Modified Logistic Regression using the EM Algorithm for Reject Inference

Billie S. Anderson, Ph.D.
Research Statistician
SAS Institute
100 SAS Campus Dr.
Cary, NC 27513
(919)531-3687 (work)
Billie.Anderson@sas.com

J. Michael Hardin, Ph.D.
Senior Associate Dean
The University of Alabama
Box 870221
Tuscaloosa, AL 35487
(205)348-8901
mhardin@cba.ua.edu

Abstract.
Credit Scorecards are commonly built using data available within an organizations transactional database. Such data, however, will only contain information for those applicants who were 'accepted' or previously awarded credit by the organization; data will not be available for those applicants who were 'rejected'. The use of *reject inference* to adjust credit scorecard models for the missing data represented by rejected loan applications is common practice and several approaches are used in today's financial industry.

Logistic regression is used to linearly estimate the probability of an applicant being a bad credit risk. In this paper we propose an alternative method for estimating the parameters in a logistic regression model for use as a reject inference technique. This paper presents a method for estimating the regression coefficients in a binomial regression when the response variable is missing for the rejected applicants, and the missing data mechanism is considered to be missing at random (MAR) (Little and Rubin 1987). The authors assume that the characteristic variables are fully observed and use the EM (Expectation-Maximization) algorithm to estimate the regression coefficients. The modified logistic regression (EM logistic model) in which the parameters are estimated using accepted and rejected applicants' data will be compared to a standard augmentation procedure that uses only accepted applicant's data to estimate the

parameters of a logistic regression model. The performance of the EM logistic model and augmentation will be evaluated through a simulation study.

Keywords: reject inference EM algorithm

1. Introduction

Logistic regression is one of the most common credit scoring models used in the credit industry (Hand and Henley 1997; Thomas 2000). Logistic regression is used to linearly estimate the probability of an applicant being a bad credit risk. In this paper we propose an alternative method for estimating the parameters in a logistic regression model for use as a reject inference technique. This article presents a method for estimating the regression coefficients in a binomial regression when the response variable is missing for the rejected applicants, and the missing data mechanism is considered to be missing at random (MAR) (Little and Rubin 1987). The authors assume that the characteristic variables are fully observed and use the EM algorithm to estimate the regression coefficients. The modified logistic regression (EM logistic model) in which the parameters are estimated using accepted and rejected applicants' data will be compared to a standard augmentation procedure that uses only accepted applicant's data to estimate the parameters of a logistic regression model. The performance of the EM logistic model and augmentation will be evaluated through a simulation study.

Section 2 develops notation and the complete-data model for the EM logistic model. In Section 3 we derive the E- and M- steps of the EM algorithm for the EM logistic model. In Section 4 the augmentation procedure used to compare to the EM logistic model is described and details of the simulation study are discussed. Section 5 presents the details of how the simulation study was evaluated. Section 6 gives the results of the study and section 7 concludes with final remarks.

2. Model and Notation

Suppose $\mathbf{X}$ denotes a $((n \times m) \times p)$ a data set matrix of characteristic information, with *i*th row $x_i = (x_{i1}, ..., x_{ip})$ where x_{ij} is the value of the variable x_j for applicant i in which there are n accepted applicants and m rejected applicants for a total of $n+m$ applicants. The 1 is included in $\mathbf{x}$ to indicate the intercept of the model will be used. The outcome vector $\mathbf{Y}$, will be a ((n x m) x 1) vector denoting whether the loan for applicant j was a good or bad loan and y will denote the observed value of $\mathbf{Y}$. We will assume the outcome variable has been coded as 0 (bad loan) or 1 (good loan). For the n accepted applicants, y will be observed and for the m rejected applicants y will be missing. Suppose $y_1, ..., y_n, y_{n+1}, ...y_m$ are independent applicant observations, where each y_j has a

Bernoulli distribution with probability of default p_j, j=$1, ..., n, n+1, ..., m$ where p_j takes the form:

$$p_j = P(Y_j = 1 \mid \mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1+\exp(\mathbf{x}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{x}\boldsymbol{\beta})} \qquad (1)$$

and where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ denotes the p-vector of regression coefficients that are to be estimated. The equation in (1) is known as the logistic model (Neter, Kutner, Nachtsheim and Wasserman 1996).

The function that expresses the probability of the observed data as a function of the unknown parameter vector $\boldsymbol{\beta}$ is the likelihood function. In order to implement the EM algorithm, the complete and incomplete likelihoods must be found. The incomplete data log-likelihood would be the likelihood function for the accepted applicants only given by:

$$\log L(\boldsymbol{\beta}) = \sum_{j=1}^{n} y_j \mathbf{x}_j \boldsymbol{\beta} - \ln[1+\exp(\mathbf{x}_j \boldsymbol{\beta})], \qquad (2)$$

and the complete data log-likelihood is the following:

$$\log L_C(\boldsymbol{\beta}) = \sum_{j=1}^{n} y_j \mathbf{x}_j \boldsymbol{\beta} - \ln[1+\exp(\mathbf{x}_j \boldsymbol{\beta})] + \sum_{j=n+1}^{m} y_j \mathbf{x}_j \boldsymbol{\beta} - \ln[1+\exp(\mathbf{x}_j \boldsymbol{\beta})]. \quad (3)$$

The derivation of the incomplete and complete data log-likelihood functions can be found in the Appendix.

The goal of maximum likelihood estimation is to find the unknown model parameters $\boldsymbol{\beta}$ that maximize the complete data log-likelihood given in (3). The complete data log-likelihood contains the unobserved outcome variable, y_j where $j=n+1,\ldots,n+m$, for the rejected applicants. Since the outcome variable is missing for the rejected applicants, the maximization procedure that finds $\boldsymbol{\beta}$ will be accomplished via the EM algorithm.

3. Estimation of the Parameters via the EM Algorithm

The complete data model specified in (3) treats the outcome variable, y_j, $j=n+1,\ldots,n+m$, for the rejected applicants as missing data. Thus, maximum likelihood estimates of $\boldsymbol{\beta}$ can be obtained via the EM algorithm by finding the parameter estimates, $\boldsymbol{\beta}$, that maximize the expected complete log-likelihood function, where the expectation is taken over the missing data given the observed data. The derivations of the E- and M-steps are given in the following paragraphs.

Using some initial value for $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^{(0)}$, the E-step requires the calculation of $H(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)}) = E\{\log L_c(\boldsymbol{\beta}) \mid \mathbf{x}, \boldsymbol{\beta}^{(0)}\}$, the expectation of the complete-data log-likelihood $\log L_C(\boldsymbol{\beta})$, conditional on the observed data and the initial fit $\boldsymbol{\beta}^{(0)}$ for $\boldsymbol{\beta}$. This step is performed here by simply replacing each unobserved outcome variable for the rejected applicants, y_j, $j=n+1,\ldots,n+m$, by its expectation conditional on $\mathbf{x}_j$, given by

$$E(Y_j \mid \mathbf{x}_j, \boldsymbol{\beta}^{(0)}) = P(Y_j = 1 \mid \mathbf{x}_j, \boldsymbol{\beta}^{(0)}) = \frac{1}{1+\exp(-\mathbf{x}_j\boldsymbol{\beta})}. \qquad (4)$$

That is, each y_j, $j=n+1,\ldots,n+m$ is replaced by the initial estimate of the posterior probability that the jth rejected applicant with characteristic vector $\mathbf{x}_j$ is a bad loan, and the $\boldsymbol{\beta}$ given in (4) is computed from the accepted applicants for the initial iteration of the E-step.

The M-step maximizes the function in (3). To find the value of $\boldsymbol{\beta}$ that maximizes the function in (3), we will use iteratively re-weighted least squares (IRLS). IRLS is a numerical algorithm that maximizes any specified objective function using a standard weighted least squares method (Kotz and Johnson 1983). For the problem at hand, the iterative solution to finding the value of $\boldsymbol{\beta}$ can be outlined as follows:

1. Choose initial estimates of the regression coefficients, such as the estimates of $\boldsymbol{\beta}$ using the accepted applicants only.

2. At each iteration, k, update the regression coefficients:
$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} + (\mathbf{X}\grave{}\mathbf{V}^{(k-1)}\mathbf{X})^{-1}\mathbf{X}\grave{}(\mathbf{y} - \mathbf{p}^{(k-1)})$$
where
X is the data set matrix of characteristic information for the accepted and rejected applicants,
y is the observed outcome vector (containing 0's and 1's),
$\mathbf{p}^{(k-1)}$ is the vector of fitted response probabilities for the previous iteration, the ith entry of which is
$$p^{i,(k-1)} = \frac{1}{1+\exp(-\mathbf{x}\boldsymbol{\beta}^{(k-1)})}$$
$\mathbf{V}^{(k-1)}$ is a diagonal matrix, with diagonal entries $p^{i,(k-1)}(1-p^{i,(k-1)})$.

3. Repeat steps 2 until $|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}|$ is close to zero.

The **X** data matrix contains accepted and rejected applicant characteristic information, so **X** contains no missing information. However, the vector **Y** has missing entries for the rejected applicants. In step 2, the problem of missing entries in **Y** is handled by computing the posterior probabilities of a rejected applicant being a bad loan as given in (4). Once the posterior probabilities are computed, the rejected applicant is allocated as a good loan (i.e., $y_j=0$) if the probability is less than the percentage of bad loans in the data or a bad loan (i.e., $y_j=1$) if the probability is more than the percentage of bad loans in the data. In essence, the cutoff probability to determine how to allocate rejected applicants as good or bad loans is determined by the prior probability of the data. This way of determining the cutoff probability has been used in the literature before (Banasik, Crook and Thomas 2003, Georges 2007).

The E- and M- steps are repeated until convergence is attained. Convergence of the algorithm is attained when there is a significantly small difference between $\log L(\boldsymbol{\beta}^{(k+1)}) - \log L(\boldsymbol{\beta}^{(k)})$.

4. Augmentation versus the EM Logistic Model as Reject Inference Techniques.

A standard augmentation procedure will be used to compare to the EM logistic model. Augmentation is a straightforward procedure to implement, consisting of two steps. The first step is to build a model on the accepted applicants in which the dependent variable being modeled is the good/bad status of the loan and then to apply this model to the rejected applicants to infer their probability of default. Second, a cutoff probability is determined, classifying the rejects as good or bad loans, and this information is added back into the model derived from the accepted population and re-modeled.

This study involves comparing the augmentation procedure to the EM logistic model through a simulation study. In summation, the following are the simulation steps.

1. Build a statistical model (logistic regression) using the accepted applicants with the SAS® German credit scoring data.
2. Use the model from step 1 to determine the good/bad loan status for the rejected applicants.
3. Sample the accepted and rejected applicants in such a way as to mimic true prior probabilities in credit scoring data sets.
4. Split the sampled data into in-house and future data sets.
5. Build the reject inference models (augmentation and EM logistic) using the in-house data set.
6. Apply the models to the future data set.
7. Test the performance of the reject inference techniques on the future data set.

Once the model is used in steps 1 and 2 to determine the good/bad status of the rejected applicants, steps 3-7 will be repeated many times in order to determine if using the EM logistic model performs better than augmentation as a reject inference technique. Figure 1 presents a graphical view of the steps involved in the simulation study.

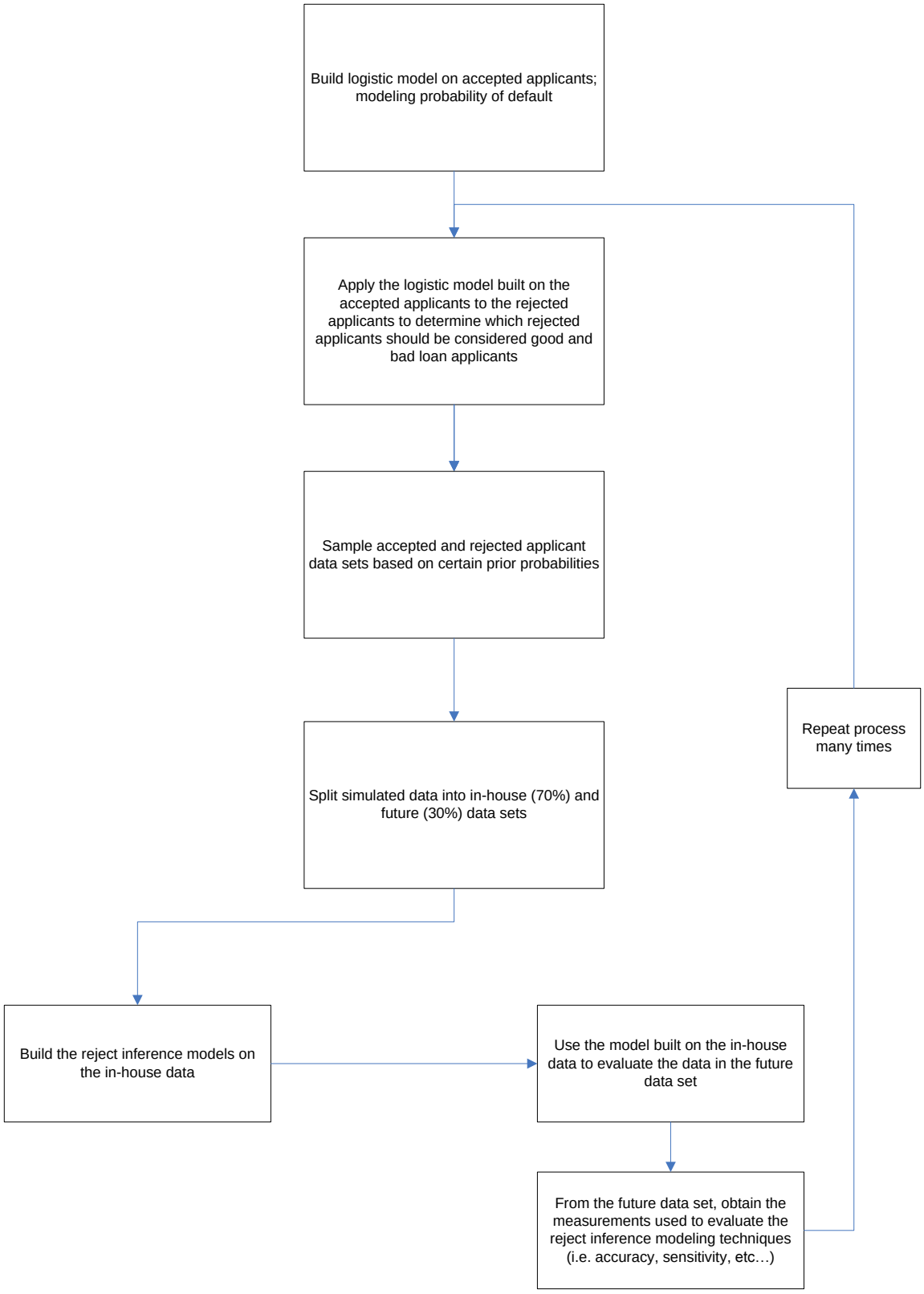

Figure 1: Flow Chart of the Simulation Process

4.1 Data

The data used in this simulation study will be the SAS® German credit scoring data. The data set consists of 3,000 accepted applicants and 1,500 rejected applicants. The characteristic variables used from the SAS® German credit data set are as follows: AGE, INCOME, TMJOB1, TMADD, EC_CARD, CHILDREN, PERSON_H, NMBLOAN, LOANS, FINLOAN, BUREAU, and RESID. Table 1 describes the variables used in the study. GB is the binary outcome variable that denotes whether the applicant was a good or bad loan. EC_CARD is a binary variable indicating whether or not the applicant has a checking bank card. NMBLOAN is the number of outstanding loans the applicant has with the financial institution where the applicant is applying for credit. NMBLOAN ranges from 0 to 2 loans. LOANS is the number of outstanding loans that the applicant has with the financial institution or any other financial organization. LOANS range from 0 to 9 loans. FINLOAN is a binary variable indicating whether or not the applicant has fully paid any outstanding loans the applicant may have with the financial institution.

Table 1: Variables Included in the Logistic Regression Model

Variable	Label	Measurement Level
GB	good/bad	categorical
AGE	age	continuous
INCOME	income	continuous
TMJOB1	time at job	continuous
TMADD	time at address	continuous
EC_CARD	bank card	categorical
CHILDREN	number of children	categorical
PERSON_H	number of persons in household	categorical
NMBLOAN	number of outstanding loans	categorical
LOANS	number of running loans	categorical
FINLOAN	number of finished loans	categorical
BUREAU	credit bureau rating	categorical
RESID	residential status	categorical

4.2 Credit Scoring Model

Based on the 3,000 accepted applicants, a statistical model using logistic regression is built to predict the probability of default. The dependent variable being modeled is the good/bad indicator variable, denoted by GB. GB=1 (bad status), if the applicant has defaulted, and GB=0 (good status) if the applicant has not defaulted. A logistic regression model was built using the accepted applicants' data, and this model was used to score (classify) the rejected applicants in order to determine a good/bad status for the rejected applicants. The classification from the logistic model will be used as the 'true' good/bad classification for the rejected applicants. The logistic model was applied only one time to the rejected applicants in order to determine their 'true' good/bad status. Accepted and rejected applicants were allocated in this fashion so that the 'true' good/bad status of the rejected applicants can be known in order to evaluate the reject inference procedures.

4.3 Sampled Applicants

The original German credit data consisted of half good and half bad loans in the accepted applicant population. The German credit data set has been oversampled (Siddiqi, 2006). Oversampling refers to cases where the proportion of good and bad cases in the data is different from the actual population. In such a case, the sample used to demonstrate the reject inference techniques will be adjusted so that the sample bad rate reflects a true-life credit scoring data set. The sample bad rate will be referred to as prior probabilities, and several prior probabilities will be examined in this simulation study. The data will be sampled based on several different prior probabilities. The first prior probability will assume a bad rate among the accepted applicants of 0.05 with a corresponding bad rate among the rejected applicants of 0.20. It is typical practice for the bad rate among the rejected applicants to be 2 to 4 times that of the bad rate among the

accepted applicants (Siddiqi, 2006). This simulation study will use a bad rate for the rejected applicants that are 4 times that of the accepted applicants. The next prior probability will assume a 0.10 bad rate among the accepted applicants with a corresponding 0.40 bad rate among the rejected applicants.

4.4 Cutoff Values

In each of the reject inference techniques, the end result is a posterior probability of default for each applicant. Based on these posterior probabilities, the applicant will be classified as a good or bad credit risk. The cutoff values used in this study will be 0.10, 0.15, and 0.20. If the applicant has a posterior probability of default below the cutoff value, the applicant will be designated as a good loan; otherwise the applicant will be deemed a bad loan.

4.5 In-House and Future Data Sets

Once the data has been sampled using the appropriate priors, the data set is split randomly into in-house and future data sets by using proc surveyselect within the SAS® software. The in-house data set consists of 70% of the data and the future data set consists of 30% of the data. A 70-30 split of the data is common in credit scoring practices (Siddiqi, 2006). The reject inference techniques will be built from the in-house data sets and the techniques will be evaluated on the future data set.

5. Model Performance Measures

Measuring the performance of the reject inference techniques is a key issue. Once the reject inference methods are performed, these techniques must be compared to each other to determine effectiveness. The problem is how to measure model performance. Naturally, the most effective model should be chosen for use in updating the credit scoring models. The measures used to

evaluate model performance are misclassification measures and the Kolmogorov-Smirnoff (KS) statistic.

5.1 Misclassification

While the statistical models designed in the simulation study are used to predict the probability of an applicant being good or bad, there is always the chance that an actual good applicant will be classified as a bad and therefore rejected, and vice versa. A confusion matrix is one such measure used to gauge the level of misclassification and used to compare different models (Siddiqi, 2006). The confusion matrix compares the number of true goods and bads against the number of predicted goods and bads for a particular statistical model. An example confusion matrix is shown in Table 2. From the confusion matrix, diagnostic measures such as accuracy, sensitivity and (1-specificicity) can be calculated. In this instance, accuracy is a measure of the overall performance of the modeling procedure. Accuracy is the percentage of goods and bads the statistical model correctly classifies. Using Table 1, accuracy is calculated as (A+D)/(A+B+C+D). Sensitivity is a measure of how many good loans the model classified correctly. Sensitivity is a percentage which can be calculated as A/(A+C). (1-specificity) is the percentage of loans that were incorrectly classified by the model. For purposes of this research, (1- specificity) will be defined as a false negative. In credit scoring, a false negative can be more costly than a false positive. A false negative occurs when a good loan is classified as a bad loan. (1-specificity) will be calculated as C/(A+C).

	Truth	
Predicted:	**Good**	**Bad**
Good	True Positive (A)	False Positive (B)
Bad	False Negative (C)	True Negative (D)

Table 2. Example of a confusion matrix

During each replication of the simulation study, the misclassification statistics will be collected. The simulation size of the study was based on an accuracy level of 0.50. The accumulated statistics will be analyzed from the future data sets for three groups; accepted applicants, rejected applicants, and all (accepted and rejected applicants). Conclusions will be drawn based on the empirical results.

6. Results

Figures 2-4 display the results from the simulation study with a prior probability of 0.05 among the accepted applicants (with a corresponding 0.20 prior probability for the rejected applicants). Each of the three graphs displays the model performance measures plotted against the three different cutoff values for the three different groups of interest (all applicants, accepted applicants, and rejected applicants). Each point in the plot represents the average model performance measures from performing the simulation study. From Figure 2, the EM logistic model outperforms the augmentation procedure for all groups across all cutoff values. Figures 3 and 4 show the EM logistic model also performs better in terms of sensitivity and (1-specificity) for all groups across all cutoff values.

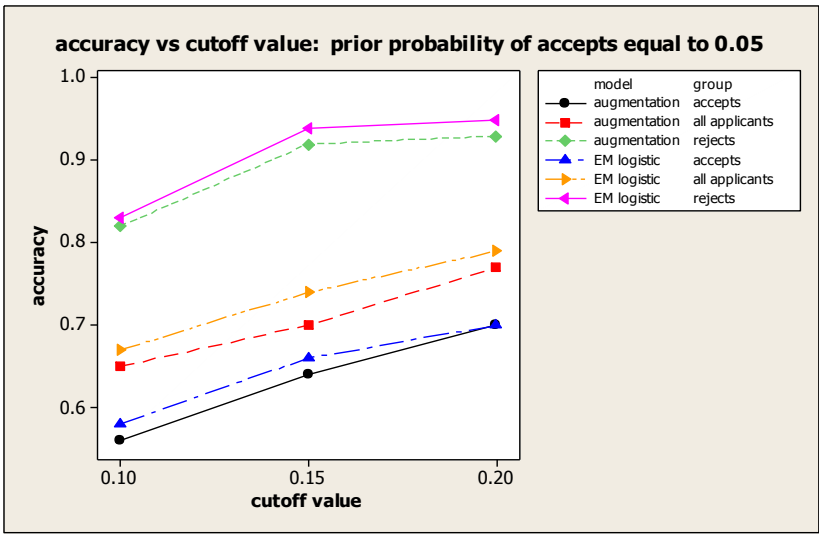

Figure 2: Accuracy versus cutoff values: prior probability of accepts equal to 0.05

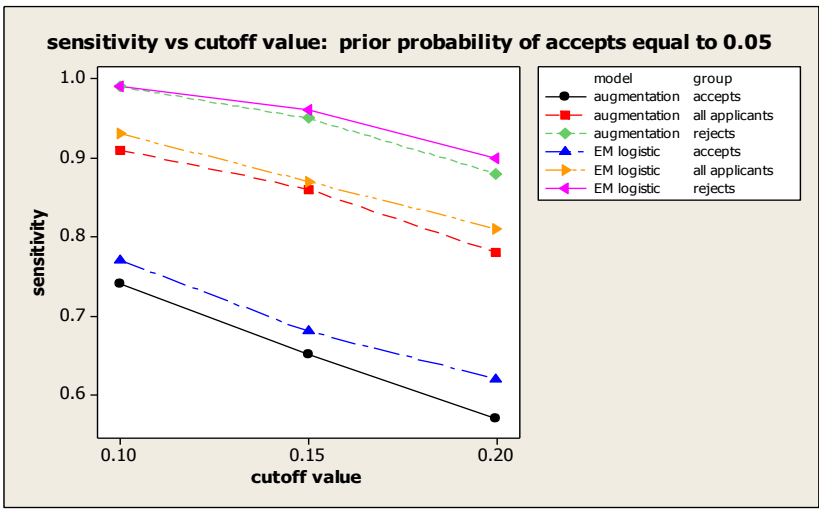

Figure 3: Sensitivity versus cutoff values: prior probability of accepts equal to 0.05

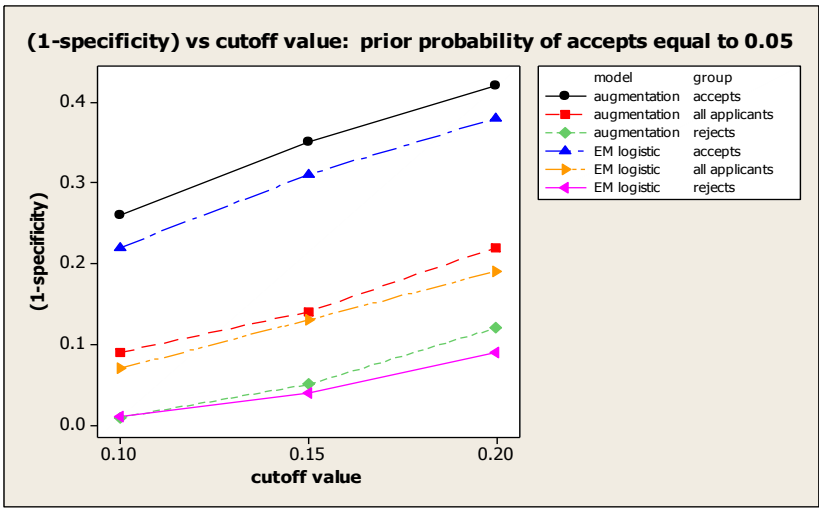

Figure 4: (1-specificity) versus cutoff values: prior probability of accepts equal to 0.05

Figures 5-7 display the same information as those in Figures 2-4, except using a 0.10 prior probability for the accepted applicants (with a corresponding 0.40 prior probability for the rejected applicants). As shown in Figures 5-7, when the percentage of bad loans is increased in both the accepted and rejected population, the EM logistic model outperforms the augmentation procedure in terms of all three model performance diagnostics for all groups of interest and across all cutoff values.

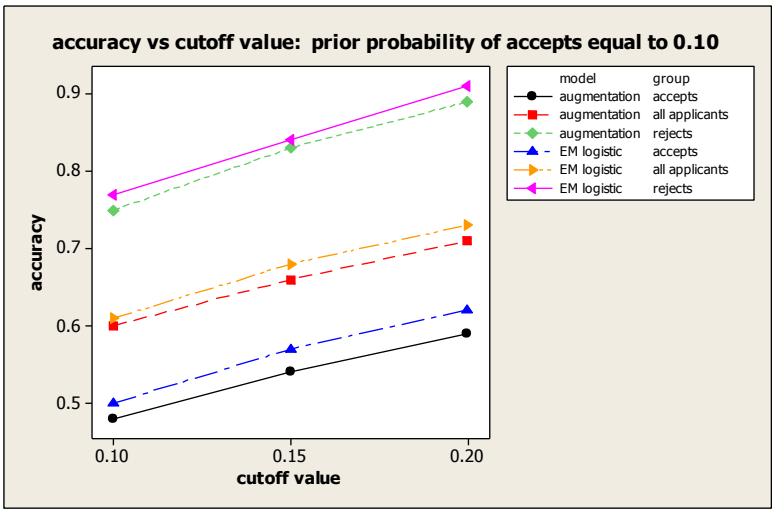

Figure 5: Accuracy versus cutoff values: prior probability of accepts equal to 0.10

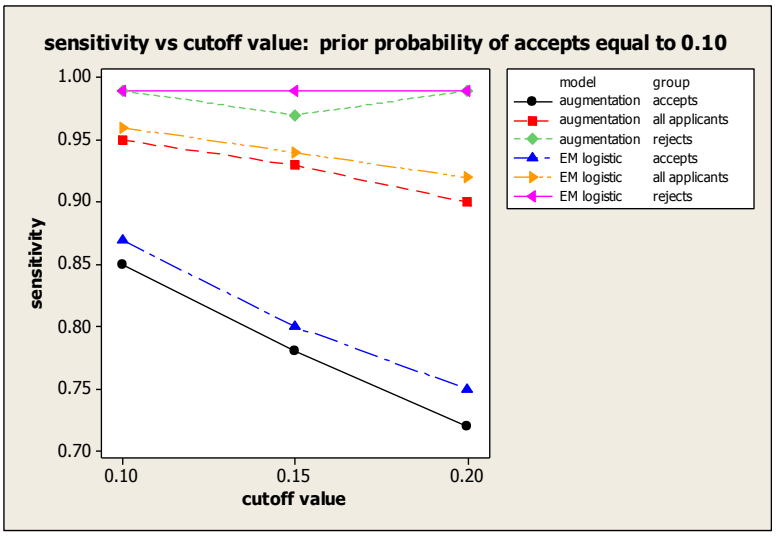

Figure 6: Sensitivity versus cutoff values: prior probability of accepts equal to 0.10

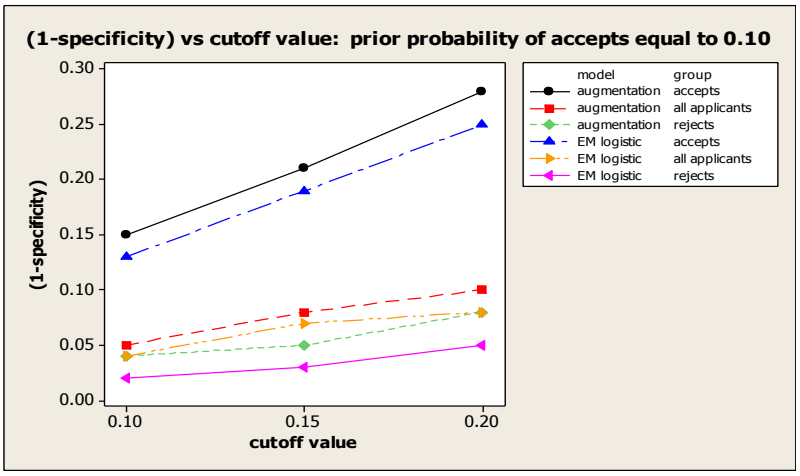

Figure 7: (1-specificity) versus cutoff values: prior probability of accepts equal to 0.10

7. Conclusions

We have presented a method for estimating parameters in a logistic regression model when some of the outcome variables are missing and the missing data mechanism is MAR. The major innovation of this method is that the EM algorithm can be used to handle the missing responses for the rejected applicants. To determine how well this estimation procedure worked in a reject inference context, a simulation study was performed which compared the EM logistic model to the standard augmentation reject inference procedure. As indicated in the simulation, the EM

logistic model outperformed the augmentation procedures across all settings of the simulation for each model diagnostic under study.

Based on the simulation study, the data analyzed, and the reject inference techniques under study in this paper, the following conclusions are qualifiedly drawn:

1. There is considerable support that using a mathematical algorithm such as the EM algorithm to perform reject inference can provide a more accurate credit scoring model over the standard augmentation procedures which are currently in practice.
2. Because data used in the simulation was limited to banks and financial institutions, no direct inferences concerning the applicability of the models to other economic segments can be drawn. It is possible, however, that the models can be useful in other rapidly growing areas of the market having similar economic characteristics.
3. Officials in banks and finance companies should consider the use of the EM logistic model as an alternative reject inference technique within their institution in order to update their credit scoring models due to the successful results presented in this study.

Certain implications can be drawn from the above conclusions: (1) for many lenders, profitability can be enhanced through a reduction of risk, that is, using a credit scoring model that is more selective in choosing between good and bad credit risks could bring fruitful profits to a financial institution; and (2) many lenders could likely benefit from the adoption of more advanced methods of reject inference in the management of their credit operations, in both individual credit decisions and in the aggregate.

A final implication, and one for which further research is needed, concerns the potential cost savings resulting from the use of an improved reject inference technique that helps better distinguish between good and bad credit risks compared to the current industry standard (i.e.,

augmentation). For example, from the results of this study, it was shown that on average, for particular settings of the simulation, the EM logistic model outperformed the augmentation procedure by as much as 2% in terms of accuracy. The question of interest is how much, in terms of profit, does that 2% increase in accuracy produce for a financial institution. As far as the author is aware, there are no published studies or research which discusses/shows how an improved reject inference technique would yield actual profit dollars for a financial institution. Statistics on the profitability of improved reject inference techniques should indicate whether potential added profitability to a financial institution is real or theoretical.

References

Banasik, J., Crook, J. N., and Thomas, L. (2003), "Sample Selection Bias in Credit Scoring Models," *Journal of the Operational Research Society*, 54, 822-832.

Georges, J. (2007), *Applied Analytics Using SAS® Enterprise Miner™ Course Notes*, Cary: SAS Institute Inc.

Hand, D. J., and Henley, W. E. (1997), "Statistical Classification Methods in Consumer Credit Scoring: A Review," *Journal of the Royal Statistical Society Series A*, 160, 523-541.

Horton, N. J., and Laird, N. M. (2001), "Maximum Likelihood Analysis of Logistic Regression Models with Incomplete Covariate Data and Auxiliary Information," *Biometrics*, 57, 34-42.

Kotz, S., and Johnson, N. L. (1983), *Encyclopedia of Statistical Sciences* (Vol. 4), New York: John Wiley.

Little, R. J. A., and Rubin, D. R. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. (1996), *Applied Linear Regression Models* (Vol. 3rd), New York: McGraw-Hill.

Siddiqi, N. (2006), *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, Hoboken, New Jersey: John Wiley.

Thomas, L. C. (2000), "A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers," *International Journal of Forecasting*, 16, 149-172.

Appendix: Derivation of the Complete and Incomplete Data Log-Likelihood Functions for the Modified Logistic Regression Reject Inference Technique.

Logistic regression involves directly modeling the probabilities of group membership. In multiple logistic regression, it is assumed that each observation j, $j=1,\ldots,n,n+1,\ldots,n+1,\ldots,n+m,$ are independent observations such that $Y_j|\mathbf{x} \sim$ Bernoulli(p_j) where the logistic model takes the form

$$p_j = P(Y_j = 1 \mid \mathbf{x}) = \frac{\exp(\mathbf{x\beta})}{1+\exp(\mathbf{x\beta})} = \frac{1}{1+\exp(-\mathbf{x\beta})}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ denotes the unknown p-vector of parameters to be estimated.

Since each Y_j is a Bernoulli random variable, where:

$$P(Y_j = 1 \mid \mathbf{x}) = p_j$$

we can represent its probability distribution as follows:

$$f(Y_j) = p_j^{\,y_j}(1-p_j^{\,y_i}) \qquad Y_j = 0,1;\ j = 1,...,n,n+1,...,n+m. \tag{2}$$

The observed (incomplete) data log-likelihood is the likelihood function that contains the n accepted applicants. Since Y_j observations are independent, the likelihood function is:

$$g(Y_1,...Y_n) = L(\boldsymbol{\beta}) = \prod_{j=1}^{n} f_j(Y_j) = \prod_{j=1}^{n} p_j^{\,y_j}(1-p_j)^{y_j}. \tag{3}$$

It will be easier to work with the logarithm of the likelihood function. Taking the logarithm of (A.3) yields:

$$\log L(\boldsymbol{\beta}) = \sum_{j=1}^{n} \log(p_j^{\,y_j}(1-p_j)^{y_j}). \tag{4}$$

Substituting in the formula for p_j and distributing the logarithm produces:

$$\log L(\boldsymbol{\beta}) = \sum_{j=1}^{n} y_j \mathbf{x\beta} - y_j \log(1+\exp(\mathbf{x\beta})) - (1-y_j)\log(1+\exp(\mathbf{x\beta})). \tag{5}$$

Certain terms in (A.5) cancel to produce the simplified observed data log-likelihood function:

$$\log L(\boldsymbol{\beta}) = \sum_{j=1}^{n} y_j \mathbf{x\beta} - \log(1+\exp(\mathbf{x\beta})). \tag{6}$$

The complete data log-likelihood is the likelihood function that contains the accepted and rejected applicants. The likelihood function that contains the accepted and rejected applicants can be expressed as:

$$g(Y_1,...Y_n,Y_{n+1},...,Y_m) = L_C(\boldsymbol{\beta}) = \prod_{j=1}^{n} p_j^{\,y_j}(1-p_j)^{y_j} + \prod_{j=n+1}^{m} p_j^{\,y_j}(1-p_j)^{y_j}. \tag{7}$$

Again, it is easier to work the logarithm, so taking the logarithm of (A.7) produces:

$$\log_C L(\boldsymbol{\beta}) = \sum_{j=1}^{n} \log(p_j^{\,y_j}(1-p_j)^{y_j}) + \sum_{j=n+1}^{m} \log(p_j^{\,y_j}(1-p_j)^{y_j}). \tag{8}$$

Substituting in the formula for p_j and distributing the logarithm yields:

$$\begin{aligned}\log L_C(\boldsymbol{\beta}) = &\sum_{j=1}^{n} y_j \mathbf{x}\boldsymbol{\beta} - y_j \log(1+\exp(\mathbf{x}\boldsymbol{\beta})) - (1-y_j)\log(1+\exp(\mathbf{x}\boldsymbol{\beta})) + \\ &\sum_{j=n+1}^{m} y_j \mathbf{x}\boldsymbol{\beta} - y_j \log(1+\exp(\mathbf{x}\boldsymbol{\beta})) - (1-y_j)\log(1+\exp(\mathbf{x}\boldsymbol{\beta})).\end{aligned} \tag{9}$$

Certain terms in (A.9) cancel to produce the simplified complete data log-likelihood function:

$$\log L_C(\boldsymbol{\beta}) = \sum_{j=1}^{n} y_j \mathbf{x}\boldsymbol{\beta} - \log(1+\exp(\mathbf{x}\boldsymbol{\beta})) + \sum_{j=n+1}^{m} y_j \mathbf{x}\boldsymbol{\beta} - \log(1+\exp(\mathbf{x}\boldsymbol{\beta})) \tag{10}$$