International Journal of Mathematics and Statistics Studies

Vol.1, No.3, pp. 29-37, September 2013

Published by European Centre for Research Training and Development UK (www.ea-journals.org)

COMPARISON OF JACKKNIFE AND RESUBSTITUTION METHODS IN THE ESTIMATION OF ERROR RATES IN DISCRIMINANT ANALYSIS

G.A. Osuji and S.I Onyeagu
Department Of Statistics, Nnamdi Azikiwe University, Awka

D.D. Ekezie
Department Of Statistics, Imo State University, Owerri

ABSTRACT:*Most often, in classification procedures, error rates or probability of misclassification are assessed. Because in real life application of classification rules or methods, some errors of misclassification can be more costly than the others. In this work, two methods of estimating error rates, namely; the Jackknife and resubstitution methods are examined using ten samples of size* $n_1 = n_2 = 30, 900$ *from the population pair* $p_1 = (.3,\ .3,\ .3)$ *and* $p_2 = (.4, .4, .4)$. *From the results obtained from the experiments, we observed that the resubsitution method performed* better than the Jackknife method in estimating the exact probabilities of misclassification.

KEYWORDS: Jackknife, Resubstitution, Misclassification, Error Rates.

INTRODUCTION

Two relevant methods as regards to the assessment of bias and standard error are the bootstrap and Jackknife methods. In discriminant analysis, many works have been done as on how to estimate the probability of misclassification and or the apparent error rate. Closely related to the Jackknife is the idea of cross validation which is mostly used in the more specialized area of model choice and assessment performance of a prediction or allocation rule. (see krzanowski 1993). However, this method of estimation (Jackknife) have not been applied extensively on the Full multinomial, First and Second order Bahadur and the optimal procedures.

In this procedure, each sample member is omitted in turn from the data, there by generating n separate samples each of size $n - 1$. The advantage of this method when it is being applied is that it requires only one explicit inversion, namely, the inversion of the sample variance-covariance matrix based on the entire sample. The new rule which we are proposing based on the Jackknife method will be applied on two different population structures in order to see how it performed.

The jackknife procedure omits one observation, develops the classification function using all other observations (n1 + n2 – 1) and uses the classification function to classify the excluded observation. This process is repeated for each of the observations. The estimates of error rates obtained are unbiased estimates of error rates for a classification rule based on n1 + n2 – 1 observation. Using the jackknife method for the full multinomial, first and second order Bahadur and the optimal procedures require no matrix inversion. Using the jackknife method on the linear discriminant function involves the computation of n1 + n2 -1 discriminant functions and n1 + n2 -1 matrix inversions. Since the computation of each discriminant function requires a matrix inversion, this method was devised to reduce the number of inversions. Otherwise for large dimensional problems this would be too time consuming. The proposed method requires only one explicit inversion, namely, the inversion of the sample variance –covariance matrix based on the entire sample

REVIEW OF RELATED LITERATURE

Let $\widehat{\Theta}_{(i)}$ denote the value of $\widehat{\Theta}$ obtained from the i^{th} of these samples, that is when the i^{th} sample member is omitted from the calculation, and let Θ denote the average of the n values $\widehat{\Theta}_{(i)}$. Then the Jackknife estimate of the standard error of $\widehat{\Theta}$ is given by

$\widehat{\Theta}_j = \{(\frac{n-1}{n})\sum_{i=1}^{n}(\widehat{\Theta}_i - \overline{\Theta})^2\}^{1/2}$ (krzanowski 1993)

Krzanowski further stated that because of the explicit instructions for obtaining each of the sub-samples in the Jackknife procedure, it is possible to find an analytical expression for $\widehat{\Theta}_j$ in those cases where ϑ has a simple algebraic form.

Bartlett (1951) also used an identity given by $B = A + UV'$ and obtained

$$B^{-'} = \frac{A^{-'}U.V'A^{-'}}{1+V'A^{-'}U}$$

where A and B are square non-singular matrices while U is a column vector and v' is a row vector.

This inverse and/or identity we shall use in our proposed method.

Goldstein and Dillion (1978) also noted that Lachenbruch's estimate is sometimes in accurately referred to as a Jackknife estimate. Following a suggestion made by Miller (1974), the actual and the apparent error being biased estimates of t^* are logical candidates for the Jackknife method. In its most simplistic form, let $\widehat{\Theta}$ be an estimate of a parameter Θ based on a sample of size N-1 formed by deleting the j^{th} observation.

Define

$\widehat{\Theta}_j = N\widehat{\Theta} - (N-1)\widehat{\Theta}_{-j}$, $j = 1, 2, \ldots,$ N

Then the estimator

$$\hat{\theta}_j = \frac{1}{N}\sum_{j=1}^{N}\hat{\theta}_j = N\hat{\theta} - \frac{(N-1)}{N}\sum_{j=1}^{N}\hat{\theta}_j$$

which has the property that it estimates the order $\frac{1}{N}$ term from a bias of the form

$$E(\hat{\theta}) = \theta + \frac{\alpha_1}{N} + \theta \frac{1}{N^2}$$

If $\hat{\theta}$ represents the actual error $t(\hat{D})$ based on a sample of size $N = N_1 + N_2$ from the mixed population, then the Jackknife actual error is given by

$t(\hat{D}) = Nt(\hat{D}) - \frac{N-1}{N}\sum_{j=1}^{N} t_{-j}(\hat{D})$ where $t_{-j}(\hat{D})$ is the actual error based on $N_1 + N_2 - 1$ observations.

Again, for normal approximation arguments in determining estimates of actual error see Cochran and Hopkins (1961), Hills (1966) and for a full discussion of the Jackknife and the boostrapping methods see Efron (1982).

PROPOSED METHOD

The proposed method is an extension of the Jackknife method. For a k-population case, we have the following structures

___|_________)[_________|________)[_______|________)[__________)[____|_______)[_

μ_1 μ_2 μ_3 ... μ_k

However, let $X_1,...,X_n$ be a sample from population π_1 and $Y_1,...,Y_n$ be a sample from population π_2. Let $\bar{X}$ *and* S_x be the mean vector and variance-covariance matrix of the sample from $\pi_1, \bar{Y}$ and S_y be the mean vector and variance covariance matrix of the sample from π_2.

$$S_x = \sum_{i=1}^{n_1} (x_i - \bar{x})(x_i - \bar{x})'$$

$$S_y = \sum_{i=1}^{n_2} (y_i - \bar{y})(y_i - \bar{y})'$$

The pooled covariance matrix for the two samples is

$$S = \frac{(n_1 - 1)S_x + (n_2 - 1)S_y}{n_1 + n_2 - 2}$$

Suppose we leave out the k^{th} observation X_k from π_1, then the new mean of the sample from π_1 becomes

$$\bar{X}_k = \frac{\sum_{\substack{i=1 \\ i\neq k}}^{n_1} X_{1i}}{n_1 - 1} = \frac{n_1\overline{X} - \overline{X_{1k}}}{n_1 - 1} \quad i \neq k$$

The variance will be

$$(n_1 - 2)S_x^k = \sum_{\substack{i=1 \\ i\neq k}}^{n_1} \left(X_{1i} - \overline{X}_{ik}\right)\left(X_{1i} - \overline{X}_{ik}\right)^{\prime} i \neq k$$

$$= \sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_{ik}\right)\left(X_{1i} - \overline{X}_{ik}\right)^{\prime} - \left(X_k - \overline{X}_k\right)\left(X_k - \overline{X}_k\right)^{\prime}$$

We then have

$$\left(X_{1i} - \overline{X}_k\right) = X_i - \frac{n_1\overline{X} - X_k}{n_1 - 1} = \frac{n_1\left(X_i - \overline{X}\right) + \left(X_k - X_i\right)}{n_1 - 1}$$

and

$$\left(X_k - \overline{X}_k\right) = X_i - \frac{n_1\overline{X} - X_k}{n_1 - 1} = \frac{n_1\left(X_k - \overline{X}\right)}{n_1 - 1}$$

and the new covariance will be

$$(n_1 - 2)S_{1k} = \sum_{i=1}^{n_1} \left(X_1 - \overline{X}_k\right)\left(X_i - \overline{X}_k\right)^{\prime} i = k$$

$$= \sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_k\right)\left(X_{1i} - \overline{X}_k\right)^{\prime} - \left(X_{ik} - \overline{X}_k\right)\left(X_{ik} - \overline{X}_k\right)^{\prime}$$

$$\left(X_{1i} - \overline{X}_k\right) = X_{1i} - \frac{n_1\overline{X} - X_k}{n_1 - 1}$$

$$= \frac{n_1\left(X_{1i} - \overline{X}\right) + \left(X_k - X_{1i}\right)}{n_1 - 1}$$

$$\left(X_k - \overline{X}_k\right) = X_k - \frac{n_1\overline{X} - X_k}{n_1 - 1} = \frac{n_1\left(X_k - \overline{X}\right)}{n_1 - 1}$$

$$\sum_{i=1}^{n_1}\left(X_{1i}-\overline{X}_{ik}\right)\left(X_{1i}-\overline{X}_{1k}\right)^{/}=\sum_{i=1}^{n_1}\left[n_1\left(X_{1i}-\overline{X}\right)+\left(X_k-X_{1i}\right)\right]\frac{n_1\left(X_{1i}-\overline{X}\right)+\left(X_k-X_{1i}\right)}{\left(n_1-1\right)^2} \quad \text{…. (3.1)}$$

Again

$$\sum_{i=1}^{n_2}\left(X_k-\overline{X}_k\right)\left(X_k-\overline{X}_k\right)^{/}=\left(\frac{n_1}{n_1-1}\right)^2\left(X_k-\overline{X}\right)\left(X_k-\overline{X}\right)^{/}$$

$$=\left(\frac{n_1}{n_1-1}\right)^2\left(X_kX_k^{/}-2\overline{X}^{/}+\overline{X}\,\overline{X}^{/}\right)$$

Equation 3.1 now reduces to

$$\sum_{i=1}^{n_1}\left(X_i-\overline{X}_k\right)\left(X_i-\overline{X}_k\right)^{/}=\left(\frac{1}{n_1-1}\right)^2\left[n_1^{\,2}(n_1-1)S_x-2n_1X_k\overline{X}^{/}+2n_1^{\,2}\overline{X}\,\overline{X}^{/}+(1-2n_1)\sum_{i=1}^{n_1}X_iX_i^{/}\right]$$

Therefore

$$(n_1-2)S_{1k}=\sum_{i=1}^{n_1}\left(X_i-\overline{X}_k\right)\left(X_i-\overline{X}_k\right)^{/}-\left(X_k-\overline{X}_k\right)\left(X_k-\overline{X}_k\right)^{/}$$

$$=\frac{1}{(n_1-1)^2}\left[n_1^{\,2}(n_1-1)S_x-2n_1X_k\overline{X}^{/}+2n_1^{\,2}\overline{X}\,\overline{X}^{/}+n_1X_kX_k^{/}+(1-2n_1)\sum_{i=1}^{n_1}X_iX_i^{/}\right]-\left(\frac{n_1}{n_1-1}\right)^2$$

$$X_kX_k^{/}-2X_k\overline{X}^{/}(n_1-2)(n_1-1)^2S_{ik}=\quad n_1^{\,2}(n_1-1)S_x-2n_1X_k\overline{X}^{/}+2n_1^{\,2}\overline{X}\,\overline{X}^{/}+n_1X_kX_k^{/}+(1-2n_1)$$

$$\left[(n_1-1)S_x+n_1\overline{X}\,\overline{X}^{/}\right]-n_1^{\,2}X_k\overline{X}_k+2n_1^{\,2}X_k\overline{X}^{/}-n_1^{\,2}\overline{X}\,\overline{X}^{/}$$

$$(n_1-2)(n_1-1)^2S_{1k}=(n_1-1)^3S_x+n_1(1-n_1)\left(\overline{X}\,\overline{X}^{/}-2\overline{X}_k\overline{X}+X_k\overline{X}_k^{/}\right)$$

$$=(n_1-1)^3S_x+n_1(1-n_1)\left(\overline{X}-\overline{X}_k\right)\left(\overline{X}-\overline{X}_k\right)$$

Finally we have

$$(n_1-2)S_{ik}=(n_1-1)S_x-\frac{n_1}{(n_1-1)}\left(X_k-\overline{X}\right)\left(X_k-\overline{X}\right)\text{...............3.2}$$

But

$S_{1k}=\dfrac{(n_1-2)S_x^{\,k}+(n_2-1)S_y^{\,-1}}{n_1+n_2-3}$ is the pooled covariance matrix when the sample point X_{1k} is left out. Note that

$$(n_1+n_2-3)S_k = (n_1-1)S_x + (n_2-1)S_y{}^{-1} - \frac{n_1}{n_1-1}\left(X_k-\overline{X}\right)\left(X_k-\overline{X}\right)^{/}$$

$$= (n_1+n_2-2)S - \frac{n_1}{n_1-1}\left(X_k-\overline{X}\right)\left(X_k-\overline{X}\right)^{/}$$

$$S_k = \frac{(n_1+n_2-2)S}{(n_1+n_2-3)} - \frac{n_1}{n_1-1}\frac{\left(X_k-\overline{X}\right)\left(X_k-\overline{X}\right)^{/}}{n_1+n_2-3}$$

From Bartletts indentify

$$B = A + UV^{/}$$

$$B^{-1} = \frac{A^{-1} - A^{-1}U.V^{/}A^{-1}}{1+V^{/}A^{-1}U}$$

Let $A = \frac{(n_1+n_2-2)S}{n_1+n_2-3}$ and

$V = \overline{X} - X_k$, we then have that

If X_k is from π_1, then

$$\left(S^{1k}\right)^{-1} = \frac{n_1+n_2-3}{n_1+n_2-2}S^{-1} - \frac{(n_1+n_2-3)}{n_1+n_2-2}S^{-1}\frac{n_1}{n_1-1}\frac{\left(X_{1k}-\overline{X}\right)\left(\overline{X}_1-\overline{X}_{1k}\right)}{(n_1+n_2-2)}$$

$$1+\frac{n_1}{(n_1-1)(n_1+n_2-2)}\left(\overline{X}_1-\overline{X}_{1k}\right)S^{-1}\left(X_{1k}-\overline{X}_1\right)$$

and the discriminant function computed without sample point X_{1k} is given by

$$D_{1k}(X_{1k}) = \frac{(n_1-n_2-3)}{n_1+n_2-2}\left\{X_{1k}-\tfrac{1}{2}\left(\overline{X}_1+\overline{X}_2\right)+\frac{\left(X_{1k}-\overline{X}_1\right)}{2(n_1-1)}\right\}\left[(S_{1k})^{-1}-\left(\overline{X}_1-\overline{X}_2\right)-\frac{\left(X_{1k}-\overline{X}_1\right)}{(n_1-1)}\right]$$
…3.3

From equation 3.3, estimates of probabilities of misclassification are computed by summing the number of cases that were classified from each population and dividing by the number in each population which is regarded as cross- validation.

SAMPLING EXPERIMENTS AND RESULTS

In order to compare the performance of the two methods of estimating error rates, namely, the Jackknife and the Re-Substitution methods, we generated 10 samples of size $n_1 =$

n_2 =30, 900 from the population pair $p_1 = (.3,.3,\ .3)$ and $p_2 = (.4,.4,.4)$ with which we obtained results for five methods of classification namely optimal rule, full multinomial rule,first order bahadur procedure,second order bahadur procedure and procedure and procedure based on the linear discriminant function;

Table 4.1

$n_1 = n_2 = 30$

$P(MC) = 0.4320$

OPTIMAL		FULL		FIRST		SECOND		LDF	
RESUB	JACK	RESUB	JACK	RESUB	JACK	RESUB	JACK	RESUB	JACK
0.3444	0.3440	0.3333	0.2846	0.3333	0.3125	0.3333	0.2846	0.3333	0.3154
0.4666	0.4626	0.4000	0.3974	0.4500	0.4545	0.4000	0.3974	0.4000	0.4482
0.3755	0.3719	0.3500	0.3463	0.3500	0.3538	0.3500	0.3463	0.3500	0.3470
0.4500	0.4460	0.4167	0.4080	0.4500	0.4386	0.4167	0.4080	0.4500	0.4407
0.3534	0.3530	0.3500	0.3465	0.3667	0.3622	0.3500	0.3463	3500	0.3516
0.3833	0.3839	0.3833	0.3785	0.3833	0.3828	0.3833	0.3807	0.3833	0.3817
0.3567	0.3572	0.3167	0.3142	0.3500	0.3407	0.3167	0.3142	0.3500	0.3481
0.3803	0.3800	0.3667	0.3640	0.3667	0.3682	0.3667	0.3640	0.3667	0.3661
0.3833	0.3841	0.3500	0.3470	0.3833	0.3798	0.3667	0.3555	0.3833	0.3809
0.4167	0.4176	0.3500	0.3484	0.4167	0.4160	0.3833	0.3790	0.4167	0.3821
0.3910	0.3900	3617	0.3535	0.3850	0.3809	0.3667	0.3576	0.3783	0.3762

Table 4.2

$_1 = \quad _2 = 900$

$(\quad) = 0.4320$

OPTIMAL		FULL		FIRST		SECOND		LDF	
RESUB	JACK	RESUB	JACK	RESUB	JACK	RESUB	JACK	RESUB	JACK
0.4445	0.4445	0.4456	0.4455	0.4456	0.4455	0.4455	0.4455	0.4472	0.4463
0.4152	0.4152	0.4139	0.4138	0.4167	0.4166	0.4139	0.4138	0.4167	0.4166
0.4194	0.4194	0.4089	0.4088	0.4089	0.4088	0.4089	0.4088	0.4089	0.4088
0.4320	0.4320	0.4250	0.4249	0.4350	0.4349	0.4250	0.4349	0.4350	0.4349
0.4173	0.4173	0.4128	0.4127	0.4128	0.4127	0.4128	0.4127	0.4128	0.4127
0.4108	0.4108	0.4111	0.4111	0.4111	0.4111	0.4111	0.4111	0.4111	0.4111
0.4133	0.4133	0.4050	0.4049	0.4106	0.4105	0.4106	0.4105	0.4106	0.4105
0.4254	0.4250	0.4244	0.4244	0.4261	0.4304	0.4261	0.4261	0.4350	0.4349
0.4243	0.4241	0.4122	0.4122	0.4122	0.4122	0.4122	0.4122	0.4122	0.4122
0.4482	0.4171	0.4233	0.4233	0.4267	0.4266	0.4239	0.4238	0.4267	0.4266
0.4250	0.4219	0.4182	0.4182	0.4205	0.4209	0.4190	0.4199	0.4216	0.4214

CONCLUSION

From the results obtained from our experiment, we observed that the resubstitution method performed better than the Jackknife method in estimating the exact probability of mis classification. This is not surprising because in classification using binary variables, we do not classify observations parse, but observations with response patterns 000, 010, 001, and so on. Leaving out one observation from a response pattern will not make much difference in the error rates. More often than not also, one encounters response patterns with zero observations which make it impossible to apply the Jackknife method.

REFERENCES

Bartlett,M.S.(1951) " An inverse matrix adjustment arising in discriminant analysis" Annals of Mathematical Statistics, 22 page 107-111.

Cochran, W.G., Hopkins, C.E. (1961) "Some classification problems with multivariate qualitative data" Biometrics, 17, 10 - 32 .

Efron , B. (1975) "The Efficiency of logistic Regression compared to Normal Discriminant Analysis", Journal of America Statistical Association, 70, 892 – 898.

Goldstan ,M., and Dillion, W.R. (1978), "Discrete Discriminant Analysis", John Wiley and Son's, INC new York

Hills, M. (1967) "Discrimination and allocation with discrete data" J.Roy.Stat.Soc. C 16,237-250.

Krzanowski, W.J. (1993) "Principles of Multivariate Analysis,", Oxford University Press Inc. new York.

Lachenbruch, P.A. and Mickey, R.R., (1968) "Estimation of Error rates in Discriminant Analysis"., Technometrics 10, 1-11.

Onyeagu, S.I and Adegboye, O.S. (1996) "Some methods of Estimating the Probability of misclassification in Discriminant Analysis" Journal of the mathematical Association of Nigeria ABACUS vol. 24 no 2 page 104-112.

Onyeagu, S.I.(2003) "A first course in Multivariate Statistics," Mega concept, Awka Nigeria.

Onyeagu S.I. (1997) "Derivation of the optimal classification rule for discrete variables" Unpublished Ph.D dissertation submitted to the University of Ilorin. Nigeria.