Bias Correction in Risk Assessment When Logistic Regression is used with An Unevenly Proportioned Sample between Risk and Non-Risk Groups

Timothy H. Lee, SPS Payment Systems
Donald T. Searls, University of Northern Colorado

Linear Logistic Regression is a simple but a very powerful tool to assess the likelihood of being in one "category" for an observation with specific independent characteristic values, i.e., when the response variable is dichotomous and the data is replicated, the conditional probability, that an observation belongs to one of the two categories given independent characteristic values, can easily be estimated through Logistic Regression. For various reasons, stratified sampling, sometimes, causes a different sample proportion between the two groups from the population. Many statistical packages allow their users to adjust weights to fix this bias problem as an option in using the Logistic Procedure. The users, however, would experience more computing cost by using the option. In many cases, the purpose of the biased sampling is for computational economy and if the computing cost stays the same, using the biased sample with adjusted weights is not advantageous.

In this study, simple bias correction without using adjusted weights is explained using simulated bankruptcy data. Since the method can be used for any software without adjusting weights, computational economy can be achieved with unbiased results.

1. Introduction

Two group classification techniques are instrumental in many cases of decision making in business, finance, and marketing, etc. For example, when credit grantors extend credit, they need to assess each applicant's credit worthiness or risk for the extension of credit. In marketing analysis, they want to target more potentially responsive populations for direct mailing. These examples are typical cases where the dependent variable is binary, i.e., risk versus non-risk, or response versus non-response. Logistic regression analysis, parametric or nonparametric discriminant function analysis, and neural net are usual candidate tools for such cases. These methods are known to be comparable one to another in terms of classification accuracy. Each of these has merits and demerits depending on the user's point of view such as cost, purpose of analysis, etc. Logistic regression, for many reasons, often has been preferred to other methods, especially to discriminant function analysis. Press and Wilson (1978) made empirical applications to compare logistic regression and discriminant function analysis using breast cancer data and population change data of the U.S. They concluded that Logistic regression outperforms linear discriminant function analysis when the normality assumption is violated. Fienberg (1980), also, mentioned the superiority of logistic regression over discriminant function analysis in case of non-normal populations. In reality, the normality assumption is not easily met, especially in most of the credit or demographic profile data. One of the advantages of using the logistic regression model is that it provides the likelihood of being in one group for an observation given characteristic profile values.

Let E be an event that an observation is from one category and a vector $\underline{x}$ be the characteristic values of the observation. Then, the logistic regression model is

$$p(x) = \Pr\{E \mid \underline{x}\} = 1/[1+ \exp\{ - (\alpha + \underline{\beta}\underline{x})\}],$$

where $(\alpha, \underline{\beta})$ are unknown parameters that are to be estimated from the sample. This model is used to classify an observation into one of the two mutually exclusive categories based on $\underline{x}$.

In actual analysis, the binary dependent variable, usually coded 0 or 1 for event or non-event, is regressed on $\underline{x}$.

1.1 Sample Bias

In many cases of two group classification, the proportion of one group is far smaller than the other.

For instance, the proportion of cancer patients among the population, or the proportion of bankrupt accounts in a portfolio is observed to be very low. In such a case, analysts would rather choose stratified random sampling than simple random sampling. For instance, n observations are taken randomly from the event population, and m observations are taken from the non-event population. The sample ratio between event and non-event in such a sample is quite different from that in the population. For classification purposes, such an uneven proportion shouldn't be a problem, because a classification model developed on an unevenly proportioned sample would work as well as a model developed on an evenly proportioned sample. Such sampling scheme saves sampling cost and in using the data, later on, will bring a reduction of computing cost as well. Immeasurability is, of course, sometimes a cause of the uneven proportion. In this study, we like to consider sample bias in the sense of an uneven or distorted ratio between two mutually exclusive categories.

1.2 Model Bias

If a logistic regression model is derived based on a biased sample, the estimated probability of event given $\mathbf{x}$ would be either underestimated or overestimated even though the model has almost the same classification power as that derived from unbiased data. Let's consider, as a more detailed example, a case when bankruptcy is an event. That is, a model is developed to assess likelihood of bankruptcy given a vector of characteristic values. The risk assessment is biased if the model is developed by using biased data.

2. Analysis of Data

In this study, we used simulated bankruptcy data from Moody's Industrial Manuals 1968-1972 to expand our discussion. The data set has 4 independent variables, x_1 = (cash flow)/(total debt), x_2 = (net income)/(total assets), x_3 = (current assets)/(current liabilities), and x_4 = (current assets)/(net sales). The dependent variable is coded as 0 for bankruptcy and 1, otherwise.

For illustration, let's assume that the proportion of the event (bankruptcy) is 1/50 (=0.02) in a portfolio. A logistic model was derived using a biased development sample which has proportion of event (bankruptcy) 1/3 ($\cong$ 0.33). The parameter estimates on the biased sample were

$$(\alpha', \beta_1', \beta_2', \beta_3', \beta_4')$$

$$= (\ 2.8603, -\ 3.6938, -\ 1.7649, -\ 1.7286,\ \ 0.4760\ ).$$

Figure-1 in the Appendix presents plots between estimated risk versus observed risk for the biased sample. A smooth curve produced by the authors' robust smoother (1990) is superimposed to enhance the visual information. Figure-2 presents the same plots on an unbiased sample which has the same proportion as the population. We can observe that there is, in Figure-1, a strong linear relationship (almost a 45 degree line with some endurable noise) between observed and estimated risks, while, in Figure-2, there is no linearity between the two values and it presents a bias assessment of risk. In most cases, the bias is leaning toward over estimation. That is, when the proportion of the event is very low such as bankruptcy, the sample proportion of the event is usually far higher than the population proportion and may result in an overestimation of risk unless an adjustment is made in the process of estimation.

3. Bias Correction

We can consider two kinds of corrections, i.e., a priori adjustment and a posteri correction.

3.1 A priori Adjustment

One of the easy ways of a priori adjustment is to assign proper weights based on the sampling fraction, f = n/N, where, n and N are sample and population sizes, respectively. If, in the case of stratified sampling, f is 0.5 for a stratum, the corresponding sample weight $1/f$ =2 will be assigned in the estimation procedure. This kind of adjustment is allowed, in most of the commercial software, for the price of additional computing cost. To compute estimates of the parameters, Iteratively Reweighted Least Squares (IRLS) or similar methods are used. For example, IRLS for k+1 response categories is used, in SAS, as in the following:

Let $\mathbf{Z}_j = (Z_{1j}, \ldots, Z_{(k+1)j})^t$ be a multinomial vector such that

$$Z_{ij} = 1 \quad \text{if } Y_j = i$$

$= 0$ otherwise , for $j = 1, ..., n$

(In two group case, $k = 1$ and Y is a binary response variable)

Let $\mathbf{p}_j = E(\mathbf{Z}_j)$, $\mathbf{V}_j = Cov(\mathbf{Z}_j)$,
and

$\gamma^t = (\alpha_1, \alpha_2, ..., \alpha_k, \beta)$.

And, let $\mathbf{D}_j$ be the matrix of partial derivatives of $\mathbf{p}_j$ with respect to γ. Then, the estimating equation for the regression parameters is

$$\Sigma_j \mathbf{D}_j^t \mathbf{W}_j (\mathbf{Z}_j - \mathbf{p}_j) = 0,$$

where $\mathbf{W}_j = w_j \mathbf{V}_j^-$, w_j is the weight of j-th observation, and $\mathbf{V}_j^-$ is a generalized inverse of $\mathbf{V}_j$. $\mathbf{V}_j^-$ is chosen as the inverse of the diagonal matrix with $\mathbf{p}_j$ as the diagonal. The parameters are estimated iteratively as

$$\gamma'_{m+1} = \gamma'_m + (\Sigma_j \mathbf{D}'^t_j \mathbf{W}'_j \mathbf{D}'_j)^{-1} \Sigma_j \mathbf{D}'^t_j \mathbf{W}'_j (\mathbf{Z}_j - \mathbf{p}'_j)$$

Where $\mathbf{D}'_j$, $\mathbf{W}'_j$, and $\mathbf{p}'_j$ are evaluated values of $\mathbf{D}_j$, $\mathbf{W}_j$, and $\mathbf{p}_j$ at γ'_m.
If the likelihood evaluated at γ'_{m+1} is less than that evaluated at γ'_m, then γ'_{m+1} is recomputed using half the value of the second term of the right hand side.

As was discussed, by assigning proper weights, if it is allowed, or by replicating $1/f$ (to the nearest integer) times, if weighting is not allowed, the sample bias problem in risk assessment can be easily overcome with additional expense.

Our interest, however, is not in a priori adjustment but in a posteri correction. When a model is developed already and the development data is no longer available, or redevelopment causes unexpected inconvenience or cost, posterior correction based on minimal information about the population would be an economical and efficient alternative.

3.2 A posteri Correction

This approach is used to alleviate a biased risk estimation due to an uneven sampling fraction by computing a simple correction factor. For illustration, assume a situation that a probability model is derived using biased data and it is applied in an application data set. The application data is not used for the derivation of the model. We assume, further, that the proportion of the event in the application data will be approximately the same as that of the population. The probability of the event predicted will be biased and it should be corrected. To simplify the discussion, let's define the following:

p: population proportion of events

p': sample proportion of events in a biased data set

ϕ': estimated likelihood of an event for given $\mathbf{x}$ on an application data using the biased model developed on the biased data set

m: number of events observed at ϕ' in the biased data set

n: number of non-events observed at ϕ' in the biased data set

M: total number of events in the biased data set

N: total number of non-events in the biased data set

Further, let:

f' = m/M (Relative frequency of event at ϕ' in the biased data set)

g' = n/N (Relative frequency of non event at ϕ' in the biased data set)

Then, the likelihood of event for an observation estimated on the application data, even though the data is not biased, would be,

$$\phi' \cong m / (m + n)$$

$$= f' * M / (f' * M + g' * N) \quad \text{..... (1)}$$

Since the model was derived on the biased data, ϕ', the conditional probability given characteristic values $\mathbf{x}$, is biased although it is calculated on the application data. It always results in the same likelihood for $\mathbf{x}$ and implies the same likelihood as if it were calculated on the biased sample.

The true likelihood of event at ϕ' can be calculated by,

$$\phi = p * f / [p * f + (1-p) * g] \quad \text{..... (2)}$$

,where f and g are population relative frequencies for event and non-event, respectively.

The problem is how to estimate (or approximate) ϕ in (2) using ϕ' in (1).
One necessary condition that can be easily proven empirically is that

$$f' \cong f \text{ and } g' \cong g \text{ for any } \phi' \text{ and } p'.$$

From (1), using above condition,

$$[\phi']^{-1} - 1 = (g'/f') * (N/M)$$

$$\cong (g/f) * (N/M) \quad \text{..... (3)}$$

By multiplying (M/N) * [(1-p)/p] and adding 1 on both sides of (3),

$$\{(\phi')^{-1} - 1\} * (M/N) * [(1-p)/p] + 1$$

$$\cong [f * p + g * (1-p)] / (f * p) \quad \text{..... (4)}$$

From (2) and (4), we get,

$$\phi \cong \{ [(1-\phi')/\phi'] * (M/N) * [(1-p)/p] + 1\}^{-1},$$

or by using the fact that (M/N) = [p'/(1-p')], we get

$$\phi \cong \{ [(1-\phi')/\phi'] * [p'/(1-p')] * [(1-p)/p] + 1\}^{-1}_{f}$$

The last formula is for bias correction. It shows that the biased likelihood ϕ' can be corrected easily and the only necessary information about the population is the proportion of the event. The formula was applied to the estimated likelihood of event (estimated risk) in Figure - 2 and the corrected risk and observed risk is plotted in Figure - 3. A strong linear relationship is found between the estimated risk and the observed risk, particularly for an observed risk under 20%. This is the region where most of the observed risks occur. This shows that the biased risk is corrected.

4. Discussions

As mentioned above, Logistic regression is a very popular tool in classification analysis. Especially in the two group case such as risk versus non-risk analysis, it is very instrumental in assessing risk level for an observation in a portfolio. An uneven proportion, however, will cause a biased estimation. In business applications, the size of the risk group is usually small compared to the portfolio size. For example, in developing a bankruptcy forecasting model for a portfolio, the number of bankruptcies is very low so all the bankruptcies are taken into the development sample along with a certain number of non-bankruptcies. Even though the resulting model has good separation power when measured by the Kolmogorov-Smirnov test, Apparent Error Rate, or Kull-back Leibler information value, etc., the risk measured by the model would be overly assessed. For worse scenarios, redevelopment of the model is impossible because the original data was purged, or the biased model is installed on the system already and is in production mode. In such cases, a posteri correction is very handy.

Even when the weight option is available in using statistical software, if the weight assigned to one group is too large compared to the other, such as the bankruptcy prediction case, the resulting estimates of risk may not be accurate when round-off error or wrong direction of convergence is cumulated in the process of the iterative reweighted algorithm of the logistic procedure. If such a situation is expected, both the weight assignment and the above correction algorithm can be used for a test.

Acknowledgements

The authors wish to acknowledge Dr. Sam Houston of University of Northern Colorado for his careful review of the paper.

References

Press, S.J. and Wilson, S. (1978), *Choosing between Logistic Regression and Discriminant Analysis*, Journal of the American Statistical Association, Vol. 73, 699 - 705

Fienberg, Stephen (1980), *The Analysis of Cross-Classified Categorical Data*, second ed., The MIT Press.

Lee, Timothy and Searls, Donald (1990), *Two Stage Smoothing of Scatterplots*, ASA Proceedings, Statistical Graphics Sec.

SAS user's Guide, Version 6 (1990), Vol. 2, 1071 - 1126.

Charles W. Therrien (1989), *Decision, Estimation, and Classification: an introduction to pattern recognition and related topics*, John Wiley & Sons.

Richard A. Johnson and Dean W. Wichern (1982), *Applied Multivariate Statistical Analysis*, Prentice-Hall.

Ronald H. Randles and Douglas A. Wolfe (1979), *Theory of Nonparametric Statistics*, John Wiley & Sons.

William G. Cochran (1977), *Sampling Techniques*, third ed., John Wiley & Sons.

Appendix

Figure - 1:
Estimated Risk versus Observed Risk on the biased development data set

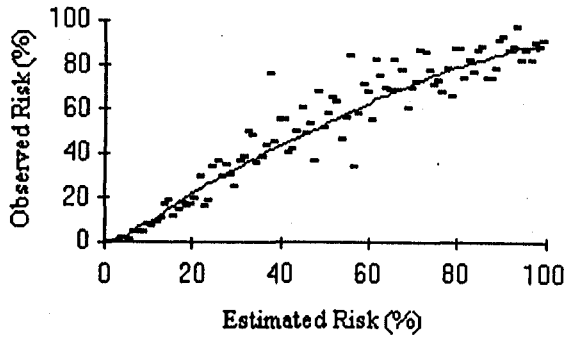

Figure - 2:
Estimated Risk versus Observed Risk on an unbiased application data set

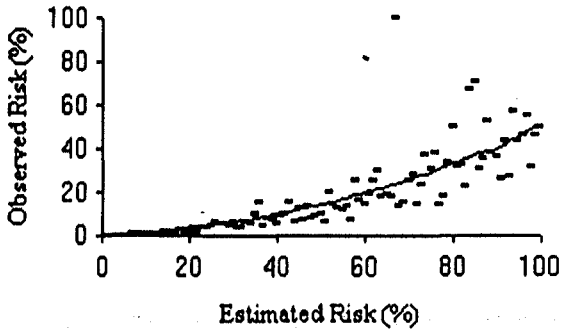

Figure - 3:
Corrected Risk versus Observed Risk on an unbiased application data set