Journal of Economics
and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

comparison Bennett's inequality and regression in determining the optimum sample size for estimating the Net Reclassification Index (NRI) using simulation

Researcher: Balasim Saadoun Jasim
Diyala education directorate
Ministry of Education

Prof. Dr. Dejela I. MAHDI
College of Administration and Economics
University of Baghdad

Received: 21/1/2020 Accepted: 21/6/2020 Published: June/ 2021

Abstract :

Researchers have increased interest in recent years in determining the optimum sample size to obtain sufficient accuracy and estimation and to obtain high-precision parameters in order to evaluate a large number of tests in the field of diagnosis at the same time. In this research, two methods were used to determine the optimum sample size to estimate the parameters of high-dimensional data. These methods are the Bennett inequality method and the regression method. The nonlinear logistic regression model is estimated by the size of each sampling method in high-dimensional data using artificial intelligence, which is the method of artificial neural network (ANN) as it gives a high-precision estimate commensurate with the data type and type of medical study. The probabilistic values obtained from the artificial neural network are used to calculate the net reclassification index (NRI). A program was written for this purpose using the statistical programming language (R), where the mean maximum absolute error criterion (MME) of the net reclassification network index (NRI) was used to compare the methods of specifying the sample size and the presence of the number of different default parameters in light of the value of a specific error margin (ε). To verify the performance of the methods using the comparison criteria above were the most important conclusions were that the Bennett inequality method is the best in determining the optimum sample size according to the number of default parameters and the error margin value.

Keywords: optimum sample size, Bennett‘s inequality, artificial neural network, regression method, net reclassification index, mean maximum absolute error.

1-Introduction

Diagnostic or screening tests are used to detect the patient's illness in medical fields. The accuracy of these tests can be assessed through all types of traditional statistical methods such as Sensitivity and Specificity. In recent studies, it has become more and more necessary to evaluate accuracy when adding new information such as the new biomarker (new biomarkers) or the structure of the new paradigm that may be added to the basic diagnostic method.

In order to study the performance of diagnostic accuracy to obtain an appropriate data set with reasonable sample size. We may not obtain a sample size with statistical efficiency to achieve significant results with high accuracy. On the other hand, it is not possible to conduct a study with a very large sample size leading to a loss of accuracy in the desired results And effort, time and material cost. There are several methods for calculating the optimum sample size in the areas of social, economic and medical sciences, especially in the field of diagnostic medicine.

Therefore, the question of estimating the sample size, in this case, has become more difficult and complex. Therefore it is necessary to use methods of calculating the sample size that is commensurate with the high-dimensional data and also commensurate with the criteria of accuracy, as was suggested (Pencina et al. in the year 2008) the net reclassification index (Net reclassification index) (NRI), as this indicator is used to improve diagnostic accuracy when adding new vital signs to basic vital signs [Pencina, D'Agostino, Vasan: 2008] and obtaining a new model that includes basic variables in addition to the new signs, and let new vital signs be on, For example, genes belonging to a specific protein of breast cancer patients These signs have been added to the basic vital signs, for example, age, gender, heart rate, body temperature, blood pressure, respiratory rate, etc. To improve the accuracy of the diagnosis for people with this disease and to judge about all this process, we use an accuracy criterion which is the net reclassification index (NRI), so it must be determined. Appropriate sample size to measure diagnostic accuracy using this index.

This research aims to find the best sample size to obtain the best estimate of parameters in the presence of high-dimensional data (HDD) based on the best reclassification network indicator (NRI) as a statistical measure to obtain accuracy in diagnosis and thus obtain the most accurate diagnosis of new vital indicators added to basic vital indicators In order to obtain an accurate diagnosis or medical examination.

2. Sample Size Determination Methods

In this research, we will use two methods to determine the optimum sample size in the presence of high-dimensional data, which are Bennett's inequality method and the regression method and by using a binary non-linear logistic regression model to estimate the Net Reclassification Index (NRI). These methods can be listed as follows.

1.2 - Inequalities Method

Sometimes the application of normal approximation using the central limit theory is undesirable, and the symmetric unimodal distribution may not be appropriate to the shape of the distribution in order to describe the estimated parameters in finite samples, especially when the true parameters are within the naturally bounded.

2.1.1 – Bennett's inequality (Bennet)

Proposed Bennett's in the year (1962) Bennett's inequality, and I used this inequality for the jth of the estimators where the following formula was obtained: -

$$Pr(|\hat{\theta}_j - \theta_j| > \epsilon) \leq 2exp\left(-\frac{n\epsilon}{W}\left[\left(1+\frac{v_j}{w\epsilon}\right)\log\left(1+\frac{w\epsilon}{v_j}\right)-1\right]\right)$$

To achieve an error probability bound (α), we may bound each estimation error probability with(α/p) using the Bonferroni correction.

Note that the right side of the inequality is a decreasing function of the variance of v_j . [Bennett: 1962]

As (v_j) represents the maximum values of variance vector for each estimator (j) to achieve the limits of error probability for the level of significance (α) from the above probability formula, we obtain a formula for calculating the sample size as in the following formula: -

$$n^* = \left\{\frac{\epsilon}{W}\left[\left(1+\frac{v}{W\epsilon}\right)\log\left(1+\frac{W\epsilon}{v}\right)-1\right]\right\}^{-1}\left(\log p + \log\frac{2}{\alpha}\right) \dots .(1)$$

Assume ϵ> 0, and a very small value is close to zero. [Pencina, D'Agostino, Vasan: 2008]

2.2 - Regression Method

(Jiang and Li) proposed (2017) a regression method for calculating the sample size in the case of high-dimensional data in order to determine the biomarkers, supposing that we have an experimental data set with a size (n_o) (the default size is from previous experience). Practically, this training data set can be obtained either from previous studies or by conducting a pilot study. [Jiang, Li: 2008]

This method can be illustrated with the following steps: -

(1) - Generate k random resamples from an experimental data set. For repetition k = 1, ……, K . We generate a random number N_{jk} through uniform distribution of the over $[n_o/2\,,n_o]$ And obtain a sample of size N_{jk} .We refer to the parameters for θ_j that are estimated based on kth from a sample that represents $\hat{\theta}_{jk}$.

And using averages that

$$\bar{\theta}_j = \frac{\sum_{k=1}^{k}\sqrt{N_{jk}}\hat{\theta}_{jk}}{\sum_{k=1}^{k}\sqrt{N_{jk}}}$$

Then the error is calculated from the estimation process for each sample K.

$\epsilon_{jk} = \left|\widehat{\theta}_{nj} - \overline{\theta}_j\right|$

(2) - Therefore, when estimation error is taken, we obtain:

$\epsilon_j = \mathrm{E}\left|\widehat{\theta}_{nj} - \overline{\theta}_j\right| \approx \mathrm{C}_j / \sqrt{N_{jk}}$

Dividing the constant vector is approximately equal to the root of the given sample size.

$Var(\epsilon_j) \propto 1/\sqrt{N_{jk}}$

Then we get K of the arranged pairs for the sample size and the estimation error from the first step, Where the sample size is N_{jk} (independent variable) and the estimation error ϵ_{jk} as a response (dependent variable), thus we produce a regression model as in the following formula: - [Jiang, Li: 2008]

$$\epsilon_{jk} = \frac{\beta_j}{\sqrt{N_{jk}}} + \frac{e_{jk}}{\sqrt{N_{jk}}} \quad , k = 1, 2, \dots \dots ., k \quad \cdots \cdots \cdots (2)$$

. Where : (β_j) unknown parameters

e_{jk}: random error with a mean of zero.

We estimate the vector parameters (β_j) using the (Weighted least Squares Estimation) (WLS) method, we obtain:-

$$\widehat{\beta}_j = \frac{1}{K}\sum_{k=1}^{k} \sqrt{N_{jk}\epsilon_{jk}}$$

Depending on the regression model in the second step, we can calculate the sample size Nj for any error value (ϵ), and assuming that Z j is the significance levels$1 - \alpha/p$ of the following amount:

$$\sqrt{N_{jk}\epsilon_{jk}} - \widehat{\beta}_j \quad , k = 1, 2, \dots \dots ., k$$

By employing the significance levels, sample sizes and estimated parameters previously obtained in the regression model to obtain error values, as follows:

$$\epsilon = \frac{\widehat{\beta}_j}{\sqrt{N_j}} + \frac{Z^j}{\sqrt{N_j}} \quad , k = 1, 2, \dots \dots ., k \qquad \cdots \cdots \cdots (3)$$

The values of (Nj) are as follows:

$$N_j = \left(\frac{\widehat{\beta}_j + Z^j}{\epsilon}\right)^2$$

From the values of(Nj), we withdraw the sample size, which represents the largest value of (Nj) , as follows:-

$$n = max\{N_j : 1 \leq j \leq p\} \qquad \cdots \cdots \cdots (4)$$

3. Binary logistic regression model

Logistic regression is one of the elements of a set of models called the general linear models' group. The logistic regression model describes the relationship between the response variable, which is of the intermittent type, and the explanatory variables are of the intermittent or continuous type or mixing between them. It is also considered an important method in data classification and allocation to specific groups, and It has the relationship between the response variable (y) and the explanatory variables (Xi), whether its type is quantitative or qualitative non-linear, where the response variable (y) is binary response Bernoulli distribution assuming one of the two values (1,0), and that the probability of occurrence Price (π_i) and the probability of non-response $(1-\pi_i)$. As it is considered a special case of GLRM models, the probability density function can be written as follows: - [Scott, 2002]

$$p(Y=y_i)=\pi_i^{y_i}(1-\pi_i)^{1-y_i} \qquad \cdots\cdots\cdots(5)$$

Where

yi = 0,1

yi: binary response variable.

πi: the probability of a response when yi = 1.

Expecting a response variable represents the probability of a response.

$$E(y_i)=p(Y=1)\;=\pi_i \qquad \cdots\cdots\cdots(6)$$

The variance of the response variable, according to the Bernoulli distribution.

$$V(y_i)=\pi_i(1-\pi_i) \qquad \cdots\cdots\cdots(7)$$

Let $x_1, x_2, \ldots\ldots x_k$Be a set of explanatory variables that n represents the number of observations of these variables that form the matrix X.

$$X=(x_{ij})_{n\times k}$$

Where

i = 1,2, ⋯⋯, n, (n) represents the sample size.

j = 1,2, +, k + 1, (k) is the number of explanatory variables, p = k + 1 represents the number of parameters.

If yi = [y1, y2, ...… yn] is a random sample from the two-response variable and yi ∈ {0,1}.

Thus, the formula for the logistic regression model is as follows:

$$y_i=\pi_i+e_i \qquad \cdots\cdots\cdots(8)$$

Where

πi: represents the logistic regression function (probability of response).

$$\pi_i=\frac{e^{\beta_0+\sum_{j=1}^{p}X_{ij}\beta_j}}{1+e^{\beta_0+\sum_{j=1}^{p}X_{ij}\beta_j}} \qquad \cdots\cdots\cdots(9)$$

Where

β_j: vector of the parameters of degree (1 × p) if i=1,2,...,n , j=1,2,....,p

(Xij) = {xi1, xi2,… xij}: an array of explanatory variables of degree (n x p)

n: represents the number of views.

p: represents the number of explanatory variables.

Whereas, the random error limit (ei) is divided into Bernoulli distribution with a mean of zero, as in the formula: -

$$e_i=y_i-\pi_i \qquad \cdots\cdots\cdots(10)$$

Taking the expected for two sides of equation (10), we obtain:-

$$\mathrm{E}(\mathrm{e_i}) = \pi_\mathrm{i} - \pi_\mathrm{i} = 0 \qquad \cdots\cdots\cdots(11)$$

The variance of the random error limit is equal to the variance of the dual response variable.

$$\mathrm{V}(\mathrm{e_i}) = \pi_\mathrm{i}(1 - \pi_\mathrm{i}) \qquad \cdots\cdots\cdots\cdots(12)$$

Therefore, the random error limit has a mean of zero and a variance of πi (1-πi). It is noted that the variance of the error limit depends on the values of the response probability (πi). [Shen and Gao, 2008: 4]

3.1 - Linear transformation to the binary logistic regression function

Because of the negative bends in the logistic regression parameters, which affect these properties of the parameters and the predictable response values, many statisticians resort to linear transformation by using the Logit function (Logit) to remove the curvature of its parameters where the results are The test is misleading because the parameters are not distributed naturally and are biased. Their variations are not as minimal as possible. In (1944), the researcher (Berkson) found a logarithmic relationship in order to convert the relationship between the explanatory variables and the probability of the response (π_i) to a linear relationship by converting the probability pr (Y = 1│X) to a function whose duration is from (-∞, + ∞) This function is called the odds ratio which is given by the following formula: -

$$\frac{\mathrm{pr}(\mathrm{Y}=1|\mathrm{X})}{\mathrm{pr}(\mathrm{Y}=0|\mathrm{X})} = \frac{\mathrm{pr}(\mathrm{Y}=1|\mathrm{X})}{1-\mathrm{pr}(\mathrm{Y}=1|\mathrm{X})} \qquad \cdots\ \cdots\ \cdots(13)$$

Until we get a function with a range of (-∞, + ∞), we take the natural logarithm of the odds ratios (Odde), and it is as follows:-

$$\mathrm{logit}\ (\pi_\mathrm{i}) = \mathrm{Ln}\left(\frac{\pi_i}{1-\pi_i}\right) \qquad \cdots\cdots\cdots(14)$$

$$\mathrm{Ln}\left(\frac{\pi_i}{1-\pi_i}\right) = \mathrm{Ln}\left(e^{\beta_0+\sum_{j=1}^{p} X_{ij}\beta_j}\right) \qquad \cdots\cdots\cdots(15)$$

$$\mathrm{Ln}\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} X_{ij}\,\beta_j \qquad \cdots\cdots\cdots(16)$$

The form can be rewritten as in the following form:-

$$\mathrm{Ln}\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_\mathrm{u}\mathrm{U} + \beta_1\mathrm{x}_{\mathrm{i}1} + \beta_2\mathrm{x}_{\mathrm{i}2} + \cdots + \beta_\mathrm{j}\mathrm{x}_{\mathrm{ij}}\,, \qquad \cdots\ (17)$$

Where

I=1,2,.....n , j=1,2,.....,p

U: is a baseline variable .

Xi1, Xi2,…, X (ij): represent the observations of the explanatory variables.

$\beta_0, \beta_1, \beta_2, \ldots, \beta_j$: logistic regression model parameters. This means that the binary logistic regression model became linear using the Login function [Kleinbaum, Klein, 2002]

4. Artificial Neural Network (ANN)

Artificial neural networks (ANN) are networks of artificial neurons connected to act as simple processing elements for a specific task. There are different types of networks, including multi-layer networks (Multi-Layer NetWork) that have been widely used to solve regression problems, as the multi-layer neural network model contains an input layer (source nodes), or one or more hidden layers (arithmetic nodes) and the output layer) Account contract). Whereas the input layer has the function of receiving the input signals, the hidden layers are responsible for processing and disseminating the received (input) signals and their output. [Haykin, 2008]

In order to obtain efficient and highly accurate prediction capabilities, we will use the artificial neural network called the hidden monolayer neural network while bypassing the communication layer. The aim is to obtain estimates for the binary logistic regression model and employ these values in the net reclassification index (NRI) to obtain the estimated values of the indicator and then determine the best method for determining the optimum sample size.

Some multi-layer neural networks may have special connections called (Skip Layer), where some of the input signals directly connect to the output layer, i.e., the implementation of the hidden layer. [Pencina, D'Agostino, Steyerberg: 2011].

$$y_i = \phi_i\left(b_i + \sum_{l=1}^{m} W_{kl}X_l + \sum_{i=1}^{n} W_{il}.\,\phi_k\left(\sum_{j=1}^{m} W_{kj}X_j + b_j\right)\right) \cdots\cdots (18)$$

Equation (9) shows the artificial neural network with skipping the communication layer where the result of the input node (k) is handled by the output node (l) and (logistic activation function) will be used, which takes the following formula: -

$$\phi(z) = \frac{1}{1+e^{-z}} \quad \cdots\cdots\cdots\cdots (19)$$

Depending on the activation function, the neural network outputs will be binary values confined between (0.1). Also, this network represents the direct contact between the input node and the output node. This network is called the hidden monolayer neural network, with the communication layer being skipped. [Habibnia, Maasoumi, 2008]

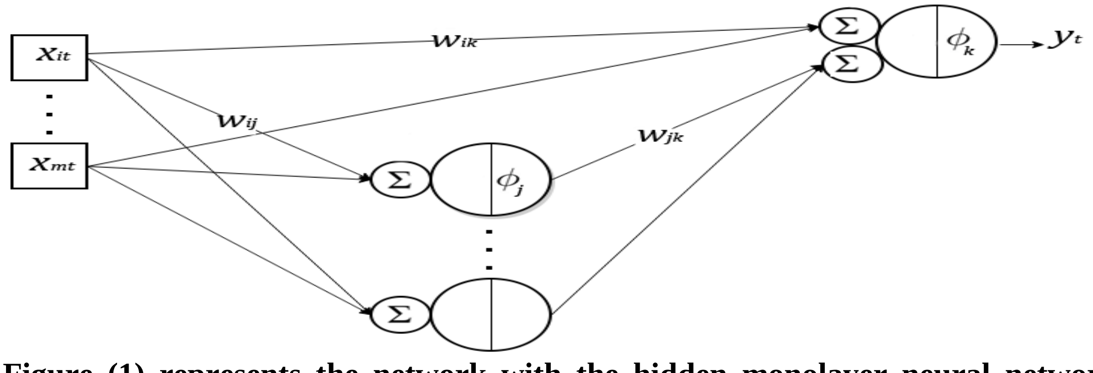


Figure (1) represents the network with the hidden monolayer neural network with the connection layer being bypassed

5. Measurements of the accuracy of diagnosis and classification

We will use one of these important measures, the Net reclassification index scale, and then obtain the best sample size that gives the most accurate diagnosis of the new biomarkers added to the basic biomarkers for diagnosis or a medical examination. [Pencina, D'Agostino, Vasan: 2008]

6. Net reclassification index

It was suggested by (Pencina and others) in (2008), and it is considered one of the common statistical measures to evaluate the new vital indicator added to the basic model in case of multiple categories.

We use the net reclassification index in the classification in the case of multiple categories, and the formula for the net reclassification index (NRI) is as follows: -

$$S_J = \sum_{m=1}^{M} \omega_m \, P\{p_m(\mathcal{M}_2) = max\, p(\mathcal{M}_2), p_m(\mathcal{M}_1) \neq maxp(\mathcal{M}_1)|Y = m\} \cdots (20)$$

Suppose we have (p) biomarkers, and we want an NRI for these scores for a sample view. We refer to the net reclassification index (NRI) with the symbol (Sj) where (j = 1, 2, ..., p). We refer to his estimate through this sample with the symbol $(\widehat{S_J})$ Moreover, the probabilities of this indicator are estimated in two categories based on the two-response logistic regression model. In the case of multiple categories, its probabilities were estimated by relying on the multi-response logistic regression model as it is considered non-estimated Unbiased, and the estimated NRI is as follows: -

$$\widehat{S_J} = \sum_{m=1}^{M} \frac{\omega_m}{n_m} \sum_{i=1}^{n} I\{\hat{\pi}_{mi}(\mathcal{M}_2) = max\hat{\pi}_i(\mathcal{M}_2), \hat{\pi}_{mi}(\mathcal{M}_1) \neq max\hat{\pi}_i(\mathcal{M}_1),\ Y_i = m\} \cdots\cdots\cdots (21)$$

Where

$\hat{\pi}_{mi}(\mathcal{M}_{2j}) = max\hat{\pi}_i(\mathcal{M}_{2j})$**: represents the probabilities of the new model.**

$\hat{\pi}_{mi}(\mathcal{M}_1) = max\hat{\pi}_i(\mathcal{M}_1)$ **: represents the probabilities of the old (basic) model.**

$\hat{\pi}_i(\mathcal{M}_j) = \left(\hat{\pi}_{1i}(\mathcal{M}_j), \hat{\pi}_{2i}(\mathcal{M}_j), \dots\dots\dots, \hat{\pi}_{Mi}(\mathcal{M}_j)\right)$ **: represents the estimated probabilities.**

Using the central limit theory (CLT), we can show $\sqrt{n}(\hat{S} - S) \overset{d}{\to} N(0, \sigma_{\hat{S}}^2)$. The variance of the scale (($\widehat{S_J}$)) can be obtained from the following relationship: -

$$\sigma_{\hat{S}}^2 = \sum_{m=1}^{M} \frac{w_m^2}{p_m} d_m - \sum_{i=1}^{M}\sum_{j=1}^{M} w_i w_j d_i d_j \qquad \cdots\cdots (22)$$

Where

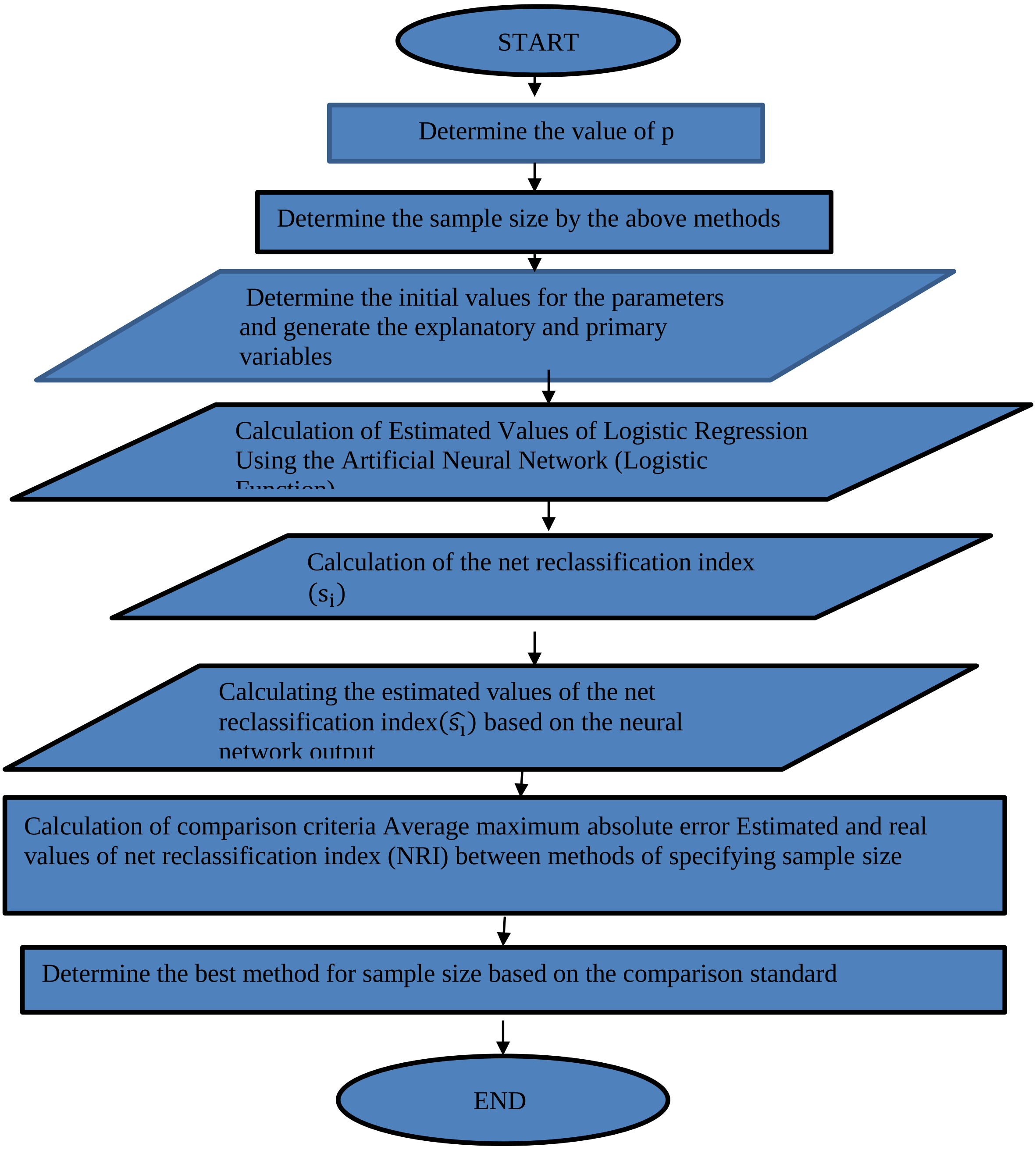
$d_m = P(\rho_m(\mathcal{M}_1) \neq maxp(\mathcal{M}_1), \rho_m(\mathcal{M}_2) = maxp(\mathcal{M}_2)|Y = m)$

is considered the best rating accuracy measure with the largest value of the net reclassification index (NRI). [Li, Fine, : 2008]

7. Simulation

It is a method that includes the use of a theoretical mathematical model and its analogy to the real model that represents the problem studied or as defined by (Naylor) It is a method used to control the experience to be studied and know its characteristics and properties using mathematical and logical susceptibility available in the electronic calculator or as others have known a method by which an alternative model is found Similar to the real model without trying to get the real model itself. [Al-Batal: 2008]

Figure (2) an algorithm showing the simulation mechanism to compare methods of determining the sample size


START
Determine the value of p
Determine the sample size by the above methods
Determine the initial values for the parameters and generate the explanatory and primary variables
Calculation of Estimated Values of Logistic Regression Using the Artificial Neural Network (Logistic Function)
Calculation of the net reclassification index (s_i)
Calculating the estimated values of the net reclassification index$(\hat{s}_i)$ based on the neural network output
Calculation of comparison criteria Average maximum absolute error Estimated and real values of net reclassification index (NRI) between methods of specifying sample size
Determine the best method for sample size based on the comparison standard
END

((p): represents the number of features, i.e., the number of explanatory variables (X's))

7.1 Stages of building simulations

To analyze data using simulation experiments, it is necessary to define the stages of building simulation, which are as follows: -

7.1.1 Generate explanatory variables

First Case: -

Explanatory variables are generated according to the multivariate normal distribution according to the following formula: -

$\mathbf{X_1, X_2, \dots, X_p \sim MN(0, \Sigma)}$

As that

$\Sigma = (\sigma_{ij})_{p\times p}\ , \sigma_{ii} = 1\ \ , 1 \le i, j \le p\ \ , \sigma_{i-1,i} = \sigma_{i,i-1} = 0.5\ \ , 2 \le i \le p\ \ , \sigma_{ij} = 0$

As for the main variable, it is generated from the standard normal distribution, as follows: -

$\mathbf{U \sim N(0, 1)}$

Second Case:-

The explanatory and baseline variables U are generated according to the standard multivariate normal distribution and according to the following formula: -

$\mathbf{U, X_1, X_2, \dots, X_p \sim MN(0, 1)}$

7.1.2 -Defining the initial parameters

The default values for simulation experiments are described according to the following table: -

Table (1-4): The default values for the initial parameters for the first case

$(Case\ one)$	= pعدد المعلمات	$(parameter)$ $\beta = (\beta_0, \beta_u, \beta_1, \dots, \beta_p)$
1	$p = 50$	$\beta = (1.5, 2, 0.5)\ , \beta_j = 0\ , j = 4, \dots, 50$
2	$p = 150$	$\beta = (1.5, 2, 0.5)\ , \beta_j = 0\ , j = 4, \dots, 150$
3	$p = 300$	$\beta = (1.5, 2, 0.5)\ , \beta_j = 0\ , j = 4, \dots, 300$

Table (2-4): The default values for the initial parameters for the second case

$(Case\ one)$	= pعدد المعلمات	$(parameter)$ $\beta = (\beta_0, \beta_u, \beta_1, \dots, \beta_p)$
1	$p = 100$	$\beta = (2, 2, 3)\ , \beta_j = 0\ , j = 4, \dots, 100$
2	$p = 400$	$\beta = (2, 2, 3)\ , \beta_j = 0\ , j = 4, \dots, 400$
3	$p = 600$	$\beta = (2, 2, 3)\ , \beta_j = 0\ , j = 4, \dots, 600$

7.1.3 -Response Variable Calculation

The probability response variable is calculated based on the logistic regression model as follows: -

$$\log\frac{p}{1-p}=\beta_0+\beta_u U+\beta_1 X_1+\cdots+\beta_p X_p \quad \cdots\cdots\cdots(23)$$

As that

U: is a baseline variable .

β_u: parameter of the baseline variable.

7.1.4- Generate the Reclassification Network Index

After the probability model has been generated in paragraph (7.1.3), the probabilistic value is used to calculate the reclassification network index, which represents the confusion between two probable logistical models, one of which includes the main variable and the probable response variable symbolized by $\hat{\pi}(\mathcal{M}_1)$ which takes the formula The following: -

$$\hat{\pi}_i(\mathcal{M}_1)=\log\frac{p}{1-p} \quad \cdots\cdots\cdots(24)$$

$$\log\frac{p}{1-p}=\beta_0+\beta_u U \quad \cdots\cdots\cdots(25)$$

The other includes the primary variable in addition to the explanatory variables and the response variable, symbolized by $\hat{\pi}(\mathcal{M}_{2j})$, which takes the following formula: -

$$\hat{\pi}_i(\mathcal{M}_{2j})=\log\frac{p}{1-p} \quad \cdots\cdots\cdots(26)$$

$$\log\frac{p}{1-p}=\beta_0+\beta_u U+\beta_1 X_1+\cdots+\beta_p X_p \quad \cdots\cdots\cdots(27)$$

As that

U: is a baseline variable.

β_u: parameter of the baseline variable.

7.1.5- Discuss simulation experiments

Statistical Programming Language (R) program was used where the average greater absolute standard (MME) of the Net Reclassification Network (NRI) index was adopted, and the results were as follows: -

Table (3-4): Comparison of methods for determining the sample size the first case

NRI	$(P=50)$ $\beta=(1.5,2,0.5\,),\beta_o=0.5,\beta_u=0.5$		
Methods		Bennt. Inq.	regression
$(\varepsilon=0.1)$	n^*	1570	520
	$MAE\ \hat{\varepsilon}$	0.00371892	0.0100436
	(MME)	0.02242589	0.03654729
$(\varepsilon=0.05)$	n^*	6181	2221
	$MAE\ \hat{\varepsilon}$	0.001607057	0.007394658
	(MME)	0.01652275	0.02422569

Table (4-4): Comparison of methods for determining the sample size the first case

NRI	$(P = 150)$ $\beta = (1.5, 2, 0.5)\,, \beta_o = 0.5, \beta_u = 0.5$		
Methods		Bennt. Inq.	regression
$(\varepsilon = 0.1)$	n^*	1797	638
	$MAE\ \hat{\varepsilon}$	0.002937762	0.005637807
	(MME)	0.01585712	0.04388563
$(\varepsilon = 0.05)$	n^*	7075	2190
	$MAE\ \hat{\varepsilon}$	0.00124253	0.002439609
	(MME)	0.0112162	0.03802348

Table (5-4): Comparison of methods for determining the sample size the first case

NRI	$(P = 300)$ $\beta = (1.5, 2, 0.5)\,, \beta_o = 0.5, \beta_u = 0.5$		
Methods		Bennt. Inq.	regression
$(\varepsilon = 0.1)$	n^*	1940	674
	$MAE\ \hat{\varepsilon}$	0.004069711555	0.005864626
	(MME)	0.01760712	0.03585019
$(\varepsilon = 0.05)$	n^*	7638	2188
	$MAE\ \hat{\varepsilon}$	0.0009105066	0.002693432
	(MME)	0.006812624	0.01849395

Table (6-4): Comparison of methods for determining the sample size the second case

NRI	$(P = 100)$ $\beta = (2, 2, 3)\,, \beta_o = 0.5, \beta_u = 0.5$		
Methods		Bennt. Inq.	regression
$(\varepsilon = 0.1)$	n^*	1713	519
	$MAE\ \hat{\varepsilon}$	0.003597973	0.008815351
	(MME)	0.01546763	0.06086168
$(\varepsilon = 0.05)$	n^*	6745	2165
	$MAE\ \hat{\varepsilon}$	0.001754779	0.004324181
	(MME)	0.01323156	0.02997787

Table (7-4): Comparison of methods for determining the sample size the second case

NRI	$(P = 400)$ $\beta = (2, 2, 3)\,, \beta_o = 0.5, \beta_u = 0.5$		
Methods		Bennt. Inq.	regression
$(\varepsilon = 0.1)$	n^*	2000	793
	$MAE\ \hat{\varepsilon}$	0.002442682	0.005820985
	(MME)	0.01793922	0.03590148
$(\varepsilon = 0.05)$	n^*	7872	3174
	$MAE\ \hat{\varepsilon}$	0.001615918	0.001876937
	(MME)	0.008292873	0.009594903

Table (8-4): Comparison of methods for determining the sample size the second case

NRI	$(P = 600)$ $\beta = (2, 2, 3\,), \beta_o = 0.5, \beta_u = 0.5$		
Methods		**Bennt. Inq.**	**regression**
$(\varepsilon = 0.1)$	n^*	**2083**	**777**
	$MAE\ \hat{\varepsilon}$	**0.002515294**	**0.009956108**
	(MME)	**0.0190346**	**0.05703476**
$(\varepsilon = 0.05)$	n^*	**8202**	**2564**
	$MAE\ \hat{\varepsilon}$	**0.002148653**	**0.00388981**
	(MME)	**0.008870116**	**0.043663**

Analysis of the results

Simulation results for the first case showed that table number (3-3), (4-3), (5-3), resp -ectively, when the number of parameters (p = 50) (p = 150) (p = 300) and the parame- ter values ($\beta = (1.5, 2, 0.5\,), \beta_o = 0.5,\ \beta_u = 0.5$) The best way to determine the sample size when the margin of error value (ε = 0.1,0.05) is the regression method since it has the highest mean largest absolute error (MME).

Simulation results for the second case showed that table number (6-3), (7-3), (8-3), respectively, when the number of parameters (p = 100) (p = 400) (p = 600) and the parameter values ($\beta = (2, 2, 3\,), \beta_o = 0.5,\ \beta_u = 0.5$) The best way to determine the sample size when the margin of error value (ε = 0.1,0.05) is the regression method since it has the highest average largest absolute error (MME).

8 . Conclusions and recommendations

8.1 - Conclusions

Through simulation experiments and the results presented, the researcher concluded the following: -

1- Through the experimental side, the results proved that the best way to determine the sample size in both cases is the (Regression) method and at the margin of error value (ε = 0.1,0.05)

2- There is an inverse relationship between the sample size and the margin of error, as we note that the smaller the margin of error, the greater the sample size and vice versa.

3- The best estimate of hypothetical parameters for high-dimensional data was obtained by using the Regression method because it has the highest mean maximum and greatest absolute error (MME).

8.2 - Recommendations

In light of the conclusions that we reached through the research, the following recommendations can be included:

1- We recommend using the Regression method to determine the sample size in the case of high-dimensional data and for studies, especially in the medical and diagnostic fields.

2- We recommend using the Net Reclassification Index (NRI) for medical studies of diagnostic medicine.

3- Using different methods other than artificial neural networks to estimate the dual logistic regression model and thus estimate the net reclassification index (NRI).

9. Sources

[1]-Al-Battal, Ahmed Hussein, Ahmed, Issam Kamel, Khudair, Al-Bara 'Abdul-Wahab (2008), "Using Simulation in Teaching Simple Linear Regression", Anbar University Journal of Pure Sciences, Third Issue, Volume Two.

[2]- Bennett, G., (1962). " Probability inequalities for the sum of independent random variables" . Journal of the American Statistical Association. Vol. 57, No. 297, pp. 37–39.

[3]-Herbert, J., & Forrest, D. (1984)." Linear Probability, Logit, and Probit Models". Sage Publications, Inc.

[4]- Habibnia, A., & Maasoumi, E. (2019) ." Forecasting in Big Data Environments: an Adaptable and Automated Shrinkage Estimation of Neural Networks (AAShNet)".pp.1-26

[5]- Jiang B.,& Li J., (2017) ."Sample size determination for high dimensional parameter estimation with application to biomarker identification". Computational Statistics and Data Analysis, 118, 1–12.

[6]- Kleinbaum, D. G., & Klein, M. (2002)." Logistic Regression A Self-Learning Text". 3d ,Springer.

[7]- Li, J., and Fine, J., (2008). "Roc analysis with multiple classes and multiple tests: methodology and its application in microarray studies". Biostatistics 9, 566–576.

[8]- Pencina, M., D'Agostino, R., and Vasan, R., (2008)." Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond ". Statistics in Medicine 27,pp- 157–172

[9]- Pencina, M., D'Agostino, R., Steyerberg, E., (2011)." Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers". Statistics in Medicine 30,pp- 11–21 .

[10]- Riple, B. D. (1996) ." Pattern Recognition and Neural Network ". Cambride University .

[11]- Scott, M. (2002). "Applied Logistic Regression Analysis". sage.

[12]- Shen, J., & Gao, S. (2008) ." A Solution to Separation and Multicollinearity in Multiple Logistic Regression". Journal of Data Science : JDS,

[13]- Ying, Y. (2014)." McDiarmids inequalities of Bernstein and Bennett forms ". pp.1-6.

مقارنة متباينة بينيت و الانحدار في تحديد حجم العينة الامثل لتقدير مؤشر اعادة التصنيف الصافي (NRI) باستعمال المحاكاة [*]

الباحث: بلاسم سعدون جاسم
مديرية تربية ديالى وزارة التربية
blasm5517@gmail.com

أ.د. دجلة إبراهيم مهدي
كلية الإدارة والاقتصاد جامعة بغداد
dr.dejela.mahdi@gmail.com

Received: 21/1/2020 Accepted: 21/6/2020 Published: June/ 2021

مستخلص البحث

زاد اهتمام الباحثين في السنوات الاخيرة بتحديد حجم العينة الامثل للحصول على دقة وتقدير كافيين وللحصول على معالم عالية الدقة وذلك لتقييم عدد كبير من الاختبارات في مجال التشخيص في ان واحد.

تم في هذا البحث استعمال طريقتان لتحديد حجم العينة الامثل لتقدير معالم البيانات ذات الابعاد العالية . وهذه الطرائق هي طريقة متباينة بينت وطريقة الانحدار . يتم تقدير انموذج الانحدار اللوجستي الثنائي اللاخطي بحجم عينة كل طريقة في حالة بيانات عالية الابعاد باستعمال الذكاء الاصطناعي وهي طريقة الشبكة العصبية الاصطناعية (ANN) كونها تعطي تقدير عالي الدقة بما يتناسب مع نوع البيانات ونوع الدراسة الطبية . يتم توظيف القيم الاحتمالية التي تم الحصول عليها من الشبكة العصبية الاصطناعية في حساب مؤشر اعادة التصنيف الصافي (NRI) , تم كتابة برنامج لهذا الغرض باستعمال لغة البرمجة الاحصائية (R) حيث تم الاعتماد على معيار متوسط اكبر خطا مطلق (MME) لمؤشر شبكة اعادة التصنيف الصافي (NRI) للمقارنة بين طرائق تحديد حجم العينة وبوجود عدد المعلمات الافتراضية مختلفة في ظل قيمة هامش خطا معين (ε) . للتحقق من اداء الطرائق باستعمال معايير المقارنة اعلاه حيث كانت اهم الاستنتاجات هي ان طريقة متباينة بينيت هي الافضل في تحديد حجم العينة الامثل باختلاف عدد المعلمات الافتراضية وقيمة هامش الخطأ .

المصطلحات الرئيسة للبحث :حجم العينة الامثل , متباينة بينيت , الشبكة العصبية الاصطناعية , طريقة الانحدار , مؤشر اعادة التصنيف الصافي , متوسط اكبر خطا مطلق

[*] بحث مستل من رسالة ماجستير.