IMA Journal of Mathematics Applied in Business & Industry (1993/4) **5**, 45–55

Can reject inference ever work?

D. J. HAND AND W. E. HENLEY*

Department of Statistics, The Open University, Milton Keynes MK7 6AA, U.K.

The true good/bad status of applicants accepted for credit is ultimately known. However, the status of rejected applicants will never be known. 'Reject inference' is the process of inferring the status of applicants who have been rejected. This paper reviews methods of reject inference, and describes some new approaches. Three classes of method are described: (i) methods based on extrapolating a model built on the accepted applicants into the reject region; (ii) methods based on the distribution of the rejected applicants; (iii) methods using supplementary information. In particular, we conclude that the distribution of the rejected applicants cannot assist reject inference unless additional assumptions are made.

1. Introduction

When applicants apply for credit, they are assessed using a scoring instrument and either accepted or rejected. The accepted ones are followed up, and their true good or bad credit-worthiness can be determined. However, the rejected ones are not followed up, so that their true status is unknown. 'Reject inference' is the process of attempting to infer the true creditworthiness status of the rejected applicants.

There are several reasons for being interested in this. A straightforward one is a wish to determine the number of good credit risks rejected by the scoring instrument. Another, more subtle, involves the effort to develop improved scoring instruments. In general, the only data available on which to develop new instruments will be: (a) the applicants' responses to the individual items comprising the scoring instrument (these will be available for both accepted and rejected applicants), and (b) the true good or bad status of those accepted (but not of those rejected). Unfortunately, as outlined below, instruments based solely on the accepted applicants could be biased. This observation has led to the idea that, if one can use reject inference to infer the status of those rejected, then perhaps this information could be combined with the known status of those accepted to reduce the bias in the new instrument.

Reject inference has attracted a great deal of interest, and many methods have been proposed. Much of the work seems to have been based on a poor understanding of what can be achieved using the rejects. The aim of this paper is to clarify this, and to describe some proposed methods of reject inference.

For simplicity, we shall discuss only the case of two classes: those with high probability of defaulting on the credit repayments (the 'bads') and those with a low probability (the 'goods'). In particular, we shall not consider a continuum of such probability classification, and we shall not consider different kinds of defaulter. Nor, in this paper, shall we discuss models based on changes of state between good

* Correspondence regarding this paper should be addressed to the first author.

and bad—as far as we are concerned, each credit applicant belongs to just one of the two classes and is fixed in that class. Thus, for the purposes of this paper, each applicant may be regarded as having a true class.

2. Extrapolation from the accepts

Let $\mathcal{X}$ be the set of characteristics measured for each applicant. This set could include such categories as age (perhaps grouped as <25, 25–29, 30–39, 40–49, 50–59, 60 or above), time at present address (<1 year, 1–5 years, >5 years), and type of accommodation (rented flat, owned flat, owned house). The possible responses to each such characteristic are termed *levels*—so that, for example, the characteristic 'accommodation' has three levels of response. The initial accept/reject decision is based on the pattern of these levels of response.

The probability that an applicant with a particular vector $\boldsymbol{x}$ of characteristic levels is good will be denoted by $P(\mathrm{g} \mid \boldsymbol{x})$ and the probability density function of the good-risk applicants will be denoted by $p(\boldsymbol{x} \mid \mathrm{g})$ (with corresponding notation for bad).

A fundamental approach to reject inference is straightforward extrapolation from the accept region of the $\mathcal{X}$ space over the reject region. That is, using just the sample of 'accepts' (their known $\boldsymbol{x}$ vectors and their known true good/bad status), one constructs a model to predict the probability of being good in the accept region, and then extrapolates this over the reject region.

Of course, the extent of truncation (the size of the reject region) will determine how good a model can be fitted. A large reject region means there is little information on which to base the model. Moreover, problems will arise with any extrapolation method if the $P(\mathrm{g} \mid \bullet)$ function differs in form between the accept and reject regions, or if the error variation is such that a highly accurate model cannot be constructed in the accept region. The latter would mean that the form of the model could not be accurately specified. For example, the accept data may not permit one to reject the hypothesis of a straight-line predictor in favour of a quadratic predictor. This being the case, extrapolating beyond the accept data into the reject region could well produce inaccurate results as the departure from linearity becomes more extreme.

As far as the extrapolation goes, one might base it on either of the two fundamental approaches to classifier design. In the first approach, of which logistic regression is an example, one estimates $P(\mathrm{g} \mid \bullet)$ directly. In the second, illustrated by classical linear discriminant analysis, one estimates $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$ and then derives $P(\mathrm{g} \mid \bullet)$ using Bayes' theorem:

$$P(\mathrm{g} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \mathrm{g})P(\mathrm{g})}{p(\boldsymbol{x} \mid \mathrm{g})P(\mathrm{g}) + p(\boldsymbol{x} \mid \mathrm{b})P(\mathrm{b})},$$

where $P(\mathrm{g})$ and $P(\mathrm{b})$ are the overall proportions of goods and bads in the population. The theory is standard and is explained in many places, for example Hand (1981).

Now, for our purposes, there is a vital distinction between the two approaches (Dawid 1976). This is that a sampling fraction which varies arbitrarily across the $\mathcal{X}$ space will lead to distortion of methods that estimate the probabilities via $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$ but not of methods which directly estimate $P(\mathrm{g} \mid \bullet)$.

For example, sparse sampling (based on x alone) from one side of a normal distribution will lead to an asymmetric distribution, and so methods that assume normality of the resulting $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$ distributions will be biased. However, such sparse sampling will not influence $P(\mathrm{g} \mid \bullet)$: the proportion of goods at any particular x will remain the same at that x whatever the sampling fraction, so that $P(\mathrm{g} \mid x)$ will not be distorted.

In our situation, the distribution of known goods is truncated (containing only applicants from the accept region). The same applies to the distribution of bads. This truncation will bias any model-based estimate of $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$ and, hence, of estimates of $P(\mathrm{g} \mid \bullet)$ based on them. However, truncation will not bias methods based on estimating $P(\mathrm{g} \mid \bullet)$ directly.

It would be possible to derive approaches based on $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$ but modified to make allowance for the truncation; however, since these would depend on the validity of assuming (for example) multivariate normality under no truncation, it is unlikely that they would be very robust in practice.

The consequence of all this is that, if the accept/reject decision is based just on the scores in $\mathcal{X}$, then one might expect methods such as logistic regression to extrapolate reasonably well beyond the accepts, but methods such as classical discriminant analysis to perform badly. The latter would require extensive adjustment to account for the truncation resulting from the rejection. A number of empirical studies have demonstrated the truth of these observations: for example, Eisenbeis (1978) cites work by Avery showing how truncation of normal populations can cause bias in methods based on discriminant analysis.

Of course, both approaches will be adversely influenced if an incorrect model form is adopted.

Extrapolation from a model built on the accept region and based on the $P(\mathrm{g} \mid \bullet)$ is thus an appealing approach, and one which forms a natural baseline against which to assess the other methods outlined below.

3. Using the rejects

If one does not assume known distributional forms for the $\mathcal{X}$ values of the goods and bads, $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$, then it is simple enough to show that the $\mathcal{X}$ values of the rejects (for whom the true good/bad status is unknown) do not contain information about the parameters of the $P(\mathrm{g} \mid \bullet)$ function.

To see this, let S be a random variable taking values g and b and let X be the random vector of characteristics. Then the joint probability function of S and X is

$$\mathrm{P}[(S, \mathrm{X}) = (s, x)] = P(s, x) = P(s \mid x; c)p(x; d),$$

where c and d are parameters for the indicated distributions. We wish to estimate $P(\mathrm{g} \mid \bullet; c)$, (or equivalently $P(\mathrm{b} \mid \bullet; c) = 1 - P(\mathrm{g} \mid \bullet; c)$). If we do not assume a known form for $p(\bullet; d)$, then this does not contribute to the likelihood. That is, the $\mathcal{X}$ values do not contain information about the parameters of $P(\mathrm{g} \mid \bullet; c)$. This is true for both accepts and rejects; but in particular it means that using the rejects cannot lead to an improved scoring rule.

If, on the other hand, we do assume particular forms for the separate distributions

of the $\mathcal{X}$ values of the goods and bads, $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$, over the entire space, then the parameters in $P(\mathrm{g} \mid \bullet; c)$ intersect those in $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$ so that the $\mathcal{X}$ values, including those of the rejects, can be helpful. The parameters can then be calculated using mixture-decomposition methods (e.g. Everitt & Hand 1981). We shall discuss this possibility below.

Now suppose that some proportion of those applicants with response pattern $\boldsymbol{x}$ is accepted, where the accept/reject decision is based solely on $\boldsymbol{x}$. Then the proportion of goods among the accepts at $\boldsymbol{x}$ must equal the proportion of goods among the rejects at $\boldsymbol{x}$: no extra information on which the good/bad can be discriminated is being used.

However, if the accept/reject decision also uses extra information (such as, for example, extra characteristics not included in $\boldsymbol{x}$), then the proportion of goods among the rejects will typically not equal that among the accepts at any particular point of $\mathcal{X}$. This means that a model based solely on the accepts will normally be biased.

From another perspective, if the accept/reject decision is made in the $\mathcal{X}$ space and then a new scorecard is constructed using information in a set $\mathcal{Y}$ of characteristics, with $\mathcal{Y}$ a subset of $\mathcal{X}$, then a biased instrument will result. (Exactly the same applies if the new scorecard uses a set $\mathcal{Z}$ of characteristics where $\mathcal{Z}$ does not include all of $\mathcal{X}$.)

Moreover, the difference between the probability of being good among the accepts and the probability of being good among the rejects will normally vary between different points of $\mathcal{Y}$. Thus, to avoid bias when constructing a new instrument using an accept sample obtained from an earlier instrument, it is necessary to include all characteristics used by the earlier instrument. The fact that this is not always done may go some way towards explaining claims to have improved on the straightforward extrapolation method by using the rejects.

A simple example, in which the subset $\mathcal{Y}$ is a single characteristic, is as follows. Consider a characteristic such that low scores are associated with poor risks and high scores with good risks. Then it is likely that the scorecard will have rejected a high proportion of those with low scores. These rejections will not be based solely on the low scores on this characteristic alone, however, but on the overall score of the scorecard. Following the argument above, this implies that (conditional upon a particular level of this characteristic) the proportion of goods among the accepts will generally not equal the proportion of goods amongst the rejects. Moreover, if the instrument has any validity, one might reasonably expect that there will be a lower proportion of goods among the rejects than amongst the accepts. The consequence is that, looking at this characteristic alone, it would show apparent bias in its lower scores, in the sense that the accepts at these scores would underestimate the proportion of bads and overestimate the proportions of goods.

We explored this empirically as follows. A characteristic fulfilling the above criterion of low scores being associated with a risk of being bad is 'number of weeks since the last County Court judgement'. This had levels graded 1 (1–26 weeks), 2 (27–52 weeks), 3 (53–104 weeks), 4 (105–208 weeks), 5 (209–312 weeks), and 6 ($\geqslant$313 weeks, or no County Court judgement). As is outlined in the next section, we did have available a sample which covered the entire $\mathcal{X}$ space. We built a scorecard and used this to divide the sample into accepts and rejects. This permitted us to compare

TABLE 1

Ratio of the estimated probability of being good in the accept sample to the estimated probability of being good in the full sample, classified according to the number of weeks since the last County Court judgement

Level (number of weeks)	Full sample: proportion good	Accept sample: proportion good	Accept/Full ratio
1	0.227	0.446	1.964
2	0.304	0.469	1.542
3	0.314	0.492	1.567
4	0.378	0.552	1.460
5	0.426	0.631	1.481
6	0.556	0.692	1.245

the probabilities of being good at each level of this characteristic as calculated from the accept sample alone and as calculated from the entire sample.

The results are shown in Table 1. The first column gives the level. The next two give the proportions of goods in the full sample and accept sample respectively. And the final column gives the ratio of these two 'proportions good'. It is quite striking that the ratio of the two proportions decreases as the level increases, i.e. as the risk of being bad is expected to decrease (and indeed does, as is indicated by the trends in the middle two columns). This is precisely what our model predicted. (And, since all the proportions are well away from unity, this is presumably not just a ceiling effect.) It follows from this that one might attain improved performance if one biased the scorecard based on the accepts to fit a lower proportion of goods than was available in the accept sample, for low levels of the characteristic.

This example has used a single characteristic as the $\mathcal{Y}$ subset, but the reasoning applies more generally. And, as noted above, it could explain claims of using the rejects to improve a scorecard. The mechanism is one of adjusting the estimated $P(g \mid \bullet)$ model towards one's beliefs about the proportions—basically, a Bayesian approach.

Of course, all of the above assumes that the full sample and accept samples are taken from the same populations. If this is not the case, then things become more complicated—and there are more possibilities for improving the scorecard. These are discussed below.

We shall finish this section with a brief discussion of three methods which have been proposed and which attempt to make use of the rejects. The first is the 'augmentation' method, which nicely illustrates some of the points made above. The second method is much more modest in the use it makes of the rejects, using them solely to obtain improved estimates of covariance matrices. And the third method is the mixture-decomposition approach mentioned above.

Method 1

Hsia (1978) describes the technique, which is widely used by developers of scoring systems in attempting to take advantage of the rejects' characteristic vectors. Implicit

in it are assumptions that the accept/reject decision was made using a set $\mathcal{X}$ of characteristics, and that the new scorecard is to be built using a set $\mathcal{Y}$ such that $\mathcal{X}$ is not a subset of $\mathcal{Y}$. As we have seen above, this is likely to imply that the probability of being good among the accepts with any particular score on the new scorecard is not the same as the probability of being good among the rejects with that score.

The method begins by developing a new scorecard, using $\mathcal{Y}$, to discriminate between the accepts and rejects. Now, since $\mathcal{Y}$ does not include $\mathcal{X}$, there will be some elements of $\mathcal{X}$ not in $\mathcal{Y}$. In general this will mean that each point in $\mathcal{Y}$ will be associated with some accepts and some rejects. The augmentation method then uses the accepts among these to estimate the conditional probabilities $P(\mathrm{g} \mid y)$. The overall probabilities (of both accepts and rejects) $p(y)$ $(y \in \mathcal{Y})$ then serve as a reweighting.

The reweighted data now provide nominal distributions of goods and bads throughout the $\mathcal{Y}$ space, so that standard methods of building classifiers can be applied. Of course, if an accurately specified model was being used, then the reweighting would have no effect on the estimated parameters.

The key assumption latent in this method is that the probabilities $P(\mathrm{g} \mid y)$ for the accepts are the same as those for the rejects. This is unlikely to be the case in practice since the accept/reject decision, based on $\mathcal{X}$, which has information not included in $\mathcal{Y}$, is likely to serve to separate these two classes on the basis of the proportions of goods in them. In general, then, the augmentation method should not be expected to lead to valid estimates of good/bad probabilities.

Method 2

Reichert *et al.* (1983) compared a two-group (good accepts versus bad accepts) classical linear discriminant analysis with a three-group (good accepts, bad accepts, and rejects) classical linear discriminant analysis. They note that the inclusion of the rejects did not improve the discrimination between good and bad accepts. However, they observe that (p. 105): 'the key decision is whether to grant credit in the first place and not to identify good or bad borrowers once a loan has been granted. Thus a three-group model based on the entire population of applicants is conceptually more appropriate than a truncated two-group model.'

The first sentence here is obviously true. However, since the aim is to split the entire population into two groups, it is not clear that a three-group discriminant analysis is at all the appropriate thing to do. Constructing a model that specifically identifies characteristics of a third group which one can identify perfectly anyway, and which one would like to partition into good and bad, seems at best perverse. An extrapolation approach seems much more in tune with the objectives.

Be that as it may, classical linear discriminant analysis assumes a common covariance matrix for the three (or however many) groups in question. Thus this approach does try to make use of information in the rejects in devising the classification rules, in that the rejects are used to lead to more accurate estimates of this common covariance matrix. Having said that, it should be added that we do not believe that this will compensate for the other problems of this method outlined above.

Method 3

Classical linear discriminant analysis estimates $p(\cdot \mid \mathrm{g})$ and $p(\cdot \mid \mathrm{b})$ and then calculates

$$\frac{P(\mathrm{g} \mid x)}{P(\mathrm{b} \mid x)} = \frac{p(x \mid \mathrm{g})}{p(x \mid \mathrm{b})} \frac{P(\mathrm{g})}{P(\mathrm{b})}. \tag{3.1}$$

If one assumes, as classical linear discriminant analysis does, that $p(\cdot \mid \mathrm{g})$ and $p(\cdot \mid \mathrm{b})$ belong to a particular family of distributions, then one can estimate the parameters using both the classified cases (the accepts) and the unclassified cases (the rejects) using the EM algorithm. (The EM algorithm is an algorithm for finding maximum-likelihood solutions when there are missing data. The true good/bad status of the rejects may be regarded as missing data (Dempster *et al.* 1977)). This is an example of making the truncation explicit, noted above. Indeed, one can go even further, and use the EM approach in a mixture decomposition on entirely unclassified cases. That is, one can estimate $p(\cdot \mid \mathrm{g})$ and $p(\cdot \mid \mathrm{b})$ from rejects alone, requiring no accepts whatsoever, and then use (3.1) to estimate $P(\mathrm{g} \mid \cdot)$ throughout the whole space.

To see how this is done, define $p(\cdot)$ as the overall distribution of applicants (regardless of whether their good/bad classification is known). Now

$$p(x) = p(x \mid \mathrm{g})P(\mathrm{g}) + p(x \mid \mathrm{b})P(\mathrm{b}). \tag{3.2}$$

The left-hand side can be estimated from the overall distribution of the sample. For particular assumed values of $P(\mathrm{g})$, $P(\mathrm{b})$, and the parameters of $p(\cdot \mid \mathrm{g})$ and $p(\cdot \mid \mathrm{b})$, the right-hand side gives a completely specified distribution. Then we choose the set of parameters that leads to the smallest difference between the two sides.

This is straightforward enough, and is illustrated in a slightly different context, in Hand & Fitzmaurice (1987). And it seems like magic: even without knowledge of the good/bad classification of any accepts, we seem to have produced a reasonable estimate of the good/bad classification throughout the entire space. Predictably enough, this power has been gained at the cost of an assumption. We assumed that the distributions belonged to some parameterized family. A common assumption is that of classical linear discriminant analysis, based on multivariate normality. Unfortunately, in the credit-scoring context, with many categorical variables and markedly non-normal marginal distributions, assuming normality is unrealistic, and there does not seem to be any obvious parametric alternative. Nevertheless, this approach does seem to offer a genuine way in which advantage might be taken of the rejects, and might be worth pursuing.

4. Methods with supplementary information

As noted above, the difficulties arise because there are regions of X on which no good/bad information at all is available (the reject region). It is clear that the most straightforward way of obtaining valid reject inferences would be to obtain some information about these regions. The easiest way to do this would be to take a subsample of the cases which lie in these regions.

Naturally cost considerations enter here. Every accepted bad applicant represents a financial loss, so that somehow the information gained in terms of increased accuracy of the subsequent scorecard has to be balanced against this loss.

In this section, we assume that some such supplementary information is available. In what follows, we shall call it the 'calibration sample'. Given that such information is available, the next question to be addressed is how best to make use of it.

If we knew that the calibration sample had been randomly chosen from the population of interest (or even by some arbitrary sampling mechanism dependent solely on the characteristics spanning $\mathcal{X}$) then we could simply combine the accepts and the calibration samples and use a statistical technique based on the posterior probabilities $P(\mathrm{g} \mid x)$ of being a good risk, such as logistic regression. (But not, note, a technique based on the class-conditional distributions $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$, such as classical linear discriminant analysis, for the reasons described in detail above.) In such a case, the accept sample may add little information to that already contained in the calibration sample. On the other hand, if we are unsure about this, then either we have to compare the two samples (the accepts and the component of the calibration sample lying in the accept region) or we have to consider what aspects of the calibration sample we can reasonably expect to hold also for the new population, and see if use can be made of these. Three methods aimed at doing that are now presented.

Method 4

We first outline the idea behind this method, which we believe to be original, and then present a simple (and simplistic) example. The method assumes that several calibration samples are available. These samples are partitioned, using whatever method was used to partition the original sample, into accepts and rejects. For each calibration sample, the distribution of the probability of being good over the reject region is known (this is the definition of a calibration sample). The same is true, of course, over the accept region. Thus one can identify characterizing features of the distributions in each of the two regions for each calibration sample. Moreover, by studying the whole set of calibration samples, and the values that the characterizing features take over this set, one can estimate the relationship between the values of the features in the two regions. Finally, these relationships allow one to map from the values of the characterizing features taken on the accepts of the new sample to the values of the characterizing features of the rejects of the new sample—and hence to the form of $P(\mathrm{g} \mid \bullet)$ over the reject region.

The following example is absurdly oversimplified for reasons of exposition. Suppose that the scorecard consists of just one characteristic with three levels. Suppose also that one of these levels represents the reject region—all applicants who respond at this level are rejected—and that the other two represent the accept region. Now suppose that we have ten calibration samples. These may be obtained from different time periods, from different geographical locations, or from loans on different products, for example.

One feature is sufficient to characterize the population of rejects: this is the proportion of goods in the single level that constitutes the reject region. Similarly,

two are sufficient to characterize the accept region: the proportions of goods in each level comprising the accept region. This means that each of the ten calibration samples is described by just three numbers.

We can now use these ten samples to build a simple regression model showing the relationship between the two accept proportions and the single reject proportion. Then, for any pair of proportions, representing a population's scores in the accept region, we can predict, from the regression model, that population's score in the reject region. That is, we can predict that population's proportion good in the reject level of the characteristic.

Of course, to make this realistic, it must be extended substantially in a number of directions. First, the cross-classification of a number of characteristics must be considered. Typically this will also mean that there will be multiple cells in the reject region. Moreover it will normally mean that there will be far too many cells, in both reject and accept regions, for the characterizing features to be merely the proportions in the cells. Some more subtle summarizing statistics will be needed. Obvious suitable examples are low-order log-linear models or perhaps regression models with cells replaced by scores derived using correspondence analysis. Note also that a multivariate multiple regression or canonical correlation analysis will be needed to relate the two sets of characterizing variables together. This, in turn, implies that the number of calibration samples must be large enough to span the space of the combined set of variables (those used to describe the accept proportions and those used to describe the reject proportions) effectively, and permit sufficiently accurate estimation of the model coefficients.

This means that the calibration samples should come from different sources, such as times, locations, and products, as noted above. To some extent one can achieve this by partitioning any available calibration sample; but, of course, a trade-off is necessary between the number of such subsamples produced and the size of each one: each one must yield sufficiently accurate estimates of its characterizing feature values.

Method 5

This method, which we also believe to be original, requires a single calibration sample.

In the above, we have described extrapolation methods. These are based on building a model on the accept region and extrapolating it over the reject region. The assumption is that the model will extrapolate validly. This assumption, of course, is risky since it is a classic example of extrapolating beyond the data. Indeed, the reject region has been so designated precisely because it differs in an important way from the accept region. Thus the assumption that the model may legitimately be extrapolated over this region might be regarded as shaky. This is the sort of argument which motivated the idea that things could be improved by utilizing subjective information about the shape of the function in the reject region.

In method 5, we try to use objective information provided by the calibration sample. We begin by constructing a scorecard using the accepts who comprise the new sample. Then using the calibration sample we construct two scorecards. The first uses the entire calibration sample, including applicants in both accept and reject

regions. The second scorecard constructed from the calibration sample uses only those applicants lying in the accept region. To the extent that the extrapolation method will perform poorly, this will differ from the complete calibration sample scorecard. This difference is then used as an adjustment to the scorecard built using the accepts comprising our new sample.

As an example, suppose that linear regression is used to build the scorecards. Suppose that the vector of regression coefficients for the new sample (accepts) is $\boldsymbol{a} = (a_1, ..., a_n)$, and likewise $\boldsymbol{b} = (b_1, ..., b_n)$ for the validation-sample accepts and $\boldsymbol{c} = (c_1, ..., c_n)$ for the entire validation sample. Then a simple way of describing the relationship between $\boldsymbol{b}$ and $\boldsymbol{c}$ would be via a diagonal matrix $\boldsymbol{M}$, the elements of which are simply the ratios of the components of $\boldsymbol{c}$ to those of $\boldsymbol{b}$. That is, the ith diagonal element $\boldsymbol{M}$ is c_i/b_i. Using this matrix, an estimate of the regression coefficients for the entire new population would be $\boldsymbol{Ma}$.

Many other adjustment schemes are possible, and the identification of those which are most effective is a topic needing further work.

Method 6

In method 3 above, we described how a mixture decomposition approach could be used to estimate the parameters of $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$ by comparing the mixture of these distributions with the empirical distribution of $p(\bullet)$. If information about the true good/bad classes of some of the applicants is known, then this can be used to improve the estimate, but this was not necessary. What was necessary was that some parameterized form for the component distributions could be assumed.

Now we are assuming that there is some information about the true good/bad classes of some of the applicants—available from the calibration sample and hence available through $\mathcal{X}$ (and not restricted just to the accept region). Such information could be used to choose families for the distributions of the $p(\bullet \mid \mathrm{g})$ and $p(\bullet \mid \mathrm{b})$ distributions. As before, once the families have been chosen, estimates of the parameters can be obtained using both classified applicants and unclassified applicants via (3.2). Again, as far as we know, this approach has not been tried in practice.

5. Conclusion

Reject inference is the process of inferring the good/bad credit risk status of applicants who have been rejected. Approaches to such inference have been reviewed, dividing them into three classes: straightforward extrapolation methods, methods that attempt to utilize information in the distribution of the rejects, and methods based on supplementary information.

We have shown that it is important to distinguish between two broad classes of classification methods for use in reject inference: methods based on direct estimation of the probabilities of being good and methods based on indirect estimation via the class-conditional distributions. Of these, the latter is particularly susceptible to bias induced by the truncation implicit in the accept/reject decision.

We also pointed out the necessity of using all the characteristics used in the original

accept/reject decision when attempting to build an improved classifier. Again, without this, a biased result is likely.

Much effort has been expended on trying to utilize information in the distribution of the rejects. We pointed out that such information can, in general, only do this if additional assumptions are made.

Sometimes a calibration sample is available, giving some information on the true good/bad class of applicants throughout the characteristic space, though often these samples are not from precisely the same sample as that to which one wishes to apply reject inference. Three original methods for using such samples have been described.

Our overall conclusion is that reliable reject inference is impossible. In particular, claims that such improvements have been achieved by reject inference are based on the following circumstances.

- *Chance.* The new rule is better than the old one by luck.
- *The use of additional information*, such as approximately correct assumed distributional forms in a mixture decomposition method, or an extra 'calibration' sample, or expert skill and knowledge of the area. The latter may, of course, be subconscious.
- *Ad hoc adjustment of the rules* in a direction likely to lead to reduced bias. For example, in the augmentation method, one might reasonably assume that the rejects with a particular x vector had a lower probability of being good than the accepts with that vector.

Acknowledgement

W. E. Henley was supported in this work by a Research Studentship awarded by Littlewoods plc.

References

DAWID, A. P. 1976. Properties of diagnostic data distributions. *Biometrics* **32**, 647–58.

DEMPSTER, A. P., LAIRD, N. M., & RUBIN, D. B. 1977. Maximum likelihood from incomplete data. *Journal of the Royal Statistical Society* B **39**, 1–38.

EISENBEIS, R. A. 1978. Problems in applying discriminant analysis in credit scoring models. *Journal of Banking and Finance* **2**, 205–19.

EVERITT, B. S., & HAND, D. J. 1981. *Finite mixture distributions.* London: Chapman & Hall.

HAND, D. J. 1981. *Discrimination and classification.* Chichester: Wiley.

HAND, D. J., & FITZMAURICE, G. M. 1987. Error rate estimation by mixture decomposition. *Computers and Mathematics with Applications* **14**, 573–8.

HSIA, D. C. 1978. Credit scoring and the equal credit opportunity act. *The Hastings Law Journal* **30**, 371–405.

REICHERT, A. K., CHO, C.-C., & WAGNER, G. M. 1983. An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business and Economic Statistics* **1**, 101–14.