.

The previous lagged balance, $Y_{i,t-l}$ is deliberately not included in the model structure on the right-hand side since this changes the formulation into an *autoregressive* panel model which is not straightforward to estimate, since the predictor variable $Y_{i,t-l}$ will itself be dependent on the random effect; a typical approach to estimating such a model is to use instrumental variables [15]. However, in this case, it is not clear which instrumental variables could be used. Since previous balance is expected to be a strong driver of future balance on the same account, its exclusion from the panel model would seem a handicap. However, this is not the case since α_i acts as a relative “average balance” for account i . It takes the place of previous balance as a predictor variable, and in a more flexible way since it does not require the use of a fixed lag l . Therefore, to predict a balance for a new observation at time t using the panel model, compute

$$\mathbb{E}[y_{i,t}|x_{i,t-l}, \hat{\alpha}_i] = \hat{\alpha}_i + x_{i,t-l}\hat{\beta} \quad (2.11)$$

where $\hat{\beta}$ is the estimate from the model and $\hat{\alpha}_i$ is an estimate of the random effect for i . This can be obtained using an empirical Bayes approach. That is,

$$\hat{\alpha}_i = \mathbb{E}[\alpha_i|Y] = \int \alpha_i f(\alpha_i|Y) d\alpha_i \quad (2.12)$$

and through Bayes theorem, $f(\alpha_i|Y) \propto f(Y|\alpha_i)f(\alpha_i)$, the Bayes estimate for α_i , and using the MLE for D, V and β , the estimate for the random effect is obtained:

$$\hat{\alpha}_i = DZ_i^T V^{-1} (Y_i - X_i\hat{\beta}). \quad (2.13)$$

This estimate requires a history of data to estimate for an account. This estimate will not be available for newer accounts, and so for these cases an estimate of $\hat{\alpha}_i = 0$ is imposed.

2.2 Evaluation

A forecasting framework is used to evaluate the models. That is, given a fixed observation date d_O , the training data set is constructed from observations prior to or on this date and test data is constructed from observations after this date. However, forecasts can only be made for accounts in the test data set which have past data available in the training data set.

Balance forecasts can be made directly for the OLS regression and panel models. For the two-stage model, firstly a forecast is made based on the stage 1 classifier. If it is 0 then a balance of zero is forecast, otherwise the stage 2 regression model is used for the forecast. In this study, we have used a logistic regression model which models $P(y_{i,t} = 0 | x_{i,t-l}, y_{i,t-l})$. A threshold value is used to then forecast whether balance is zero; ie if $P(y_{i,t} = 0 | x_{i,t-l}, y_{i,t-l}) > 0.5$ then forecast $\hat{y}_i = 0$. The choice of the threshold 0.5 seemed a natural choice and leads to approximately the correct proportion of accounts being estimated with zero balance (eg for cross-section $t = 37$, 36.1% of accounts were observed with zero balance and 42.7% were predicted with zero-balance). For the mixture regression model, we cannot classify the membership for each observation normally. However, taking a Bayesian approach, the posterior probabilities for the membership of y_i belonging to the regression component, rather than zero, is given by

$$q_i = \frac{\pi_1 p_1(y_i | \theta)}{\pi_1 p_1(y_i | \theta) + (1 - \pi_1) I(y_i = 0)}. \quad (2.14)$$

Again, we take a threshold of 0.5. If $q_i > 0.5$ then forecast $\hat{y}_i = x_{i,t-l}^T \hat{\beta} + y_{i,t-l} \hat{\gamma}$ where $\hat{\beta}$ and $\hat{\gamma}$ are the estimated coefficients from the mixture regression, otherwise forecast $\hat{y}_i = 0$.

Estimates of balance, $\hat{y}_i$, are evaluated at an account-level by comparison with the true balance, y_i . The typical performance measures, mean absolute error ($MAE = \frac{1}{n} \sum_i |y_i - \hat{y}_i|$) and root mean squared error ($RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$), are used for the evaluation. For consistency, the measure is always made with respect to the monetary value of balance, and not on asinh transformed balance, since ultimately financial institutions want an accurate forecast of the monetary value, not the transform.

3 Data and data processing

The data used in this study consists of over 38,000 UK credit cards with monthly accounts and behaviour over a 3 year period from June 2008 to June 2011. The data is provided anonymously by a UK financial institution. We use month as the chronological unit in our analysis, therefore $t = 1$ refers to June 2008 and $t = 37$ refers to June 2011. Full accounting records are not available for all credit cards over this time period since the data includes credit cards opened, closed and defaulting during that period. The data includes application variables: age of applicant, employment status, tenure, months at current address and application channel. It also has behavioural variables: statement number, total outstanding balance, credit limit, cash balance, account status (ie open, closed, charged-off or fraud), delinquency, payment amount and number of payments. Total outstanding balance is the balance amount at the end of the accounting month (the date of the end of the accounting month is different for each customer), prior to payments being made by the customer into the account. This balance variable is the dependent variable used in this study. The delinquency variable measures the number of outstanding months of missed minimum payments. As with common practice, default is defined as 3 or more months of missed payments; ie delinquency ≥ 3 . Statement number is used as a proxy for account age. Usually this is fine, but there are a few occasions when statement number is not an exact measure of months open, due to administrative actions that can shift the date of statement production.

There were 12 credit cards with duplicate entries in the same calendar month. There were also 36 credit cards with a fraud account status. Also, 35 credit cards had missing values on age. These credit card accounts were removed from the data set prior to analysis. Additionally, only account records were included in the analysis where account status was open and the account was not in default. Post-default balances were not considered since the account is usually frozen at this point and balance is the subject of recovery procedures, therefore it is appropriate for LGD and EAD modelling to be used at this stage. After cleaning the data this left just over 37,000 credit cards with a mean 25.5 accounting months per credit card.

Due to commercial confidentiality constraints, full descriptive statistics of the data cannot be given. However, the distribution of balance is shown in Section 1.1. For modelling balance, apart from using the asinh transformation described in that section, we should also consider the inflation rate. However, rather than artificially adjusting the value of balances based on

official inflation figures, we will rely on the models to adjust for inflation in the estimates of the intercept and/or coefficient on lagged previous balance. Inflation rates were low during the period of the data (from 1.1% to 5.2%, ONS figures). For forecasting, if inflation is expected to be higher during the forecast period, then forecast balance should be adjusted too. Since we cannot make a prior assumption of the relationship between age and balance, the age variable is discretized into four age groups prior to inclusion in the models: under 30, 30-50, 50-65 and over 65. When the asinh transformation is used on balance, asinh is also applied to the monetary predictor variables (previous balance, credit limit, cash balance and payment amount) so that the influence of these variables is on the same scale as the outcome variable. Prior to inclusion in the models, all variables were tested for multicollinearity by computing the Variance Inflation Factor (VIF) [7]. All VIFs were less than 5, so all variables were included in the model.

4 Results

In this section, we present the model estimation and forecasting results.

4.1 Model estimation

Table 4.1 shows coefficient estimates for three models of asinh transformed balance. The OLS and mixture regression models were built on the cross-section of training data taken at $t = 37$. The adjusted coefficient of determination ($\text{adj-}R^2$) of the OLS model is 0.592 which indicates a reasonably good fit. The panel model uses all available data. All models use a lag of $l = 12$ months on behavioural variables. Since a lag of 12 months is required, account with less than 12 months of data before the modelled period are excluded from the training data.

Table 4.1 shows that several variables are associated with balance, at a 5% statistical significance level. More variables are shown to be statistically significant in the panel model, but this is most likely due to the larger sample of data available when building the panel model.

Both the OLS and mixture regression models have previous balance (asinh, lag 12 months) available as a predictor variable. The effect size demonstrates that this is by far the strongest predictor of current balance. Cash balance (asinh) also has a significant and positive association with balance in both models, but the effect size is much less. This cash balance effect was

	OLS regression	Mixture regression	Panel model
Intercept	0.420 *	1.10 *	2.10 **
Application variables			
Age of applicant:			
Under 30	-0.172 *	-0.225 *	-0.830 **
30-49 (excluded)	0	0	0
50-65	-0.0909	-0.0959	-0.245 **
Over 65	-0.417 *	-0.497 *	-0.228
Employment status:			
Employed (excluded)	0	0	0
Home maker	-0.0405	-0.0096	0.217
Retired	0.0391	0.060	0.150
Self-employed	-0.0178	-0.063	-0.0577
Student	0.148	0.228	-0.0878
Tenure:			
Home owner (excluded)	0	0	0
Council tenant	0.0902	0.0993	0.509 **
Lives with parents	-0.0061	-0.0134	0.178
Private tenant	0.176 *	0.182 *	0.444 **
Month at current address	0.0001	0.0001	0.0007 *
Application Channel:			
Mail (excluded)	0	0	0
Cold calls	-0.0144	0.0246	-0.165
Doordrops	-0.0555	-0.0775	-0.240
By inserts	0.0237	-0.0064	-0.640 **
Internet	0.0792	0.0721	0.0835
Others	-0.0903	-0.188	0.212
Behavioural variables (lagged 12 months)			
Account age (months)	-0.0019 *	-0.0006 *	-0.0129 **
Credit limit (asinh)	0.0391 *	-0.0516	0.323 **
Balance (asinh)	0.621 *	0.652 *	n/a
Cash balance (asinh)	0.0862 *	0.0775 *	0.0028
Payment amount (asinh)	-0.206 *	-0.183 *	0.0258 **
Number of payments	0.0845	0.0278	-0.0253 **
Delinquency (months)	0.0497	-0.0692	-0.0900 **
Random effect: Account id	n/a	n/a	3.69 s.d.
Probability weight π_1	n/a	0.9465	n/a

Table 4.1: Coefficient estimates for models of Balance (asinh); * 5% significance level, ** 1% significance level, s.d. = standard deviation of random effect. Statistical significance for OLS and mixture regression models is based on bootstrap method and is only measured at 5% level.

consistent across other cross-sections for OLS and mixture regression, but interestingly it is not a significant predictor in the panel model. This is most likely due to the fact that the individual-level random effect is absorbing both balance and cash balance effects.

The age of the credit card holder at application is an important factor with all other age groups indicating less balance than the excluded category of 30-49 year olds. For OLS and mixture regression, the over 65s have the smallest balance, but this is not evident when using the panel model.

Employment status does not have a statistically significant association with balance in any of the models.

Private tenants have statistically significant higher balance, compared to home owners, in all models. Additionally, the panel model shows that council tenants also have higher balances compared to home owners.

Months at current address is only statistically significant in the panel model, but it suggests this has a positive association with balance.

Application by inserts in newspapers and magazines is only statistically significant in the panel model, but it suggests that this has a negative association with balance, relative to the largest group, mailed invitations to apply.

Account age has a negative association with balance in all models. This may be an indication that individuals are more likely to use their cards when it is new.

Credit limit (asinh) has a positive and statistically significant association with balance in both the OLS and panel models. This is expected since it is a contractual upper bound on balance. However, the effect size is small in the OLS model, since the model has previous balance available as a stronger predictor variable.

Payment amount (asinh) has a negative association with balance in both the OLS and mixture regression models. A negative value of payment amount implies a positive amount paid to the credit card company. Therefore, this coefficient shows that a larger payment to the credit card company implies a larger future balance. However, for the panel model, the estimated association is reversed, although its effect size is much smaller. Additionally, number of payments also becomes statistically significant in the panel model with negative association to balance which reinforces the switch in direction of effect of payments in the panel model.

Past delinquency has a negative and statistically significant association with balance in the panel model. This is suggestive that people generally reduce their credit card balances following delinquency, perhaps in agreement with the credit card providers who will be pressing delinquent credit card holders to manage their debt. The alternative are customers who become more severely indebted, leading to default. Since these customers enter default they would not be included in this modelling exercise.

The random effect estimated for the panel model has a large standard deviation, relative to the distribution of balance (asinh) as shown in Fig 1.2. This indicates the importance of this account-level effect in explaining balance.

For the mixture regression model, the probability of an observation belonging to the degenerate component is 0.0535. This represents the portion of accounts that always having zero balance as 5.35%. If an observation belongs to the degenerate component this means the account is inactive at all times. The portion of observed accounts that are zero at all times ($t = 1$ to 37) is 7.9% which is close to the estimate given by the model. We chose the mixture regression model because we supposed the mixture would model balance, apart from Gaussian noise. Fig 4.1 shows that, with some fluctuation and some under-estimation in the lower quantiles, the distribution of residuals approximately matches a normal distribution, as we require.

These regression results will be different for different time periods, especially for the OLS and mixture regression models, since they are based on cross-sectional training data. However, the coefficient estimates for variables balance, payment amount, age groups, tenure and application channel are found to be fairly consistent in terms of statistical significance and magnitude over several alternative cross-sections.

4.2 Comparative forecasting performance

An observation date of $d_O = 15$ is used. This gives sufficient data to lag behavioural variables and allows us to forecast ahead. Therefore, for OLS and mixture regression, the cross-section at $t = 15$ is used for training, whereas the panel model can use all data from $t = 1$ to 15. We want to forecast ahead by 12 months using data independent of the training data, so we forecast balance for $t = 27$ with all models. Hence they all use the same test data set and the results are comparable. We may be interested in other forecast ahead periods, but because of the limited range of data (3 years), forecasting beyond 12 months is not possible. However, we

Figure 4.1: Q-Q plot for the mixture regression model: $t = 37$

show results for a shorter 6 months period for contrast with the 12 month forecasts. Since the OLS, two-stage and mixture regression models rely on lagged predictor variables and the panel model requires a history of data for a Bayesian estimate of the random effect, only accounts with at least 12 months of history are included in this study, or 6 months if 6 month forecasts are reported. Table 4.2 summarizes the results.

The intercept only OLS model has no predictor variables and is included as a benchmark. We would expect all models to perform better than this and the extent of improvement tells us the strength of the model for forecasting.

We know that previous balance is strongly associated with current balance and so the OLS model with only lagged balance as a predictor variable is used to determine how much improvement the other predictor variables give. Comparing this against the OLS full model (ie includes all available predictor variables), it can be seen that the improvement is modest, if not approximately the same, for both MAE and RMSE measures. Therefore, this suggests, for forecasting, it is perhaps only necessary to consider previous balance.

Stepwise general variable selection based on the AIC criterion is also used (“stepwise selec-

			6 months		12 months	
Model	Trans.	Variables	MAE	RMSE	MAE	RMSE
OLS	asinh	intercept only	961	2122	963	2143
OLS	none	intercept only	1098	1907	1110	1931
OLS	asinh	lagged balance only	718	1611	849	1917
OLS	none	lagged balance only	452	1039	616	1258
OLS	asinh	full model	677	1563	807	1856
OLS	none	full model	451	1024	611	1236
OLS	asinh	stepwise selection	679	1566	808	1857
OLS	none	stepwise selection	451	1024	612	1236
2 stage model	asinh	full model	961	2127	962	2146
2 stage model	none	full model	435	1030	584	1228
Mixture regression	asinh	full model	575	1310	681	1550
Mixture regression	none	full model	454	1028	628	1231
Panel model	asinh	full model	530	1409	611	1603
Panel model	none	full model	431	1075	543	1272

Table 4.2: Forecast performance of different statistical models for both 6 month and 12 month forecast ahead periods. Trans = Transformation. Bold shows the best performance.

tion”) to determine if a parsimonious model might perform better. However, the results are very similar to the OLS full model results and so this does not suggest any benefit.

In almost all cases, the use of the asinh transformation of balance reduces performance for both MAE and RMSE. This is in spite of it being the superior approach in terms of the distribution of outcome variable and structure of error terms in the model. In addition, modelling without transformation violates the homoscedasticity assumption of the relevant model. The reason modelling the balance directly is beneficial is that is the term which is directly evaluated by the performance measures. This is not surprising as the least square estimator is always the best linear unbiased estimator of balance for the relevant statistical model of balance.

The 6 months ahead forecasts are always better than the 12 months ahead forecasts, for all models, and this is natural since we would expect it to be more difficult to forecast further ahead into the future. In terms of the RMSE measure, the simple OLS approach is competitive with the more complex models. However, for 12 month ahead forecasts, we observe that the two-stage model improves slightly on the OLS models when no transformation of balance is used. The mixture regression model did not show improvement when using either MAE or RMSE measure. We found that the panel model estimator did not consistently converge

correctly for the full model structure ¹. To resolve this problem, we attempted to reduce model complexity by removing two of the weakest predictor variables, cash balance and payment amount. This solution worked and led to consistent model convergence. This panel model gave the best performance using the MAE measure. This is a consequence of its use of a larger training data set, but is also a consequence of the way it uses the empirical Bayes estimate of the random effect in the forecast, instead of using lagged balance as a predictor variable. We tested this hypothesis by producing forecasts using a simple 0 estimate for the random effect. As expected, this produced poor results (eg for 12 month forecasts, and with no asinh transformation, MAE and RMSE were 1055 and 1789 respectively).

Further analysis taking different values of d_O (15 to 25), and hence different training and test data, have shown that these results are consistent over time in terms of the ranking of the statistical models. However, there are some periods of test data that give generally better forecast results than others.

5 Conclusion

In this study we have modelled credit card balance using several alternative model structures and assessed the performance of the models based on how well they forecast balance at account-level. An initial univariate analysis of the distribution of balance suggests a “logarithm-like” transformation using the inverse hyperbolic sine (asinh) function to deal with negative values. Additionally, it is apparent that credit card balance is a mixture of several distributions: in particular, a degenerate component to deal with zero balances, along with a skewed distribution of positive balances. This structure suggested the use of two-stage and mixture regression models. We also considered a panel model structure in order to take advantage of data over time for multiple accounts. The results of Section 4.1 suggest several drivers of balance with the strongest being previous (lagged) balance. In order to avoid building a complicated autoregressive panel model, we utilized an empirical Bayes estimate of the random effect to act as a proxy for previous balance. This strategy was successful since the panel model gave good results.

For forecasting, the use of the asinh transformation does not give better performance. This result is consistent with findings for LGD [16, 1] and is a consequence of the fact that the

¹In particular, we used the `lme` and `glmmPQL` packages in R. Using `lme` often led to a convergence error, whilst using `glmmPQL` with a Gaussian link function led to a biased coefficient estimate on cash balance.

models built on transformed balance will be biased estimators of balance. Both the two-stage and mixture regression models perform well and are competitive with the standard OLS model. However, Table 4.2 demonstrates that the random effect panel model is clearly the best model based on using the MAE measure of performance. This is a consequence of the larger volume of data available for training and also due to the way the empirical Bayes estimate of the random effect is used in forecasting. Arguably, MAE is the most appropriate measure for this problem, since differences in financial values are being measured, and it makes less sense to work from square monetary difference as is the case with RMSE.

There remain some limitations of the models that would form the basis of future work. Firstly, although the panel model elicits some advantage from modelling balance based on past training data, rather than just a cross-section, it is expected that using data that is too old would actually have a detrimental effect, due to population drift. For this reason, further investigation into the optimal time period to use in panel model training would be valuable. The optimal time period for best predictive performance for our data is 3-6 months of latest data. Secondly, the error term and random effect in the panel model are both assumed to be normally distributed. Given the distribution of the dependent variable, balance, these assumptions are unrealistic. For this reason further work developing two-stage and mixture regression models within the framework of panel models may be beneficial. Thirdly, although lagged behavioural predictor variables are included, further variables could be considered to model the dynamics of balance. In particular, whether there is a vintage effect to balance would be an interesting question. Also, modelling associations between balance and economic conditions could be studied by including macroeconomic variables in the panel model framework [2]. This would be very useful if revenue predictions during recession periods were required as part of stress testing.

To our knowledge, this paper is one of the first published studies into modelling balance, unconditional on default. This is an important problem motivated by the need for accurate revenue prediction and optimal customer/risk management. We have proposed several alternative methods and tested them on a real data set of UK credit cards.

References

- [1] T. Bellotti and J. Crook. Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28:171–182, 2012.

- [2] J. Crook and T. Bellotti. Time varying and dynamic models for default risk in consumer loans. *Journal of the Royal Statistical Society series A (Statistics in Society)*, 173:279–468, 2010.
- [3] R.D. De Veaux. Mixtures of linear regressions. *Computational Statistics and Data Analysis*, 8:227–245, 1989.
- [4] S. Finlay. Credit scoring for profitability objectives. *European Journal of Operational Research*, 202:528–537, 2010.
- [5] G. Jimenez, J. A. Lopez, and J. Saurina. EAD calibration for corporate credit lines. *Federal Reserve Bank of San Francisco*, 2009.
- [6] N.L. Johnson. Systems of frequency curves generated by methods of translation. *Biometrika*, 36:149–176, 1949.
- [7] P. Kennedy. *A guide to econometrics (5th ed.)*. Blackwell, 2003.
- [8] N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, December 1982.
- [9] M. Leow and J. Crook. A two stage mixture model for predicting EAD. In *Credit Scoring and Credit Control Conference XIII*, www.business-school.ed.ac.uk/waf/crc_archive/2013/4.pdf, 2013.
- [10] P. Ma, J. Crook, and J. Ansell. Modelling take-up and profitability. *Journal of the Operational Research Society Issue 3*, 61:430–442, 2010.
- [11] M. Qi. Exposure at default of unsecured credit cards (working paper). *Office of the Comptroller of the Currency Publications*, July 2009.
- [12] R. Quandt and J. Ramsey. Estimating mixtures of normal distributions and switching regression. *Journal of American Statistical Association*, 73:730–738, 1978.
- [13] R. Taplin, H. M. To, and J. Hee. Modeling exposure at default, credit conversion factors and the Basel II Accord. *Journal of Credit Risk*, 3:75–84, 2007.
- [14] H. Terris. Cards contributing more bank income. *American Banker*, April 2008.

- [15] M. Verbeek. *A guide to modern econometrics (2nd ed.)*. Wiley, 2004.
- [16] B. H. Yang and M. Tkachenko. Modeling exposure at default and loss given default: empirical approaches and technical implementation. *Journal of Credit Risk*, 8:81–102, 2012.