Statistical Modelling 2005; **5**: 145–158

Measuring customer quality in retail banking

David J Hand and Martin J Crowder
Department of Mathematics, Imperial College London, London

Abstract: The retail banking sector makes heavy use of statistical models to predict various aspects of customer behaviour. These models are built using data from earlier customers, but have several weaknesses. An alternative approach, widely used in social measurement, but apparently not yet applied in the retail banking sector, is to use latent-variable techniques to measure the underlying key aspect of customer behaviour. This paper describes such a model that separates the observed variables for a customer into primary characteristics on the one hand, and indicators of previous behaviour on the other, and links the two via a latent variable that we identify as 'customer quality'. We describe how to estimate the conditional distribution of customer quality, given the observed values of primary characteristics and past behaviour.

Key words: credit cards; financial delinquency; latent variables; loan default; prediction; random effects; retail banking; scorecards

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received March 2003; **revised** April 2004, November 2004; **accepted** November 2004

1 Introduction

The retail banking sector, that sector of banking concerned with credit cards, mortgages, personal loans, car finance, and so on, makes heavy use of predictive statistical models called *scorecards*. These are models that yield a score indicating how likely it is that some particular outcome will occur. They are used, for example, to predict the likelihood that a customer will default on a loan, whether or not an applicant will take up an offer, how reliable they will be with repayments, whether they will switch between mortgage suppliers, whether an application is fraudulent, and so on. A distinction is made between *application* and *behavioural* scorecards. Application scorecards are used to make decisions about whether to give customers a particular financial product in the first place (e.g., whether to grant an applicant a loan or credit card), while behavioural scorecards are used to monitor the behaviour of existing customers (e.g., to see how reliably customers keep up to date with repayments or how much they use their credit cards). This paper is mainly about application scorecards, although the ideas can be extended to handle ongoing monitoring of customer behaviour. Reviews of the literature are given in Rosenberg and Gleit (1994), Hand

Address for correspondence: DJ Hand and MJ Crowder, Department of Mathematics, Imperial College London, London SW7 2AZ, UK. E-mail: d.j.hand@imperial.ac.uk; m.crowder@imperial.ac.uk

 10.1191/1471082X05st092oa

and Henley (1997), Hand (2001a), Thomas (2000), and Thomas *et al.* (2002), and a (statistically) nontechnical introduction to the area is given by McNab and Wynn (2000).

A large number of different approaches has been used to develop models for prediction. In practical application, by far the most common is the logistic discrimination model; others include linear discriminant analysis, linear regression, classification trees, neural networks, and rule-based approaches. Despite their variety, all such models have the same generic form: they all try to predict a single outcome variable c (usually categorical) from a set of predictor variables $x = (x_1, \ldots, x_r)$. The database from which the prediction is made forms a retrospective sample from previous customers. These retrospective data contain values for both predictor variables x and outcome variable c. That is, c is an observed indicator of what eventually happened to the earlier customer: an indicator of default on a loan, a decision not to take up an offered product, an indicator of unreliability in repayments (e.g., consistent lateness), an indicator that the application turned out to be fraudulent, and so on. Variables used as predictors include such things as home status (rent or own), age, income, occupation, time at present address, time with employer, maximum number of months ever in arrears with credit card repayments, number of adverse County Court judgements, history of voluntary transfer between mortgage suppliers (known as 'churn'), and aspects of current account history.

If one is focused specifically on the particular outcome variable c, and given that c can be directly observed for previous customers or applicants, then the standard strategy is clearly attractive. However, it is not without its weaknesses, listed as follows.

(1) The database is necessarily retrospective and out of date. For example, to know whether a customer with a loan repays or defaults, one must wait until the end of the loan term, perhaps several years; for example, in the case of a mortgage the period could be 25 years or more. It is then an act of faith to believe that the distribution of (x, c) will be the same for new customers.
(2) The database is often nonrepresentative. For example, with loans, the 'good/bad' outcome will only be known for those customers actually granted a loan, the 'accepted,' not for the entire population of applicants. The resulting selectivity-bias problem has been the focus of much interest in the retail banking community (Hand and Henley, 1993; Hand, 2001b; Crook *et al.*, 2001) and methods for tackling it are used in almost every credit-scoring operation.
(3) To use a standard model, c must (eventually) be directly observable, or at least defined explicitly in terms of variables that can be directly observed.
(4) The definitions of outcome variables are often fairly arbitrary. For example, a common criterion of quality in loan repayments is that the customer never falls more than three months in arrears with repayments, but the choice of three here has no underlying rationale. The definitions are also often complex. Hand *et al.* (2001) gave a real example of current account usage in which the 'bad' class is defined as a complex nonlinear function of multiple variables, including measures of account activity, size of balance, amount overdrawn, and so on.
(5) Standard models tend to be *product-centric*; that is, separate models are built for each product, and each customer will have different scores for each product. This

is hardly ideal. One can imagine a situation in which a customer is offered a mortgage but turned down for a credit card.

(6) Finally, scorecard models are intrinsically empirical, with no notion of any underlying mechanism. This is made very apparent if one observes the operations of teams constructing banking scorecards in practice, where extensive searches through variables and explorations of grouping strategies to convert any continuous variables to categories are carried out using the current database. The consequence is that scorecards might not generalize well to new circumstances. This is of particular concern at present since we are currently witnessing a dramatic rise in consumer debt. This rise, and indeed the rate of increase in the rise of debt, has now become of major concern to governments, as well as being a major concern to individual banks and financial institutions. It implies a vulnerability to economic shocks.

There is an alternative to the above class of models, an alternative developed by the psychometric and Social Statistics communities (Bartholomew and Knott, 1999), but which appears never to have been used in the retail banking sector. Rather than trying to predict a variable that can be directly measured, this approach seeks to identify the common influence underlying a set of observed variables. If these are all measuring aspects of customer reliability, creditworthiness, financial responsibility, or *quality*, then their common aspect can be taken as a measure of this attribute. More formally, if one can hypothesize the form of the relationships between the quality and the observed variables, then one can deduce the value of the former from the values of the latter.

The particular type of latent-variable model proposed in this paper is as follows. Of core interest is q, the *customer quality*. This is itself unobservable but is related to x, a vector of observable primary characteristics, and in turn has a direct influence on y, a vector of observable behaviourable variables: simply, $x \rightarrow q \rightarrow y$. In effect, we split the previously mentioned predictor variables into an x-set and a y-set. Thus, we might take x to include home status, age, income, occupation, time at present address, and time with employer; y might then consist of arrears history, adverse County Court judgements, mortgage churn, and current account history. Clearly, the demarcation of x and y variables is to some extent arbitrary, relying on industry-based judgement.

This type of model, in which the predictor variables are partitioned into two types, those describing primary characteristics of the customer and those describing behaviour, appears not to have been previously applied within the retail banking context. And yet, it is well suited to this application. As we show below, at the very least, it represents a step away from a purely descriptive model towards a mechanistic model of how customers respond. That is, it is not a solely data-driven model, but has some underlying rationale about the nature of the variables and the customers. It is this nature of the model that gives some grounds to believe that it can go some way towards overcoming the weaknesses (1–6) listed above. For (1) and (2), if the relationship $x \rightarrow q \rightarrow y$ is fairly stable, outdatedness and unrepresentativeness of the (x, y) database should have less impact. For (3), q is not observable but can be estimated from x and y. For (5) and (6), q is not tied to any particular product, and is meant to reflect an underlying trait. Weakness (4) is a matter for the industry: q can be estimated from whatever variables are deemed to be appropriate.

Sammel *et al.* (1997) proposed a similar model, which will be compared and contrasted below. The model is also related to the Multiple-Indicator-Multiple-Cause (MIMIC) model (e.g., Bollen, 1989; Joreskog and Goldberger, 1975) sometimes used in the social sciences, though that is typically based on assumptions of multivariate normality. We make no such restrictive assumptions. The recent book by Skrondal and Rabe-Hesketh (2004) gives an extended account of latent-variable modelling.

The model is developed in detail in Section 2. In Section 3, the associated likelihood function is defined, maximum likelihood estimation considered, and a brief description of model assessment and prediction of customer quality given. One of the difficulties of working in this application area is the highly confidential nature of the data: the use of the model is illustrated on a public-domain data set in Section 4. Finally, in Section 5 some further issues are discussed. At the time of writing we are beginning to explore applying the model in practice with various financial institutions.

2 A customer quality model

Partitioning the predictor variables in the way described here leads to a model with three types of variable for each customer: $x = (x_1, \ldots, x_r)$ and $y = (y_1, \ldots, y_s)$, the recorded vectors of primary and behavioural variables, and q, the unobserved latent quality variable. The last one, q, serves as an intermediary between x and y, and in our model we assume that it is the sole intermediary: all influence of x on y is channelled through q. Specifically, we assume that

$$p(y \mid x, q) = p(y \mid q); \tag{2.1}$$

we use $p(\cdot)$ generically to denote a probability mass function or density. It is then natural to base the modelling on the two conditional distributions $p(q \mid x)$ and $p(y \mid q)$. We will assume that

$$p(q \mid x) = \phi(q; x^{\mathrm{T}}\psi, 1); \tag{2.2}$$

$$p(y \mid q) = \prod_{j=1}^{s} p(y_j \mid q); \tag{2.3}$$

$\phi(\cdot\,; \mu, \sigma^2)$ denotes the density function of $\mathrm{N}(\mu, \sigma^2)$ and ψ is a vector of regression coefficients. The normal distribution for q, given x, is in the nature of a working model: it is not crucial and can be easily replaced by a different distribution. To eliminate a potential lack of identifiability (discussed below) the scaling of q has been determined by taking its variance to be 1. In (2.3), the conditional independence of the y-components given q is typical of random-effects or frailty models: it identifies q as the sole source of correlation among the y_j.

The forms of the distributions of the y_j will depend on the nature of the behavioural variables. For example, y_j could be normally distributed (a measure of financial activity), binary (repayment or default on a loan), a Poisson count (number of delinquent events over a period), or a right-censored lifetime variable (duration

of some ongoing behaviour, not yet brought to a close). Often, $p(y_j \mid q)$ will be taken to be a generalized linear model, with parameter $\alpha_j + \beta_j q$, and then the scaling of q can be absorbed into β_j. Again, if $\mathrm{E}(q \mid x) = \psi_0 + x^{\mathrm{T}}\psi$, with a separate intercept term ψ_0,

$$\alpha_j + \beta_j q = (\alpha_j + \beta_j \psi_0) + \beta_j (x^{\mathrm{T}}\psi) + \textit{error term};$$

then, $\beta_j \psi_0$ can be absorbed into α_j, in effect taking $\psi_0 = 0$. The sign of ψ can also be changed provided that the signs of the β_j are likewise changed; this indeterminacy can be resolved by imposing a constraint such as fixing the sign of ψ_1.

The proposal of Sammel *et al.* (1997) is similar to ours here. However, there are significant differences between the two approaches in both inferential purpose and numerical treatment. Their model has (2.2) as here, but in (2.3) they make the restriction that the $p(y_j \mid q)$ are one-parameter exponential-family densities with link-transformed means

$$\eta_j = \gamma_{0j} + \gamma_{1j} q + \gamma_{2j} u_1 + \cdots + \gamma_{t+1,j} u_t,$$

in the usual notation of generalized linear models, incorporating covariates u_j. They stated that 'The primary focus is on estimating the [regression] parameters'. Our focus is quite different: we are only interested in predicting the random effects q for individuals. The other main difference is in the numerical method applied to perform maximum likelihood estimation. Since q is unobserved it needs to be integrated out in order to compute a likelihood function. We take a direct route, using numerical integration as detailed in the Appendix. In contrast, Sammel *et al.* (1997) travelled a more circuitous route, developing a modified EM algorithm for the purpose. Since it is generally not possible to obtain explicit expressions for either the M-step or the E-step, they resorted to approximations. To solve the score equation in the M-step a one-step Fisher scoring algorithm was applied; for the integration required in the E-step, either Monte Carlo or Gauss–Hermite quadrature was used, the latter being found to be more efficient. The analysis presented by Sammel *et al.* (1997) for implementation of their approach is fairly involved and there are issues concerning the accuracy of Gauss–Hermite quadrature in this context – see the Appendix here.

3 Estimation, goodness of fit and prediction

3.1 Maximum likelihood

Suppose that a data set is available, comprising observations (x_i, y_i); for the ith individual $(i = 1, \ldots, n)$ the vectors x_i and y_i have dimensions r and s, respectively. The log-likelihood function, conditional on the recorded x-values, is $l(\phi, \psi) = \sum_{i=1}^{n} l_i(\phi, \psi)$, where ϕ and ψ are the vectors of parameters governing $p(y \mid q, \phi)$ and $p(q \mid x, \psi)$, and

$$l_i(\phi, \psi) = \log p(y_i \mid x_i, \phi, \psi) = \log \int p(y_i \mid q_i, \phi) p(q_i \mid x_i, \psi) dq_i. \qquad (3.1)$$

Numerical evaluation of the likelihood function, its maximization, and the extraction of standard errors for the parameter estimates are all described in the Appendix.

3.2 Normal linear case

Suppose that the conditional distribution of y given q is s-variate normal, $N(\mu, D)$, where μ has jth component $\alpha_j + \beta_j q$ and $D = \text{diag}(\sigma_1^2, \ldots, \sigma_s^2)$; thus, $\phi = (\alpha_1, \ldots, \alpha_s; \beta_1, \ldots, \beta_s; \sigma_1, \ldots, \sigma_s)^T$. In this special case, the integration in (3.1) can be performed explicitly, either by completing the square in q in the exponent in the integrand or as follows. Write $y = \alpha + \beta q + e$, where $\alpha = (\alpha_1, \ldots, \alpha_s)^T$, $\beta = (\beta_1, \ldots, \beta_s)^T$, and $e = (e_1, \ldots, e_s)^T$ has distribution $N(0, D)$. Then $y \mid x$ has distribution $N(\mu, \Sigma)$, with $\mu = \alpha + \beta(x^T\psi)$ and $\Sigma = \beta\beta^T + D$, $\beta\beta^T$ having (j, k)th element $\beta_j\beta_k$. In the one-dimensional case, when y has only one component, there is a lack of identifiability: in that case, since $\mu = \alpha_1 + \beta_1(x^T\psi)$ and $\Sigma = \beta_1^2 + \sigma_1^2$ (a 1×1 matrix), one can set $\sigma_1 = 0$ without loss of generality because ψ can be rescaled in μ to accommodate the value of β_1 required for Σ.

It is a property of the model that y tends to be poorly predicted from x. This is because the quality of prediction in normal linear regression depends on the ratio of β to the error variance: roughly speaking, the larger the better. However, in this case, the error variance itself contains a component proportional to β^2, so the usual criterion is sabotaged. This weakness does not matter for our purpose, because we are not seeking to predict y.

For the normal case, in particular, the score vector has components

$$\frac{\partial l_i}{\partial \psi_j} = \beta^T \Sigma^{-1}(y_i - \mu_i)x_{ij};$$

$$\frac{\partial l_i}{\partial \alpha_j} = \{\Sigma^{-1}(y_i - \mu_i)\}_j;$$

$$\frac{\partial l_i}{\partial \beta_j} = -\{\Sigma^{-1}\beta\}_j + (x_i^T\psi)\{\Sigma^{-1}(y_i - \mu_i)\}_j + (x_i^T\beta)\{\Sigma^{-1}(y_i - \mu_i)\}_j;$$

$$\frac{\partial l_i}{\partial \sigma_j} = -\sigma_j(\Sigma^{-1})_{jj} + \sigma_j\{\Sigma^{-1}(y_i - \mu_i)\}_j^2$$

These formulae are useful for numerical maximization routines that require user-supplied derivatives and for constructing a covariance matrix for the parameter estimates.

3.3 Goodness of fit

Various assessments of fit can be made. For binary and categorical y-variables, the estimated probabilities, $p(y_i \mid x_i, \hat{\phi}, \hat{\psi})$ can be compared with the observed y_i-values; this is more useful if the observations can be grouped so that estimated probabilities are compared with sample proportions. For continuous y-variables, $p(y_i \mid x_i, \hat{\phi}, \hat{\psi})$ can be used to construct uniform residuals via the probability integral transform, for example,

and then a corresponding q–q plot can be inspected. For normal y-variables Pearson residuals can be constructed: for example, for the jth component,

$$r_{ij} = \frac{y_{ij} - \mathrm{E}(y_{ij} \mid x_i, \hat{\phi})}{\sqrt{\mathrm{var}(y_{ij} \mid x_i, \hat{\phi})}} = \frac{y_{ij} - \hat{\alpha}_j - \hat{\beta}_j(x_i^{\mathrm{T}}\hat{\psi})}{\sqrt{\hat{\beta}_j^2 + \hat{\sigma}_j^2}}$$

These can be plotted in the usual ways. However, such plots should be interpreted in the light of the comment above, that x might not be expected to be a good predictor of y in this model.

3.4 Prediction of q

Given x and y for an individual, a natural prediction for q is $\mathrm{E}(q \mid x, y)$. This is a function of the parameters, (ϕ, ψ) and, routinely, estimates for them can be inserted. In addition, $\mathrm{var}(q \mid x, y)$ can be computed to provide some assessment of variablity, in this case to be interpreted as the quality of prediction. In the Appendix the computation of conditional q-moments, $\mathrm{E}(q^k \mid x, y)$, is described, from which the required mean and variance can be obtained. The use of these measures will be illustrated in the example given below.

The insertion of estimates into parametric functions, informally known as 'plug-in' methods, is not always ideal for prediction. Use of the maximum likelihood estimates, for example, ignores the rest of the likelihood function. One can try to take account of estimation error, but that usually relies on asymptotic approximations, replacing the actual log-likelihood surface by a quadratic function fitted around its peak. A more inclusive approach is to evaluate the Bayesian posterior distribution of q and then, if desired, extract point quantities, such as the mean, from it. However, one must then propose a suitable prior distribution or, at least, show that the choice of prior has little effect. Also, the computational cost (employing Markov chain Monte Carlo, for instance) can be high when evaluation of the likelihood function itself is costly, as it is for the present class of models. This warrants a deeper, more comprehensive, study and we do not pursue that here.

4 Example

To illustrate the ideas described above, we have applied them to a public-domain credit-scoring data set given in Gana and Rossi (1997). These data describe 95 cases selected from Freddie Mac's loan evaluation database for the period 1993–1994; Freddie Mac is a large US mortgage provider. Information is available on six variables, to which we applied appropriate transformations. These variables were not collected for the purpose of building a model of the form described above, and one might question our partition into x and y sets, but they will serve for illustration. The primary variables are taken as

$x_1 = (\text{ratio of loan amount to value of home}) - 0.8,$
$x_2 = (\text{income score of borrower} - 4)/10,$

$x_3 = x_2^2$,
$x_4 =$ (binary variable) $- 0.2$

(the binary variable is an indicator, taking value 1 if the home is in California, 0 otherwise). The behavioural variables are taken as

$y_1 =$ transformed version of a credit score calculated by the Fair–Isaac company (the normalizing transformation applied to the score is $e^{5y/1000}/100$),
$y_2 =$ binary response variable
(taking value 1 if the loan has ever been 90 days past due date during a specified time period, and 0 otherwise).

For $p(y \mid q)$ we take the components as

$$p(y_1 \mid q) = \phi(y_1; \alpha_1 + \beta_1 q, \sigma_1^2),$$
$$p(y_2 \mid q) = \frac{\exp\{(\alpha_2 + \beta_2 q)y_2\}}{1 + \exp(\alpha_2 + \beta_2 q)}$$

That is, $y_1 \mid q$ has distribution $N(\alpha_1 + \beta_1 q, \sigma_1^2)$ and $y_2 \mid q$ is a binomial$(1, \pi)$ variate with $\text{logit}(\pi) = \alpha_2 + \beta_2 q$. The parameters comprise $\psi = (\psi_1, \ldots, \psi_4)$ and $\phi = (\alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1)$.

The result of fitting the model to the data is as follows. Maximum likelihood estimates of the parameters are obtained as

$$\hat{\psi} = (-3.05, 0.48, -3.90, -0.33)$$

with standard errors (1.11, 1.20, 7.81, 0.45), and

$$\left(\hat{\alpha}_1, \frac{\hat{\alpha}_2}{10}, \hat{\beta}_1, \frac{\hat{\beta}_2}{10}, \log\hat{\sigma}_1\right) = (0.36, -1.01, 0.07, -2.30, -2.61)$$

with standard errors (0.01, 6.25, 0.01, 14.68, 0.12); the 10^{-1} scaling of $\hat{\alpha}_2$ and $\hat{\beta}_2$ is just for numerical balance in the computations. The standard errors have been estimated from the score functions as described in the Appendix. Some of these are very large, reflecting a rather flat likelihood with parameter estimates not well defined. Variable-selection methods could be applied in the usual ways, but, for our purpose, the criterion for selection should relate to prediction of q. With the constraint $\psi_1 < 0$ imposed, the signs of the $\hat{\beta}$s seem sensible: as q increases, y_1 tends to increase and $P(y_2 = 1)$ decreases.

In Figure 1(a) a normal plot of the standardized residuals for y_1 is shown, reflecting the expected 45° line fairly well. It should be said, however, that the plot only reflects the overall distribution of residuals, which results from the normalizing transformation made, and does not necessarily imply that y_1 is well predicted from x. The dots in Figure 1(b) are 9-point moving averages of the observed binary y_2-values plotted against the

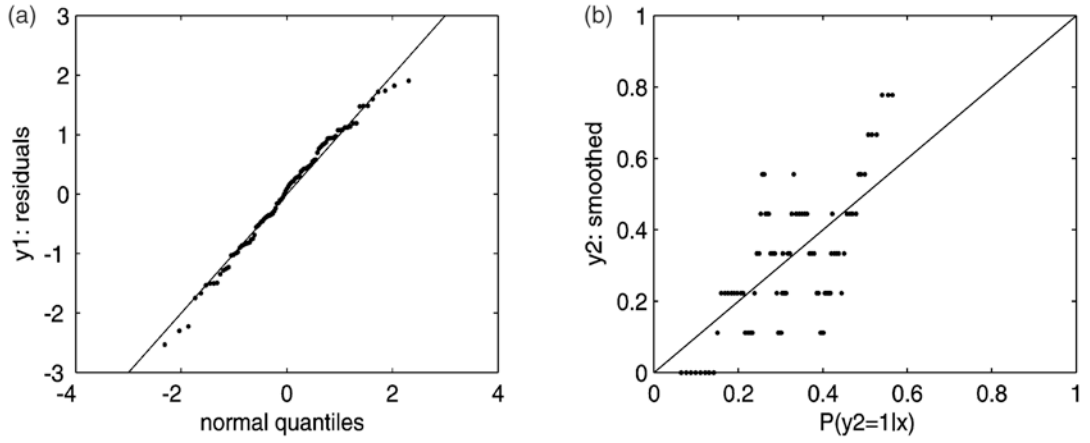

Figure 1 Plots of (a) standardized y_1-residuals, (b) smoothed y_2-values versus fitted $P(y_2 = 1 \mid x)$

corresponding estimates of $P(y_2 = 1 \mid x)$. The 9-point averaging is just a convenient way of smoothing the binary y_2-values; it is this that causes the dots to be arrayed in horizontal bands. Although the overall trend is correct, the plot confirms that individual predictions of y_2 from x are poor.

Figure 2 shows the distributions of (a) $E(q \mid x, y)$ and (b) $\text{var}(q \mid x, y)$ for the 95 sample cases evaluated at $(\hat{\phi}, \hat{\psi})$. In (a) the separation of $E(q \mid x, y)$ into two groups is striking. Upon investigation, it turns out that the left-hand group corresponds to cases with $y_2 = 1$ and the right-hand group to cases with $y_2 = 0$. It seems that, for this particular data set, y_2 alone is a strong determinator of q. Also striking, in (b), is that, whereas $\text{var}(q \mid x) = 1$ by constraint of the model, the sample values of $\text{var}(q \mid x, y)$ tend to be considerably smaller; evidently, the inclusion of y, in addition to x, in the way it is done in the model, substantially reduces the variance of the q-estimate.

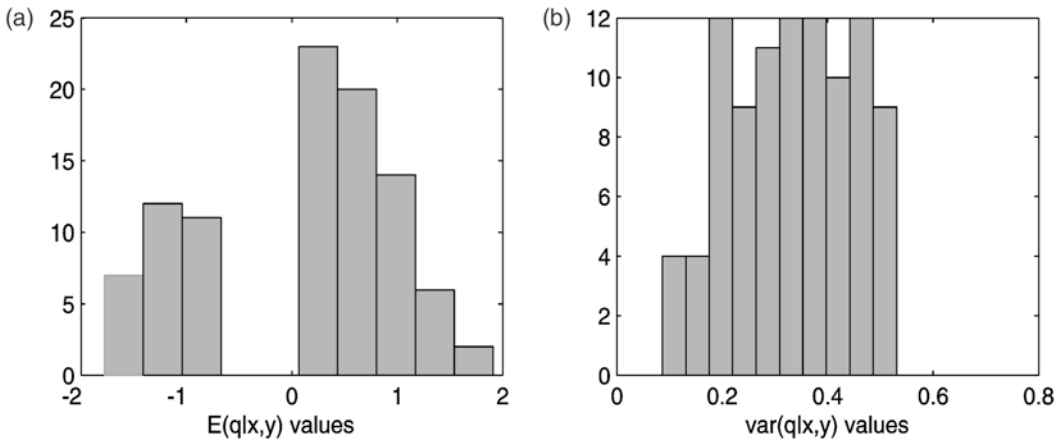

Figure 2 Histograms of (a) $E(q \mid x, y)$ and (b) $\text{var}(q \mid x, y)$ for sample cases at $(\hat{\phi}, \hat{\psi})$

5 Discussion

Financial institutions and governments across the world are concerned about the rise in consumer debt. For example, credit card debt in the UK has increased, on average, at a rate of about 11% per year since Barclaycard, the first UK credit card, was launched in 1966. Further dramatic figures are given in the review papers cited in Section 1. The potential instability arising from this, with the threat of large numbers of customers defaulting if economic conditions worsen, means that financial institutions make extensive use of predictive statistical models to choose customers and monitor their behaviour. However, almost all of these models are of a fairly rudimentary kind, not based on any theory of how customers behave, and constructed using a retrospective database. One or two moves have been made towards less restrictive types of models: for example, graphical models (Hand *et al.*, 1997; Sewart and Whittaker, 1998; Stanghellini *et al.*, 1999), but these appear not yet to have been adopted in practice.

We have developed the application of an alternative class of models for customer quality assessment. Instead of trying to predict a particular, narrowly defined, outcome measure, the model takes both primary characteristics (x) and behavioural characteristics (y) and seeks to summarize them into an overall measure of financial responsibility or quality. In particular, the model uses latent-variable ideas from psychometrics and social measurement, but extends these to include multiple-outcome variables. This model is thus a step away from ones that are purely descriptive of the data, towards ones based on a theory, albeit a weak one, of what the variables mean. More elaborate theories would require psychological modelling of consumer behaviour.

There are legal restrictions on what variables may be used in certain classes of scorecards; for example, most countries prohibit the use of sex in scorecards assessing applications for loans. The same restrictions apply here. Another important requirement of such scorecards in certain applications is that they should be 'interpretable', in the sense that some sort of explanation for the decision can be given. This is the case, for example, in 'front-end' scorecards concerned with initial loan decisions that must be reported back to the customer; it is not the case with 'back-end' scorecards, such as those concerned with attrition or fraud detection, where models can be as elaborate as one likes, for example, neural networks, rule-systems, or elaborate classification trees. This notion of 'interpretability' is very much a product-centred notion: it applies when people are applying for a particular financial product. It is less relevant in our context, where we are concerned not so much with making decisions about a specific product but with more general decisions about the customer. That is, the quality measure provides a more general measure of whether a customer is broadly attractive, that is, the sort of customer one wants to have on one's books and is prepared to make an effort to retain.

When using standard scorecards a threshold must be chosen above which an applicant will be granted the product and below which they will not. Such thresholds are based on various criteria, including estimates of the expected delinquency rate among those accepted, the overall proportion of applicants accepted, or, in more recent scorecards, the estimated future profitability of the customers. The reviews listed

above describe such issues. The measure described in this paper is not primarily intended to be used to make such narrow decisions, but rather to provide additional information about the customer to feed into customer-management programs. However, there is no reason for it not to be used for straightforward accept/reject decisions. For this, the relationship between the quality score and the outcome of interest would need to be evaluated, and then a q-threshold could be chosen on a basis similar to that used for standard scorecards.

Various extensions to the model are possible. An obvious one, in keeping with the psychometric aspect of the model, is to permit q to have more than one component. We have avoided this here because a single component has interpretative advantages. While a better fit to data can doubtless be achieved by using more components, it is important that the model should appeal to potential users who do not have a deep understanding of statistical modelling. One can think of q as the single latent variable that provides the best fit in some sense. The model yields the conditional distribution of q given x and y; from this, predicted point-values of q can be obtained, for example, the mean, which is no more complicated than the output of the simple predictive models currently used across the retail banking sector. The use of a single-component q is in line with standard random-effects and frailty modelling. A second obvious extension would be to let q vary over time, in a state-space model. This would require the use of variables that can be repeatedly measured over time, as is the case with some behavioural variables. In our initial investigations we simply try to capture some relatively static aspect of customer quality, or just its value at the time of application.

Acknowledgement

We thank the editor and referees for their careful and detailed comments.

References

Abramowitz M and Stegun IA (1965) *Handbook of mathematical functions*. New York: Dover.

Bartholomew DJ and Knott M (1999) *Latent variable models and factor analysis*, 2nd ed. London: Arnold.

Bollen KA (1989) *Structural equations with latent variables*. New York: Wiley.

Crook J, Banasik J and Thomas L (2001) Sample selection bias in credit scoring. Presented at *Credit Scoring and Credit Control VII*, Management School, University of Edinburgh.

Crouch EAC and Spiegelman D (1990) The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(x)\mathrm{e}^{-x^2}dx$: application to logistic-normal models. *Journal of the American Statistical Association*, **85**, 464–69.

Gana R and Rossi CV (1997) An empirical comparison of linear probability and logit models of mortgage default. *Fifth Conference on Credit Scoring and Credit Control*, Credit Research Centre, University of Edinburgh.

Goodwin ET (1949) The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(x)\mathrm{e}^{-x^2}dx$. *Proceedings of the Cambridge Philosophical Society*, **45**, 241–45.

Hand DJ (2001a) Modelling consumer credit risk. *IMA Journal of Management Mathematics*, **12**, 139–55.

Hand DJ (2001b) Reject inference in credit operations. In Mays E ed. *Handbook of credit scoring*. Chicago: Glenlake Publishing, 225–40.

Hand DJ and Henley WE (1993) Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry*, **5**, 45–55.
Hand DJ and Henley WE (1997) Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society A*, **160**, 523–41.
Hand DJ, McConway KJ and Stanghellini E (1997) Graphical models of applicants for credit. *IMA Journal of Mathematics Applied in Business and Industry*, **8**, 143–55.
Hand DJ, Li HG and Adams NM (2001) Supervised classification with structured class definitions. *Computational Statistics and Data Analysis*, **36**, 209–25.
Joreskog KG and Goldberger AS (1975) Estimation of a model with multiple indicators and multiple causes of a single latent variable model. *Journal of the American Statistical Association*, **70**, 631–39.
McNab H and Wynn A (2000) *Principles and practice of consumer credit management*. Canterbury: CIB Publishing.
Press WH, Flannery BP, Teukolsky SA and Vetterling WT (1988) *Numerical recipes in C*. Cambridge: CUP.
Rosenberg E and Gleit A (1994) Quantitative methods in credit management: a survey. *Operations Research*, **42**, 589–613.
Sammel MD, Ryan LM and Legler JM (1997) Latent variable methods for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society B*, **59**, 667–78.
Sewart P and Whittaker J (1998) Fitting graphical models to credit scoring data. *IMA Journal of Mathematics Applied in Business and Industry*, **9**, 241–66.
Skrondal A and Rabe-Hesketh S (2004) *Generalized latent variable modelling*. London: Chapman and Hall/CRC.
Stanghellini E, McConway KJ and Hand DJ (1999) A chain graph for applicants for bank credit. *Applied Statistics*, **48**, 239–51.
Thomas LC (2000) A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, **16**, 149–72.
Thomas LC, Edelman DB and Crook JN (2002) *Credit scoring and its applications*. Philadelphia: SIAM.

Appendix

The log-likelihood function, $l(\phi, \psi) = \sum_{i=1}^{n} l_i(\phi, \psi)$, was maximized numerically using a quasi-Newton method (Press *et al.*, 1988: Section 10.7). For this, a function is required that returns the value of $l(\phi, \psi)$ on call for any given (ϕ, ψ). To compute the individual log-likelihood contributions, $l_i(\phi, \psi)$, we need to evaluate integrals of the form

$$\int_{-\infty}^{\infty} p(y \mid q, \phi) p(q \mid x, \psi)\, \mathrm{d}q$$

Since $p(q \mid x, \psi)$ here has the form $\phi(q; x^{\mathrm{T}}\psi, 1)$, the substitution $z = (q - x^{\mathrm{T}}\psi)/\sqrt{2}$ produces

$$\pi^{-1/2} \int_{-\infty}^{\infty} g_k(z) \mathrm{e}^{-z^2}\, \mathrm{d}z$$

where $g_k(z) = p(y \mid q_z, \phi)$ with $q_z = x^{\mathrm{T}}\psi + z\sqrt{2}$. The integral can now be approximated by Gauss–Hermite quadrature (Press *et al.*, 1988: Section 4.5) as $\sum_{j=1}^{m} w_j g_k(z_j)$, where the z_j are the zeros of the mth order Hermite polynomial and the w_j are the corresponding weights.

We are grateful to the editor for drawing our attention to the paper by Crouch and Spiegelman (1990), based on the work of Goodwin (1949), and to Skrondal and Rabe-Hesketh (2004: Section 6.3.2). It turns out that Gauss–Hermite quadrature, even with 20 roots (the maximum number listed in Abramowitz and Stegun, 1965), can be insufficiently accurate for our purposes. Thus, we changed to an adaptive trapezium rule (Press *et al.*, 1988: Section 4.2). After some experimentation, we found that a sufficiently good approximation to the integral can be constructed provided that one first computes a suitable range of integration as follows. For the case here, where y consists only of normal and discrete components, we have

$$g_k(z)\mathrm{e}^{-z^2} = \mathrm{e}^{-z^2}\left\{\prod \phi(y_j^{(1)}; \alpha_j + \beta_j q_z, \sigma_j^2)\right\} p(y^{(2)} \mid q_z)$$

where the $y_j^{(1)}$ are the normally distributed components of y and $y^{(2)}$ is the vector of discrete components. Now, $\mid p(y^{(2)} \mid q_z) \mid \leq 1$ and the other factor can be manipulated into the form

$$\left\{\prod (\sigma_j\sqrt{2\pi})^{-1}\right\} \exp\left\{(1+s_1)(z-z_0)^2 + \left(s_3 - \frac{s_2^2}{1+s_1}\right)\right\}$$

in which

$$s_1 = \frac{\sum \beta_j^2}{\sigma_j^2},\ s_2 = \frac{\sum \beta_j u_j}{\sigma_j^2\sqrt{2}},\ s_3 = \frac{\sum u_j^2}{2\sigma_j^2},$$
$$u_j = \alpha_j + \beta_j(x^{\mathrm{T}}\psi),\ z_0 = \frac{s_2}{1+s_1}$$

Thus, this factor has a peak at $z = z_0$ and falls to ϵ times the peak value at $z = z_0 \pm z_1$, where

$$z_1 = \frac{\{-\log\epsilon - (s_3 - (s_2^2/(1+s_1)))\}}{1+s_1}$$

The general shape of the factor, as a function of z, is that of the normal density. We have used $(z_0 - z_1, z_0 + z_1)$ as the range for the adaptive trapezium integration, with the various tuning parameters (such as ϵ) chosen as a result of numerical trials. A more elaborate version could be implemented, to take into account the variation of $p(y^{(2)} \mid q_z)$ with z, but this has proved not to be necessary for our application. The approximations were validated against a trapezium rule with a very wide range and a very fine grid.

In Section 3.4 we based predictions of q on its conditional mean and variance given x and y. Thus, we need to compute moments $\mathrm{E}(q^k \mid x, y, \phi, \psi)$ for $k = 1$ and 2. But this is equal to

$$\int_{-\infty}^{\infty} q^k p(q \mid x, y, \phi, \psi)\,\mathrm{d}q = \int_{-\infty}^{\infty} q^k p(y \mid q, \phi) p(q \mid x, \psi)\,\mathrm{d}q \div p(y \mid x, \phi, \psi)$$

The numerator here is like the integral above, but with $q_z^k g_k(z)$ in place of $g_k(z)$, and similar considerations apply for its numerical evaluation; the denominator is just the likelihood function again.

Standard errors of the estimates, $\hat{\phi}$ and $\hat{\psi}$, may be obtained routinely from the sample information matrix: according to regular asymptotic theory, an estimate of $\mathrm{cov}(\hat{\phi}, \hat{\psi})$ is provided by $-l''(\hat{\phi}, \hat{\psi})^{-1}$. Alternatively, when the data comprise n independent observations, one can use the individual score vectors, $l_i'(\phi, \psi)$: the corresponding estimate of $\mathrm{cov}(\hat{\phi}, \hat{\psi})$ is then $\{\sum_{i=1}^{n} l_i'(\hat{\phi}, \hat{\psi}) l_i'(\hat{\phi}, \hat{\psi})^{\mathrm{T}}\}^{-1}$. The latter method is often preferable because it yields a matrix to be inverted that is at least positive semi-definite. The former method does not carry such a guarantee, usually due to numerical rounding errors, particularly when the likelihood surface at $(\hat{\phi}, \hat{\psi})$ is very flat. Numerical evaluation of the score functions was performed by differencing. For example, $\partial l_i / \partial \phi_j$ can be estimated by $\delta^{-1}\{l_i(\phi + \delta u_j, \psi) - l_i(\phi, \psi)\}$, where u_j is the jth row of the unit matrix of size $\dim(\phi)$ and δ is an appropriately small increment. The information matrix can be obtained by differencing again. An alternative method would be to calculate the required derivatives algebraically and then write extra computer code for the resulting expressions.