Available online at www.sciencedirect.com

ScienceDirect

European Journal of Operational Research 183 (2007) 1582–1594

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/ejor

Reject inference, augmentation, and sample selection

John Banasik [*], Jonathan Crook

Credit Research Centre, Management School and Economics, WRB 50 George Square, University of Edinburgh, Edinburgh EH8 9JY, UK

Received 1 January 2006; accepted 1 June 2006
Available online 25 January 2007

Abstract

Many researchers see the need for reject inference in credit scoring models to come from a sample selection problem whereby a missing variable results in omitted variable bias. Alternatively, practitioners often see the problem as one of missing data where the relationship in the new model is biased because the behaviour of the omitted cases differs from that of those who make up the sample for a new model. To attempt to correct for this, differential weights are applied to the new cases. The aim of this paper is to see if the use of both a Heckman style sample selection model and the use of sampling weights, *together*, will improve predictive performance compared with either technique used alone. This paper will use a sample of applicants in which virtually every applicant was accepted. This allows us to compare the actual performance of each model with the performance of models which are based only on accepted cases.

Keywords: Risk analysis; Credit scoring; Reject inference; Augmentation; Sample selection

1. Introduction

Those who build and apply credit scoring models are often concerned about the fact that these models are typically designed and calibrated on the basis only of those applicants who were previously considered adequately creditworthy to have been granted credit. The ability of such models to distinguish good prospects from bad requires the inclusion of delinquent credit payers in the data base. Such delinquent applicants are unlikely to have characteristics that differ radically from good applicants, yet the ability to discern those differences is the critical feature of a good model. Reject inference is a term that distinguishes attempts to correct models in view of the characteristics of rejected applicants.

Augmentation and sample selection offer potentially complementary corrections for model deficiencies that arise from the omission of rejected applicants from data bases used to build credit scoring models. Both implicitly acknowledge model deficiency arising from the unavailability of the repayment behaviour of rejected applicants. Sample selection correction may be thought of as correction for *variables* denied the model on account of rejected cases. For example, a variable may affect the probability that a case is accepted but is not included in the new model to replace it. Augmentation may be thought of as correcting for other aspects of model misspecification arising out of missing *cases*, particularly those having to do with

[*] Corresponding author.
E-mail addresses: John.Banasik@ed.ac.uk (J. Banasik), J.Crook@ed.ac.uk (J. Crook).

0377-2217/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2006.06.072

the model's functional form. For example, a linear function of some variable may quite adequately describe repayment prospects over the range of that variable observed among accepted applicants, but a hint of curvature among the less reliable applicants may seem inadequate for reliable modelling. This paper considers whether both corrections may be used simultaneously and entertains the possibility that each correction may be enhanced in the presence of the other.

Banasik et al. (2003) considered the efficacy of sample selection correction using a bivariate probit model on the basis of a rare sample where virtually all applicants were accepted. Applicants were nevertheless distinguished as to whether they would normally be accepted, so that the performance of models based on all applicants could be compared with those based only on accepted applicants. This provides a basis for discerning the scope for reject inference techniques. That paper reported distinct but modest scope for reject inference, and that the bivariate probit model achieved only a slight amount of it. Subsequent experiments using the same sample with augmentation are reported in Crook and Banasik (2004) and Banasik and Crook (2005). These suggested that augmentation actually undermined predictive performance of credit scoring models. In the discussion that follows these results are revisited in experiments slightly revised to enhance comparability and are compared with results arising from joint deployment of the two techniques. After explaining both techniques, the character of the data and its adaptation for its present application will be discussed. Then the results of the techniques used in isolation and then together will be reported.

2. Sample selection

A large literature has developed on missing data mechanisms; see for example Smith and Elkan (2003) for a Bayesian belief network approach. In this section, we concentrate only on those types of missing data mechanisms that are relevant to reject inference and which were proposed by Little and Rubin (1987). Let $D_i = 1$ if a borrower i defaults and $D_i = 0$ if he/she repays on schedule. Let $A_i = 1$ indicate that case i was accepted in the past and $A_i = 0$ if that case was not accepted. Let D_{obs} denote the values of D for cases where the repayment performance is observed, that is for cases where $A_i = 1$, and let D_{miss} denote values of D for cases where repayment performance is missing, that is for cases where $A_i = 0$. Little and Rubin classify missing mechanisms into three categories, two of which are relevant in this context (Hand and Henley, 1994). These are as follows.

2.1. Missing at random (MAR)

This occurs if

$$P(A|D_{\text{obs}}, D_{\text{miss}}, \phi) = P(A|D_{\text{obs}}, \phi), \tag{1}$$

where ϕ is the vector of parameters of the missing data mechanism. This can be written

$$P(A|D, X_2) = P(A|X_2), \tag{2}$$

where X_2 is a set of variables that will be used to model $P(A)$. The probability that an applicant is rejected (and his repayment performance is missing), given values of X_2, does not depend on his repayment performance. Since we are interested in $P(D|X_1)$ where X_1 is a set of variables that will be used to model $P(D)$ we note that Eqs. (1) and (2) are equivalent to

$$P(D|X_1, A = 1) = P(D|X_1). \tag{3}$$

The parameters we estimate from a posterior probability model (for example logistic regression) using the accepted cases only are unbiased estimates of the parameters of the population model for all cases, not merely for the accepts, *assuming the same model applies to all cases*. However, since the parameter estimates are based only on a sub sample their estimated values may be *inefficient*.

2.2. Missing not at random (MNAR)

This occurs if $P(A_i)$ is not independent of D_{miss} so D_{miss} cannot be omitted from $P(A|D_{\text{obs}}, D_{\text{miss}}, \phi)$ so

$$P(A|D_{\text{obs}}, D_{\text{miss}}, \phi) \neq P(A|D_{\text{obs}}, \phi). \tag{4}$$

This implies D cannot be omitted from $P(A|D, X_2)$ so

$$P(A|D, X_2) \neq P(A|X_2). \tag{5}$$

The probability that an application is rejected, given values of X_2, depends on repayment performance.

In MNAR we cannot deduce Eq. (3). To see this write

$$P(D|X_1) = P(D|X_1, A = 1) \cdot P(A = 1|X_1) + P(D|X_1, A = 0) \cdot P(A = 0|X_1). \tag{6}$$

Since in MNAR

$$P(D|X_1, A = 1) \neq P(D|X_1, A = 0), \tag{7}$$

$$P(D|X_1) \neq P(D|X_1, A = 1). \tag{8}$$

To parameterise $P(D|X_1)$ we must model the process which generates the missing data as well. If we do not, the estimated parameters of $P(D|X_1)$ are biased. An example of such a procedure is Heckman's model (Heckman, 1976) which, if D were continuous and the residuals normally distributed, would yield consistent estimates. A more appropriate model is that of Meng and Schmidt (1985) where $P(D|X_1)$ is modelled rather the $E(D|X_1)$, again assuming normally distributed residuals. The Meng and Schmidt model is the 'bivariate probit model with sample selection' (BVP).

To proceed further it is efficient to set up the scoring problem as follows:

$$d_i^* = f_1(X_{i1}, \varepsilon_{i1}), \tag{9}$$

$$a_i^* = f_2(X_{i2}, \varepsilon_{i2}), \tag{10}$$

where d_i^* is a continuous random variable describing the degree of default such that when $d_i^* \geqslant 0, D_i = 1$ and when $d_i^* < 0, D_i = 0$. a_i^* is a continuous random variable such that when $a_i^* \geqslant 0$, $A_i = 1$ and D_i is observed, and when $a_i^* < 0$, $A_i = 0$ and D_i is unobserved. We wish to parameterise $P(D_i)$. If we further assume $E(\varepsilon_{i1}) = E(\varepsilon_{i2}) = 0$, $\text{cov}(\varepsilon_{i1}, \varepsilon_{i2}) = \rho$, $(\varepsilon_{i1,}\varepsilon_{i2}) \sim$ bivariate normal then we have the BVP model.

Now consider various cases.

Case 1. Model 10 fits the data to be used to parameterise the new model *perfectly*. For example, in the past, the bank followed a scoring rule precisely for every applicant. Here $\varepsilon_{i2} = 0$ and so $\rho_{\varepsilon_{i1},\varepsilon_{i2}} = 0$ for all cases. Is this MAR? This depends on whether, given X_1, $P(D_i)$ in the population depends on whether the case is observed. Here we can consider two sub cases.

Case 1a. Suppose there are variables in X_2, which are excluded from X_1 but which affect $P(D_i)$. Then Eq. (7) holds and we have MNAR. If $P(D_i)$, given X_1, does not differ between the observed and missing cases, we have MAR. In the credit scoring context variables which are correlated with $P(A_i)$ and which may be in the X_2 set, but not in the X_1 set, include the possession of a County Court Judgement (CCJ). An applicant with a CCJ may be rejected so the possession of a CCJ does not appear in X_1 for the purpose of estimation. Notice that in this case the Meng and Schmidt Heckman-type model (BVP) will not make the estimated parameters more consistent than a single equation model because the source of the inconsistency that the BVP model corrects for occurs only when $\rho_{\varepsilon i1,\varepsilon i2} \neq 0$.

Case 1b. Here there is no variable in X_2 which is omitted from X_1 and which causes $P(D_i)$, given X_1, to differ between the observed and missing cases. We have MAR, not MNAR.

Case 2. Now suppose Eq. (10) does not perfectly fit the data to be used to parameterise the new model. This may occur because variables additional to those in X_2 were used to predict $P(A_i)$. In the credit scoring context such variables include those used to override the values of A_i predicted by the original scoring model. Again consider sub cases.

Case 2a. Suppose these additional variables are (a) *not included* in X_1 and (b) do affect $P(D_i)$. Then Eq. (7) holds and we have MNAR. Also, given (a) and (b) and that these variables (c) are *not included* in X_2, but (d) do affect $P(A_i)$, $\rho_{\varepsilon i1,\varepsilon i2}$ may not equal zero. In this case the BVP approach may yield consistent parameters for Eq. (9) which will not be given by a single equation model.

Case 2b. Suppose the additional variables referred to in Case 2a are (a) *included* in X_1 and (b) affect $P(D_i)$. Then Eq. (3) holds instead of Eq. (7) and we have MAR, not MNAR. Further, $\rho_{\varepsilon i1,\varepsilon i2} = 0$ and the BVP model does not yield more consistent estimates than a single equation posterior probability model.

Case 2c. In this case, the additional variables are (a) *included* in X_1 and (b) do not affect $P(D_i)$. Again Eq. (3) holds instead of Eq. (7) we have MAR not MNAR, $\rho_{\varepsilon i1,\varepsilon i1} = 0$ and the BVP model does not increase the consistency of the parameter estimates.

In short, if the assumptions of the BVP hold the technique will increase the efficiency of the estimated parameters over that achieved in a single model posterior probability model only in Case 2a.

It is worth noting, that apart from augmentation, to be described in the next section, the literature contains experiments to assess the performance of a small number of other algorithms to estimate application scoring models in the presence of rejected cases. One example is the EM algorithm (Feelders, 2000). However, implementations of the EM algorithm, like those of other imputation techniques such as Markov chain Monte Carlo methods (MCMC), have typically (but not always) assumed the missing mechanism is MAR rather than MNAR (see Schafer, 1997). In addition, the application of these techniques has been either on simulated data, which may miss the data structures typical of credit application data, or on data which does not allow a meaningful benchmark all-applicant model to be estimated.

3. Augmentation

Augmentation is a well-used technique that involves weighting accepted applicants in such a way as to synthesize a sample that fully represents rejected applicants. Its use involves tacit admission of model inadequacy whereby no single parameter set governs all applicants. Fig. 1 illustrates this intuitively by revisiting some basic principles of linear regression analysis, assuming the prevalence of a linear relationship. Part (a) suggests that extreme values in the range of an explanatory variable minimize the standard errors of the estimated parameters, but often this sample range is not a discretionary matter. Should it be restricted as in part (b) as is potentially the case for characteristics observed among accepted credit applicants, then one must be satisfied with the line estimated by those points as the best available. To weight sample observations to reflect better the mean of the explanatory variable within the general population as in part (c) is effectively to cluster observations and thereby to sacrifice *efficiency*. There was no *bias* to reduce in the first place and none after the weighting, but more *error* in the model parameters estimates probably attends such weighting. Obviously, one would not indulge in this weighting were linearity to be believed.

Fig. 2 illustrates a non-linear situation modelled linearly. Part (a) makes clear that available data do not support the discernment of curvature. Part (b) illustrates the effect of estimating with weights, presuming the presence of curvature. That might seem sensible in the credit scoring context, since the ranking of marginal applicants deserves special attention. This special concentration on marginal applicants depends on the benefits of exploiting curvature exceeding the loss of efficiency that comes

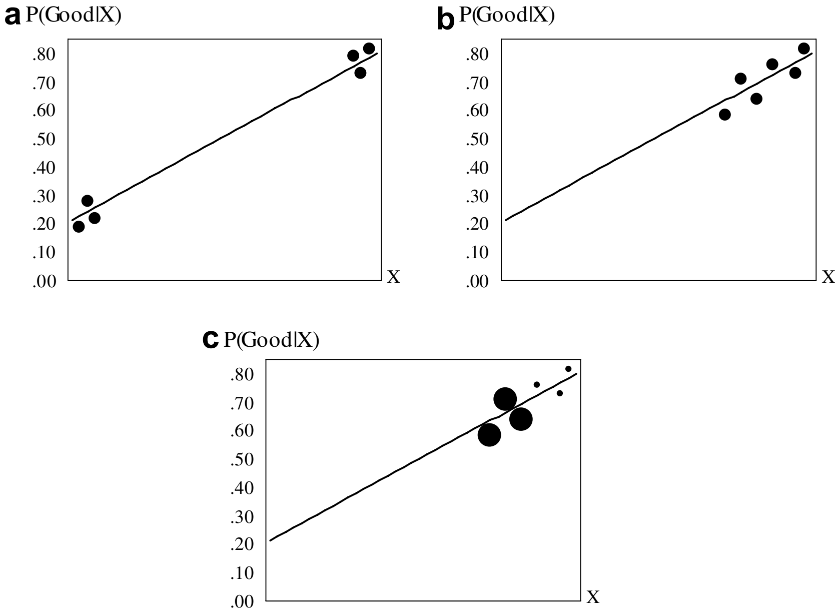

Fig. 1. Illustration of estimation scenarios for a linear relationship: (a) estimation with extreme *X* spread; (b) estimation with restricted *X* range; (c) estimation with weighting to reflect character of missing observations.

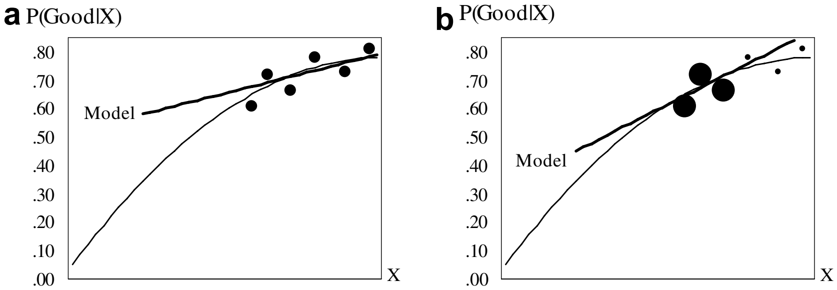

Fig. 2. Illustration of weighting to characterize different *X* ranges: (a) slope estimated over observed range; (b) slope simulates inclusion of lower *X* range.

from effectively clustering attention on a narrow range of observations.

The derivation of weighting used in the variant of augmentation deployed here was explained in Crook and Banasik (2004). In brief, it requires first the estimation of an Accept–Reject (AR) model that predicts the probability that any applicant will be among those accepted in a population. The inverse of the estimated probability equals the number of cases each accepted case in the sample represents and can be regarded as a sampling weight in the estimation of the GB model. Those accepts which have relatively low probabilities of acceptance will have relatively high weights, and since their probabilities are relatively low they may be expected to have characteristics more similar to those cases that were originally rejected than to cases which have a high probability of acceptance. Accordingly, a Good–Bad (GB) model may be estimated weighting each accepted case by the inverse of its probability arising out of the AR model. That should provide the GB model with much of the character it would have were the repayment behaviour of rejected applicants to be known and included.

Notice that since augmentation is not correcting for the possible validity of Eq. (7) it is not correcting for a missing mechanism which is MNAR. Instead it assumes the mechanism is MAR.

A couple of caveats deserve particular note in the present context of considering both sample selection and augmentation together. First, as explained above bias from omitted variables will occur (MNAR) unless the variable set of the GB model encompasses that of the AR model. However, in the analysis that follows both the AR and GB models are estimated with some variables denied the other. This permits comparable results for augmentation and sample selection, since the exclusive resort of the AR model to certain explanatory variables in sample selection is a vital feature of sample selection.[1] Secondly, augmentation is not feasible in Case 1 above, where the AR process can be modelled perfectly. Even were the probit or logistic regression equation to be estimable, it would generate unit probabilities for all accepted cases and hence undefined weights. This ability of perfect knowledge about the AR process to scuttle reject inference is a paradoxical feature augmentation shares with sample selection. As a practical matter the AR process generally depends on exclusive resort to some variables, or there are overrides[2] (a particular instance of a missing variable) in its model's application.

4. Banded data methodology

The sample available for the present analysis had virtually no rejected applicants but it did have an

[1] The present analysis differs from that presented in Crook and Banasik (2004) and in Banasik and Crook (2005) where the GB model variable set was used for the AR *model* in spite of awareness that the AR *process* depended on exclusive resort to some additional variables. In any case, an attempt to avoid bias altogether seems a vain endeavour, since augmentation is only ever reasonably used when the GB model is presumed to suffer from misspecification bias hidden by the absence of rejected applicants.

[2] An override is a case for which one or more variables additional to those in the scoring model have been used to make a decision. Of course the inclusion of such variables into an AR equation which includes merely the variables of the parameterised statistical model would improve the predictive performance of such a model. For example if a person is not on the electoral register because they have recently moved from another country and such a circumstance, when known, increased the chance of acceptance, a variable to represent this period of living abroad should be included.

indication of which applicants would normally be rejected. The credit supplier would occasionally absorb the cost of accepting poor applicants so as to have a data base that would have no need for reject inference. Table 2 demonstrates the large proportion of very poor applicants accepted on such occasions. Unfortunately, this data set indicated no scope for reject inference. Models built only upon those applicants who would normally be accepted predicted repayment behaviour of all applicants every bit as well as models built on all applicants. This probably reflected the normal acceptance threshold which would see two-thirds of applicants accepted of whom nearly 30% were ‘‘bad’’ in the sense used for development of the GB models analysed here. Such applicants were defined as those who had accounts transferred for debt recovery within 12 months of credit first being taken. Evidently models built on such accepted applicants already incorporated insights about the nature of very bad applicants as to make reject inference redundant. The influence of the acceptance threshold in determining the scope for useful application of reject inference thus became a central concern.

The credit provider supplied only the raw data, including good–bad status, and its normal accept–reject decision for each applicant. Except that most relevant variables were provided, little useful information was indicated about the nature of the normal acceptance process, so that shifting the acceptance threshold required fabrication of an acceptance process. More elaborate detail about this fabrication process appears in Banasik et al. (2003). For the present purposes suffice it to say that AR and GB variable sets described in Table 1 were determined from a process of stepwise logistic regressions using relevant dependent variables.

Table 1
Variables included in the Accept–Reject and Good–Bad models

Variable description	Good–Bad model	Accept–Reject model	Coarse categories	Minimum frequency
Time at present address		✓	8	281
B1		✓	4	242
Weeks since last county court judgement (CCJ)		✓	6	244
B2		✓	5	324
B3	✓	✓	6	453
Television area code	✓	✓	5	26
B4	✓	✓	6	496
Age of applicant (years)	✓	✓	6	201
Accommodation type	✓	✓	5	180
Number of children under 16	✓	✓	6	130
P1	✓	✓	3	377
Has telephone	✓	✓	3	1883
P2	✓	✓	6	611
B5	✓	✓	4	239
B6	✓	✓	5	320
P3	✓	✓	4	516
B7	✓		6	1108
B8	✓		6	407
B9	✓		6	1443
Type of bank/building society accounts	✓		6	188
Occupation code	✓		6	129
P4	✓		6	1108
Current electoral roll category	✓		5	458
Years on electoral roll at current address	✓		6	458
B10	✓		6	403
P5	✓		3	379
B11	✓		6	324
B12	✓		4	1163
B13	✓		4	1291
Number of searches in last 6 months	✓		4	406

B*n* = bureau variable *n*; P*n* = proprietary variable *n*; ✓ denotes variable is included.

Table 2
Sample accounting

	All sample case			Good rate (%)	Training sample cases			Hold-out sample cases		
	Good	Bad	Total		Good	Bad	Total	Good	Bad	Total
Cases not cumulated into English acceptance threshold bands to show good rate variety										
Band 1	1725	209	1934	89.2	1150	139	1289	575	70	645
Band 2	1558	375	1933	80.6	1039	250	1289	519	125	644
Band 3	1267	667	1934	65.5	844	445	1289	423	222	645
Band 4	1021	912	1933	52.8	681	608	1289	340	304	644
Band 5	868	1066	1934	44.9	579	711	1290	289	355	644
English	6439	3229	9668	66.6	4293	2153	6446	2146	1076	3222
Scottish	1543	997	2540	60.7						
Total	7982	4226	12,208	65.4						
Cases cumulated into English acceptance threshold bands for analysis										
	English sample cases									
Band 1	1725	209	1934	89.2	1150	139	1289	575	70	645
Band 2	3283	584	3867	84.9	2189	389	2578	1094	195	1289
Band 3	4550	1251	5801	78.4	3033	834	3867	1517	417	1934
Band 4	5571	2163	7734	72.0	3714	1442	5156	1857	721	2578
Band 5	6439	3229	9668	66.6	4293	2153	6446	2146	1076	3222

Normally, an AR model reflects an older GB model that determined the cases available for the new GB model. In fabricating an AR process nationality appeared as a metaphor for time. The GB behaviour of the 2540 Scottish applicants' was modelled using the variables selected for the AR model. Using the AR variable set and parameters calibrated on Scottish applicants, the remaining 9668 English and Welsh (hereafter English) applicants then received AR scores by which they were ranked and banded into five acceptance thresholds. All subsequent modelling would be restricted to English applicants.

English applicants were ranked into five bands of nearly equal size from each of which stratified random sampling determined that training and hold-out samples would have virtually the same good–bad rate. The upper part of Table 2 demonstrates the range of repayment behaviour available in the data with repayment performance in the top band nearly double that in the bottom one. All subsequent analysis uses the data as described in the lower part of Table 2 where each band includes cases in the band above it. Each of these cumulated bands then appears as a distinct potential grouping of accepted applicants. The all-inclusive Band 5 provides the basis for benchmark models against which less inclusive "accepted" applicant samples models – with and without reject inference – may be judged.

The coarse classification used in this analysis was not a feature of the provided data, but reflected preliminary analysis of GB performance over variable intervals, taking account of natural breaks among all applicants and among applicants designated as normally acceptable by the data provider.[3] For the Scottish model that was used to score and rank English applicants the coarse categories for all variables were represented by binary variables. However, both the AR and GB models that were subsequently developed for the English applicants used the weights of evidence approach whereby coarse categories within each variable appeared as specific values in that variable. This switch from a binary variable approach to a weights of evidence approach between the Scottish AR scoring process and the English AR model used to represent it prevents even a nearly perfect fit in the latter model. In spite of resort to the same variables logistic regression provides correct classification for the top four

[3] In Banasik et al. (2003) this classification was used alternatively to define binary variables and weights of evidence, and both approaches gave very similar results for models without reject inference. In this respect, the following analysis of the sample selection procedure in this paper differs from the earlier one. However, on account of collinearity problems, only the weights of evidence were used in this analysis for reject inference. A critical feature of the banding approach was that English applicants were scored using the less restrictive binary variable approach. In that earlier paper two variables were removed from both the AR and GB set in the mistaken presumption that this would be necessary to avoid a nearly perfect fit for the AR model, since the AR scores were simply fitted values using the AR variable set.

bands of only 84–95% of cases. The imperfect fit required to simulate AR model overrides was thus achieved without the need to exclude from the English AR model variables that were used in the Scottish AR process.

5. Model assessment

Classification performance depends on two features of the modelling process: its ability to rank cases and its ability to indicate or at least use an appropriate cut-off point. Overall ranking of applicants in terms of likely repayment performance is interesting, but more critical is the ranking among marginal applicants with repayment prospects that will attract deliberation. Ranking among very good applicants certain to receive credit and among very poor applicants certain to be rejected matters little unless the ranking mechanism is to be used to determine the amount of credit to give to the customer (for example setting the customer's credit limit on a credit card).

The nature of the analysis that follows may be illustrated by interpretation of Table 3 in which the application of a model's parameters estimated by each band's training sample appears. The third column represents classification success where the cut-off has been selected to equate actual and predicted numbers of goods in each band's training sample. The fourth column standardizes the results by using instead the band's hold-out sample to equate these numbers. This slightly illicit resort to the hold-out sample to obtain a parameter estimate affects results very little. The sixth column indicates the usefulness of each band's training sample ranking and cut-off applied to all applicants, including those of all lower bands. Finally, column seven shows how performance of each band's model might be improved in all-applicant prediction were the cut-off that equalizes actual and predicted good performance among the all-applicant hold-out sample to be known. Such would be approximately the case were one to somehow know what proportion of the whole applicant population is bad.

From the standpoint of reject inference two types of comparison are pertinent. First, for each band comparison of the column six result to that column's Band 5 result indicates the scope for improvement by reject inference, since it is the difference that results from availability of repayment performance by all rejected applicants. Secondly, comparison between each band's column six and seven results indicates the benefit to be had by simple awareness of the appropriate cut-off. If this cut-off is known simple modelling with accepted cases can provide this result. Column six demonstrates considerable scope for reject inference in each of the top four rows where the absence of information on rejected applicants can undermine performance. Column seven suggests that the bulk of this improvement could be had simply from awareness of the cut-off implied by knowledge of the repayment behaviour by rejected applicants. For example, the Band 1 scope for benefit from reject inference is 3.48% (i.e. 73.68–70.20) of which 2.36% (i.e. 70.20–72.56) could be obtained by knowledge of the appropriate cut-off point. To that extent one need know only the likely repayment proportion of all applicants and not the particular relationships between attributes of unacceptable applicants and repayment performance.

6. Reject inference results

Joint application of augmentation and the bivariate probit model requires a specified weighting for all cases, accepted *and* rejected alike. For accepted applicants the weights used for simple augmentation were scaled to have an average value of 1.0, the weight assigned to all rejected cases. Thus if the first $0 \ldots n$ cases are accepts and the following $(n+1) \ldots k$ cases are rejects:

Table 3
Classification using simple logistic regression

Predicting model	Own band hold-out prediction			All-applicant hold-out prediction		
	Number of cases (%)	Own band training cut-off (%)	Own band hold-out cut-off (%)	Number of cases (%)	Own band training cut-off (%)	All band hold-out cut-off (%)
Band 1	645	89.30	89.77	3222	70.20	72.56
Band 2	1289	83.40	83.86	3222	70.58	72.75
Band 3	1934	79.21	79.42	3222	71.97	73.49
Band 4	2578	75.37	75.56	3222	72.47	73.81
Band 5	3222	73.68	73.49	3222	73.68	73.49

Table 4
Overall ranking performance by area under ROC

	Own band training sample		Own band hold-out		All-applicant hold-out	
	Number of cases	Area under ROC	Number of cases	Area under ROC	Number of cases	Area under ROC
Simple logistic regression						
Band 1	1289	.8884	645	.8654	3222	.7821
Band 2	2578	.8373	1289	.8249	3222	.7932
Band 3	3867	.8141	1934	.8175	3222	.8009
Band 4	5156	.8003	2578	.8108	3222	.8039
Band 5	6446	.7934	3222	.8049	3222	.8049
Weighted logistic regression (Augmentation)						
Band 1	1289	.8468	645	.8446	3222	.7362
Band 2	2578	.7733	1289	.7647	3222	.7083
Band 3	3867	.7812	1934	.7911	3222	.7808
Band 4	5156	.7977	2578	.8097	3222	.8027
Band 5	6446	.7934	3222	.8049	3222	.8049
Simple probit						
Band 1	1289	.8893	645	.8693	3222	.7842
Band 2	2578	.8377	1289	.8252	3222	.7936
Band 3	3867	.8142	1934	.8176	3222	.8008
Band 4	5156	.8003	2578	.8107	3222	.8039
Band 5	6446	.7934	3222	.8048	3222	.8048
Bivariate probit with selection (BVP)						
Band 1	1289	.8892	645	.8674	3222	.7844
Band 2	2578	.8375	1289	.8256	3222	.7935
Band 3	3867	.8141	1934	.8178	3222	.8010
Band 4	5156	.8003	2578	.8108	3222	.8039
Band 5	6446	.7934	3222	.8048	3222	.8048
Weighted bivariate probit with selection (weighted BVP)						
Band 1	1289	.7695	645	.7324	3222	.7502
Band 2	2578	.7706	1289	.7599	3222	.7001
Band 3	3867	.7831	1934	.7936	3222	.7830
Band 4	5156	.7978	2578	.8093	3222	.8025
Band 5	6446	.7934	3222	.8048	3222	.8048

$$\begin{aligned} w_i &= p_i^{-1} \big/ n^{-1} \sum_{i=0}^{n} p_i^{-1} \quad \text{if } i \in \text{accepts}, \\ w_i &= 1 \quad \text{if } i \in \text{rejects}, \end{aligned} \tag{11}$$

where p_i = the predicted probability of acceptance, case i.

In this way, the relative weighting among accepted cases was maintained without affecting the relative weighting between accepted and rejected cases. Permitting the inverse of the probability of acceptance to be the weighting applied to rejected cases would have implied monumentally disproportionate attention to be given to the least acceptable cases among the rejects. Since use of the weighted BVP implies estimation of both an AR and a GB model, in principle the new AR model should be used to revise the weightings in a process that could iterate toward convergence. Had there been more classification success at the end of the initial iteration, this might have been attempted. However, the process of reweighting is mainly to focus attention toward more risky accepted cases, and the approximate replication of the character of all applicants is only an incidental byproduct.

Table 4 records, for each modelling approach, the area under the ROC curve which indicates the overall ranking performance achieved without reference to any arbitrary cut-off point. Logistic regression is the benchmark against which augmentation may be assessed and the comparably performing simple probit model is the benchmark for simple BVP and for weighted BVP. All results considered here deal with estimation using weights of evidence calibrated to the particular training-sample band.[4]

[4] This may seem somewhat constraining. However, the aforementioned study also considered an alternative resort to binary variables and produced similar results.

For simple BVP resort to binary variables as an alternative to weights of evidence was impeded by collinearity problems.

Consistent with the results reported in Crook and Banasik (2004) augmentation by itself provides ROC curve results quite inferior to those achieved without it. The results for BVP roughly confirm those reported in Banasik et al. (2003) except that now the slight performance improvement is non-existent. Table 5 indicates that this reflects a virtually complete absence of correlation between the AR and GB model errors even more so than previously.

The weighted BVP results represent considerable deterioration compared to a situation of no reject inference at all. The most that can be said for them is that BVP seems to have redeemed to some small extent the overall ranking results that would have occurred under simple augmentation.

Table 5
Error correlation arising from bivariate probit with selection estimation

	Simple bivariate probit with selection		Weighted bivariate probit with selection	
	ρ	Significance	ρ	Significance
Band 1	−.0321	.840	−.9908	.014
Band 2	−.0636	.645	.0355	.449
Band 3	−.1000	.303	−.0888	.722
Band 4	−.0101	.918	.1916	.348

Table 6 also confirms earlier results. In terms of classification results augmentation produces generally inferior results and in particular tends to undermine, for the upper two Bands, an ability to make good use of the Band 5 cut-off. The exception to this pattern is Band 4 where the training sample cut-off produces slightly better results and the Band 5 cut-off produces slightly worse results. For the

Table 6
Performance by correct classification

	Own band hold-out prediction			All-applicant hold-out prediction		
	Number of cases	Own band training cut-off (%)	Own band hold-out cut-off (%)	Number of cases	Own band training cut-off (%)	All band hold-out cut-off (%)
Simple logistic regression						
Band 1	645	89.30	89.77	3222	70.20	72.56
Band 2	1289	83.40	83.86	3222	70.58	72.75
Band 3	1934	79.21	79.42	3222	71.97	73.49
Band 4	2578	75.37	75.56	3222	72.47	73.81
Band 5	2578	73.68	73.49	3222	73.68	73.49
Weighted logistic regression (Augmentation)						
Band 1	645	87.75	87.60	3222	69.24	68.84
Band 2	1289	81.54	81.23	3222	68.34	67.47
Band 3	1934	79.16	79.42	3222	71.94	72.44
Band 4	2578	75.64	75.72	3222	72.84	73.49
Band 5	2578	73.68	73.49	3222	73.68	73.49
Simple probit						
Band 1	645	89.30	89.77	3222	70.11	72.75
Band 2	1289	83.32	84.02	3222	70.79	72.69
Band 3	1934	79.16	79.63	3222	71.88	73.56
Band 4	2578	75.41	75.41	3222	72.50	73.74
Band 5	2578	73.77	73.81	3222	73.77	73.81
Bivariate probit with selection (BVP)						
Band 1	645	89.30	89.77	3222	69.77	72.69
Band 2	1289	83.32	84.02	3222	70.36	72.56
Band 3	1934	79.06	79.63	3222	71.88	73.56
Band 4	2578	75.45	75.41	3222	72.53	73.74
Band 5	2578	73.77	73.81	3222	73.77	73.81
Weighted bivariate probit with selection (weighted BVP)						
Band 1	645	84.50	84.50	3222	56.80	70.64
Band 2	1289	81.54	81.69	3222	68.03	66.91
Band 3	1934	79.21	79.32	3222	71.88	72.50
Band 4	2578	75.33	75.56	3222	72.66	73.43
Band 5	3222	73.77	73.81	3222	73.77	73.81

simple unweighted BVP the results are very slightly worse, reflecting apparently inefficient resort to AR errors. Again Band 4 is the exception and again only insofar as the Band's own cut-off is used (as it normally would be).

The classification performance for weighted BVP seems very poor for the top two bands and again Band 4 provides the only exception to a finding of generally inferior performance compared to no reject inference at all. Taking Tables 4 and 6 together makes apparent what explicit crosstabulation of actual and predicted performance would convey. Overall ranking is somewhat undermined, and the indicated cut-off point serves very badly for Bands 1 and 2. Moreover, ranking in the critical region where decisions are made is also undermined by resort to this technique as indicated by comparison between the results from the simple probit with own-band cut-offs with a weighted BVP with Band 5 cut-offs. Even with that advantage this reject inference technique performs only marginally better in Band 1 (i.e. 70.64 vs. 70.11) and rather worse in Band 2.

7. The trouble with augmentation

Table 7 illustrates application of the weighting principles. The training-sample cases are ordered by acceptance probability determined by the AR model in such a way that each interval has about 129 "equivalent" probabilities. The top 1289 training cases are distinguished because these are the ones that are predicted to be accepted. In this way, the top 10 intervals include 167 rejected cases predicted to be accepted and the intervals below this include 167 accepted cases predicted to be rejected. The acceptance proportions in each interval bear a good likeness to each interval's typical acceptance probabilities given the relatively small number of cases in each.

A couple of features are very evident from Table 7. First, while 1122 correctly classified accepted cases (intervals 1–10) have the responsibility of representing all 1289 accepted cases, a large burden is put upon the 167 accepted cases wrongly predicted as rejected cases (intervals 11–15). They must represent all 5157 rejected cases (the sum of column 7 intervals 11–15). Indeed it is conceivable in principle that an accepted applicant could have an extremely small estimated probability of acceptance and thereby grab enormous attention in a weighted logistic regression. Secondly, the repayment behaviour in all but the top 129 band does not diminish radically as the acceptance cut-off point is approached. Indeed even below this point the good/bad ratio does not appear remarkably different. Accordingly, increased focus on "unacceptable" accepted cases does not provide much enhanced insight into the character of applicants with very bad repayment propensities.

Table 7
Reweighting illustration using Band 1

Interval	*P*(Accept) range within interval	Total				Training proportion			Represented by accepts
		Good	Bad	Accepts	Rejects	Cases	Accepted	Weights	
1	.99997–1.0000	126	3	129	0	129	1.00000	1.00	129
2	.99587–.99997	109	20	129	0	129	1.00000	1.00	129
3	.98302–.99587	113	16	129	0	129	1.00000	1.00	129
4	.96095–.98302	113	13	126	3	129	.97674	1.02	129
5	.93144–.96095	116	10	126	3	129	.97674	1.02	129
6	.88551–.93144	101	19	120	8	128	.93750	1.07	128
7	.82116–.88551	100	15	115	14	129	.89147	1.12	129
8	.72150–.82116	83	10	93	36	129	.72093	1.39	129
9	.60282–.72150	73	11	84	45	129	.65116	1.54	129
10	.48605–.60282	66	5	71	58	129	.55039	1.82	129
Subtotal				1122	167	1289	.87044		1289
11	.35984–.48605	48	3	51	78	129	.39535	2.53	129
12	.24927–.35984	34	2	36	93	129	.27907	3.58	129
13	.16051–.24927	20	3	23	106	129	.17829	5.61	129
14	.10240–.16051	17	3	20	109	129	.15504	6.45	129
15	.00000–.10240	31	6	37	4604	4641	.00797	125.43	4641
Total				1289	5157	6446			6446

Augmentation will provide benefit particularly when there are a large number of accepted applicants judged by an AR model to be worthy of rejection and these cases have a distinctly poor repayment performance. That should tend not to happen when the rejection rate is large – which is when reject inference seems most needed. This feature perhaps explains why Band 4 had some instances of benefit, and only small benefit at that, from reject inference.[5]

8. Conclusion

The two forms of reject inference considered here appear to provide negligible benefit whether applied in isolation or together. The nature of such negative findings is that they cannot be presented as significantly insignificant, but they arise from carefully designed experiments devised with rare data particularly suited for them. Apparent scope for reject inference in terms of the loss of accuracy that arises from modelling with a data set comprising only the more creditworthy applicants is clearly evident. In a population in which 66.6% of applicants (see Table 2) are likely to repay, a model that correctly classifies 70.2% represents a small improvement over simply accepting everyone, and the 3.48% scope for improvement possible in Band 1 represents a substantial improvement over that. The challenge is to achieve a substantial part of that scope.

An important feature of the two reject inference techniques considered here is that they are both mechanical and do not depend at all on modellers' judgement about suitable parameters. While there is nothing wrong with techniques that do depend on such judgement, appraisal of their accuracy may not easily be able to distinguish between the improvement latent in the technique as opposed to that contingent on good judgement. Even in the experiments reported in this paper it might be possible to manipulate the experiments to affect the results, for example by altering the variable selection for GB and AR models, but such arbitrary judgements have been devised with a view to the reliability of the experiment not the success of the model. The two types of judgement are distinct. Accordingly, the findings pertaining to the techniques considered here are more definitive than might be the case for others.

The findings reported above reflect the features of one data set corresponding to one context. Reject inference may very well be applied with good effect to various other contexts. Unfortunately, an ability to assess the benefit will usually be absent, since the opportunity cost of rejecting applicants can rarely be known. The data set employed here has effectively provided data on the repayment behaviour latent in all rejected applicants.

In principle it seems that the feature required of success for the two types of reject inference considered here, both separately and together, is a lot of information in the acceptance decision that pertains to the "goodness" of applicants yet is denied to the variable set of the GB model. That should tend to make focus at the lower range of acceptable applicants worthwhile and should foster correlation between the errors of the GB and AR models. These are both observable features without knowledge of the latent repayment behaviour of rejected applicants, and so should be a good indication of the prospects of benefit from applying reject inference. Unfortunately, without the knowledge of this latent behaviour, the extent of benefit will be difficult to assess.

Considerable further research is suggested by this paper. For example, it would be beneficial to experiment to see how sensitive the results are to different weighting formulae. Second one could try to relate the approaches to reject inference to the actual Bayesian network structure of the data. Third one might try Bayesian inference (see for example Smith and Elkan, 2003). Fourth one might try to identify specific conditions under which the omission of performance data for certain cases results in biased estimates of parameters and use a particular reject inference technique that is most beneficial under each set of specific conditions.

Acknowledgements

We would like to thanks two anonymous referees for very helpful suggestions. All errors are our own.

References

Banasik, J.L., Crook, J.N., 2005. Credit scoring, augmentation and lean models. Journal of the Operational Research Society 56, 1072–1091.

Banasik, J.L., Crook, J.N., Thomas, L.C., 2003. Sample selection bias in credit scoring models. Journal of the Operational Research Society 54, 822–832.

[5] Another possible explanation could be sampling error.

Crook, J.N., Banasik, J.L., 2004. Does reject inference really improve the performance of application scoring models? Journal of Banking and Finance 28, 857–874.

Feelders, A.J., 2000. Credit scoring and reject inference with mixture models. International Journal of Intelligent Systems in Accounting, Finance and Management 9, 1–8.

Hand, D.J., Henley, W.E., 1994. Inference about rejected cases in discriminant analysis. In: Diday, E., Lechvallier, Y., Schader, M., Bertrand, M., Buntschy, B. (Eds.), New Approaches in Classification and Data Analysis. Springer-Verlag, Berlin, pp. 292–299.

Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5 (4).

Little, R.J., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.

Meng, C.L., Schmidt, P., 1985. On the cost of partial observability in the bivariate probit model. International Economic Review 26, 71–85.

Schafer, J.L., 1997. Analysis of Incomplete Data. Chapman and Hall, Bury St. Edmunds.

Smith, A., Elkan, C., 2003. A Bayesian Network Framework for Reject inference. Mimeo, Department of Computer Science, University of California, San Diego.