AN ALTERNATIVE METHOD FOR DETECTION OF INFLUENTIAL OBSERVATIONS IN LOGISTIC REGRESSION MODEL BASED ON BINARY PARTICLE SWARM OPTIMIZATION

**A THESIS SUBMITTED TO INSTITUTE OF GRADUATE STUDIES
OF
NEAR EAST UNIVERSITY**

**By
KARWAN MOHAMMED EISIF**

**In Partial Fulfillment of the Requirement for the Degree of Master of Science
in
Mathematics**

NICOSIA 2021

KARWAN MOHAMMED EISIF
EISIF

AN ALTERNATIVE METHOD FOR DETECTION OF INFLUENTIAL
OBSERVATIONS IN LOGISTIC REGRESSION BASED ON BINARY
PARTICL SWARM OPTIMIZATION

NEU
2021

AN ALTERNATIVE METHOD FOR DETECTION OF INFLUENTIAL OBSERVATIONS IN LOGISTIC REGRESSION MODEL BASED ON BINARY PARTICLE SWARM OPTIMIZATION

**A THESIS SUBMITTED TO INSTITUTE OF
GRADUATE STUDIES
OF
NEAR EAST UNIVERSITY**

**By
KARWAN MOHAMMED EISIF**

**In Partial Fulfillment of the Requirement for
the Degree of Master of Science
in
Mathematics**

NICOSIA 2021

Karwan Mohammed EISIF: AN ALTERNATIVE METHOD FOR DETECTION OF INFLUENTIAL OBSERVATIONS IN LOGISTIC REGRESSION MODEL BASED ON BINARY PARTICLE SWARM OPTIMIZATION

Approval of Director of Institute of

Graduate Studies

Prof. Dr. K. Hüsnü Can BAŞER

We certify, this thesis is satisfactory for the award of the degree of Masters of Science in Mathematics

Examining Committee in Charge:

Prof.Dr. Evren Hincal	Committee Chairman,Department of Mathematics, NEU
Assist.Prof.Dr. Nuriye Sancar	Supervisor,Department of Mathematics, NEU
Assoc.Prof.Dr. Murat Tezer	Department of Primary Mathematics Teaching, NEU

I declare that all information in this document had been obtained and presented with the following academic rules and ethical conduct. As required by these rules and behavior, I have fully cited and referenced all materials and results that are not original to this work.

Name, Last name: Karwan Mohammed Eisif Eisif

Signature:

Date: 30 – 8 - 2021

ACKNOWLEDGEMENTS

I would like to express my gratitude and appreciation to my supervisor, Assist. Prof. Dr. Nuriye Sancar, for her support, encouragement, valuable suggestions throughout the preparation of this work.

I express gratitude to my family, especially my mother and wife, for encouragement, help, love, and indulgence during this thesis's preparation.

I want to extend my thanks to all my teachers who taught me during the master's degree courses and gave me vital information. As well as this, I would like to extend my gratitude to the dean and staff members who help me succeed in this thesis, especially the mathematic department's staff.

To my parents…

ABSTRACT

Binary logistic regression is a statistical model for predicting the probability of an occurrence, and is a good way to see how independent factors are related to a binary response variable. This type of model is commonly used to simulate a variety of real-world problems. The theoretical basis for understanding the logistic regression model and the mathematical equations associated with it are reviewed in this research. Correct determination of influential observations must be important part in process of modelling logistic regression since the unsuccess to detect influential observations cause misleading infrerences from the model. Existing techniques for the determination of influential data points in the literature are founded on the leave-one-out techniques. But, the findings from these single-observation based techniques are often specious because of swamping and masking problems in the existence of multiple influential data points in the dataset.

Thus, in this research the identification of the optimal group of influential observation problems has been regarded as a combinatorial optimization issue and the Binary Particle Swarm Optimization (BPSO) method has been utilized as a novel simultaneous strategy for identifying the optimal group of influential observations in the logistic model. The performance of the suggested BPSO-based method has been checked against standard diagnostic approaches by simulated studies in accordance with several evaluation criteria.

Keywords**:** Logistic regression model; Influential observations; Binary Particle Swarm Optimization; Likelihood Displacement; Masking Issue; Swamping Issue

ÖZET

Lojistik regresyon, bir olayın meydana gelme olasılığını tahmin etmek için sıklıkla kullanılan istatistiksel bir modeldir ve bağımsız faktörlerin bir ikili yanıt değişkeniyle nasıl ilişkili olduğunu görmenin iyi bir yoludur. Bu tür bir model, çeşitli gerçek dünya problemlerini simüle etmek için yaygın olarak kullanılır. Bu araştırmada lojistik regresyon modelini anlamak için teorik temel ve bununla ilişkili matematiksel denklemler gözden geçirilmiştir. Etkili gözlemlerin doğru belirlenmesi, lojistik regresyonu modelleme sürecinde önemli bir rol oynamalıdır, çünkü etkili gözlemlerin tespit edilememesi modelden yanıltıcı çıkarımlara neden olur. Literatürde etkili veri noktalarının belirlenmesine yönelik mevcut teknikler, tek gözleme dayalı tekniklerdir. Ancak, bu tek gözleme dayalı tekniklerden elde edilen bulgular, veri setinde birden fazla etkili veri noktasının varlığında maskeleme ve süpürme sorunları nedeniyle genellikle yanıltıcıdır.

Bu nedenle, bu araştırmada, etkili gözlem problemlerinin optimal grubunun tanımlanması, bir kombinatoryal optimizasyon sorunu olarak kabul edilmiş ve Ikili Parçacık Sürü Optimizasyon (BPSO) yöntemi, lojistik regresyon modelindeki etkili gözlemlerin optimal grubunu belirlemek için yeni bir eşzamanlı strateji olarak kullanılmıştır. Önerilen BPSO tabanlı yöntemin performansı, çeşitli değerlendirme kriterlerine uygun olarak simüle edilmiş çalışmalarla standart tanı yaklaşımlarına göre kontrol edilmiştir.

Anahtar Kelimeler: Lojistik regresyon modeli; Etkili gözlemler; Ikili Parçacık Sürü Optimizasyon; Olabilirlik değişim istatistiği; Maskele problemi; Süpürme problemi

TABLE OF CONTENTS

CHAPTER 4: A PROPOSED TECHNIQUE FOR PARTICLE SWARM OPTIMIZATION TO DETERMINE THE INFLUENTIAL OBSERVATIONS IN A LOGISTIC REGRESSION MODEL

CHAPTER 5: RESULT AND CONCLUSION

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

LR	Logistic Regression
MLR	Multinomial Logistic Regression
ML	Maximum Likelihood
CD	Cook's Distance
DFFITS	Difference of Fit
DFBETAS	Difference of Beta
LD	Likelihood Displacement
MSE	Mean Square Error
AUC	Area Under the Roc Curve
$\mathbf{R^2}$	Coefficient of determination
PSO	Particle Swarm Optimization
BPSO	Binary Particle Swarm Optimization

CHAPTER 1
INTRODUCTION

Regression models are important tools in order to characterize the relationship between a response (dependent) variable and one or more explanatory variables. According to distribution of response variable, there are different types of regression models. When the dependent variable has Bernoulli distribution, in other words, response variable is binary or dichotomous, logistic regression model is employed to infer association between binary response variable and the independent variables [1]. In a binary dependent variable, only two values, "0" and "1" could be taken to indicate results such as success/failure. Binary logistic regression model has become a significant method utilized to estimate the probability that the response will occur as a linear function of one or more continuous and/or dichotomous explanatory variables.

The objective of logistic regression is to discover a convenient function to depict the connection among the dichotomous features of dependent response variables and a group of independent variables.

Explained by Kleinbaum and Klein [2] the mathematical form on which the logistic model was built by $f(z) = \frac{1}{1+e^{-z}}$, $-\infty < z < +\infty$, $f(z)$ values take S-shape as z alters from $-\infty$ to $+\infty$, when z is $-\infty$, the logistic function $f(z)$ is equivalent to 0, and $f(z)$ is equivalent to 1 if z is $+\infty$. Therefore, as illustrated in Figure (2.1), the value of $f(z)$ ranges between 0 and 1, regardless of z value. The logistic model is very popular due to the logistic function $f(z)$ that ranges between 0 and 1. The model is formed to illustrate a probability, that is always a number between 0 and 1. Therefore, it is unthinkable to obtain a risk estimate either above 1 or below 0 for the logistic model. For other possible models, this is not always the same case. Thus, whenever a probability is evaluated, the logistic model is usually the first choice. As illustrated in Figure (1), when we begin at $z = -\infty$ and shift to the right, then as z rises, the value of $f(z)$ hovers close to zero for a while, then begins to rise greatly to 1, and eventually levels of around 1 as z boosts to $+\infty$. The outcome is an elongated, S-shaped figure [2].

The logistic regression model has been commonly used in a range of fields and in recent years has increased dramatically. Diagnostic tools for model adequacy determination is important part of regression modeling process [15]. There are required conditions or assumptions that must be met in this process. Because the distortion form and the required assumptions form are unreliable and misleading, this model needs to be applied. Therefore checking the model assumption is important. A significant assumption in regression model is that the constructed model is appropriate for all observations in the data. But, it is quite difficult to obtain a dataset in which the all assumptions are met. Furthermore, logistic model is not robust model. It means that even one single unusual observation is sufficient to affect the modeling process in a bad way. These observations can cause bad effect on the estimates obtained from the model. It is obvious that not all observations in the data have same role when constructing the model. Data points that dramatically affect the estimations in model are called as influential observations. An outlier observation is a data point with large residual that does not follow general trend of the observations, while influential observations are data points that have an effect on any part of the model results (parameter estimates, model adequacy and model assumption). [17, 18, 19, 20, 21, 22, 23, 24]. Hawkins' definition [19] perfectly catches the essence and spirit of the word: "An outlier is an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism".

The existence of influential observations would possibly result in distorted analysis and ambiguous results [25, 26, 27], and therefore, in the interpretation of the outcome, it is essential to be sensitive to influential observations and take account of them. The points of an outlier are extremely closely related to influential observations. These Influential observations are illustrated as points that have a noticeably higher influence on the computed values of various estimations, either separately or in combination with other observations (coefficients, standard errors, t-values etc.) [28]. Outliers and influential observations could arise during logistic regression as misclassification between binary (0, 1) answers. Points on the incorrect side of the hyperplane/classifier are referred to as misclassification [29]. It can happen when there is a significant change in the predictor (explanatory) variables, which causes a deviation in the response (labels).

Diagnostic measures are particular amounts calculated from data in order to detect influential points that may be used to eliminate or fix these influential points. Changes in the regression model can be caused by the absence of each observation, but removing the influential observations will create major changes in the model [30]. In the logistic regression model, several traditional diagnostic approaches are available to determine a single influential observation, which are DFBETAS statistics [64], DFFITS [24] Likelihood displacement statistics [41], Cook`s Distance (CD) [36]. Much of the approaches focused on the idea that the single observation is excluded from the data set to explore any improvements to regression coefficients with respect to the regression fitting. In truth, not all of these approaches can particularly be utilized to detect multiple influential observations. Due to the fundamental issues of masking and swamping, the results of these techniques are usually disappointing [31]. In case a single influential observation is not determined as an influential observation, the masking problem arises because influential observations in multiple forms cover each other effectively. In the different manner, when natural observations are mistakenly determined as an influential observation, a swamping problem arises. So it is known that it is important to use a simultaneous case method or multiple case method instead of the techniques mentioned previously. For the accurate detection of influential observations, as they are more efficient in preventing these problems as well as determining the optimal set of influential observations. Some authors introduce some ways to detect multiple influential observations. These are Burcin Coskun & O. Alpu [65] who proposed two novel multiple influential observation diagnostic measures (Generalized Cook Distance based on Generalized Standardized Pearson Residuals- GCD. GSPR and Modified Cook Distance-mCD*) for the model named as logistic regression model, A.A.M.Nurunnabi and A.H.M. Rahmatullah Imon and M. Nasser [66] proposed a new criterion for the determination of multiple influential observations in logistic regression on the basis of a generalized version of DFFITS. A.A.M. Nurunnabi, M. Nasser & A.H.M.R. Imon [67] introduced a resilient influence distance that has the ability of identifying multiple Influential observations.

A. H. M. Rahmatullah Imon a & Ali S. Hadi [33] developed a generalized version of standardized Pearson residuals (GSPR) on the basis of group deletion and then proposed a technique to determining multiple outliers. Simultaneous methods can tackle these problems by concurrently looking for ideal solutions in the search space. The simplest solution of multiple case influence observation detection systems can be generated by taking all potential influence observation combinations. This means that all potential permutations for data set observations are put together in two sub-sets; influential and non-influent, and that a choice is then taken depending upon which the best combination is formed. Although it is nearly difficult to do this practically, it needs so many potential subsets and combinations with substantial computation. Thus, it may be viewed as a combinatorial optimization problem to examine the ideal group of influencing observations. Therefore, meta-heuristic algorithms, nowadays, have been widely developed to address optimization problems, which require little or no assumptions about a problem and may search in very wide areas for possible solutions. Because of their capacity to explore the world and use local resources, population-based meta-heuristic algorithms are ideal for global searches [32]. It is to present a new technique for identifying the best collection of observations with a strong effect on the partial likelihood function and therefore parameter estimations, and also the model's predictive abilities in a Logistic regression model based on the meta-heuristic algorithm. Therefore, in this research, the naturally generated population dependent Binary Particle Swarm Optimization (BPSO) method used by Kennedy and Eberhart [51] was utilized for the multiple case analysis approach to represent the optimal collection of influential observations utilizing the objective function, designed to prevent possible masking and swamping difficulties in the Logistic Model through the use of likelihood displacement statistic. This is done in order to strengthen the logistic regression estimate with the identification and elimination of influence.

In this analysis, the aim of using the BPSO is that it has a structure which is basic, easy to use, quick and inexpensive, with few adjusting parameters and a global search strategy that is less dependent on the starting point [32], [68]. Thus, without the need for onerous computations, the optimum collection of influential observations would be determined simultaneously. BPSO is used as an experimentation method for this aim. With the aid of numerous simulation trials and actual data sets, the suggested system efficacy was checked. In Chapter 2, the logistic regression model, likelihood function, confusion matrix and area under the ROC curve are discussed. In Chapter 3, the influential observations and some traditional techniques to determine a single influential observation are given. In Chapter 4, the proposed BPSO-based method is introduced. A complete simulation study is used to demonstrate the performance of suggested BPSO-based method. In Chapter 5, result of simulation study is presented.

CHAPTER 2
LOGISTIC REGRESSION

2.1. Introduction

Regression models are important tools in order to characterize the link between a response (dependent) variable and one or more explanatory variables. According to distribution of response variable, there are different types of regression models. When the dependent variable has Bernoulli distribution, in other words, response variable is binary or dichotomous, logistic regression model is emplyed to infer partnership between binary response variable and the independent variables [1]. In a binary dependent variable, only two values, "0" and "1" could be taken to indicate results such as success/failure. Binary logistic regression model has become a significant method utilized to estimate the probability that the response will occur as a linear function of one or more continuous and/or dichotomous explanatory variables.

The objective of logistic regression is to discover a convenient function to depict the connection among the dichotomous feature of dependent response variables and a group of independent variables.

2.2 The Logistic Regression Model

The mathematical form on which the logistic model was built was explained by Kleinbaum and Klein [2] by:

$$f(z) = \frac{1}{1+e^{-z}} \quad , \ -\infty < z < +\infty \qquad (2.1)$$

$f(z)$ values take S-shape as z alters from $-\infty$ to $+\infty$, when z is $-\infty$, the logistic function $f(z)$ is equivalent to 0, and $f(z)$ is equivalent to 1 if z is $+\infty$. Therefore, as illustrated in Figure (1), the value of $f(z)$ ranges between 0 and 1, regardless of z value.

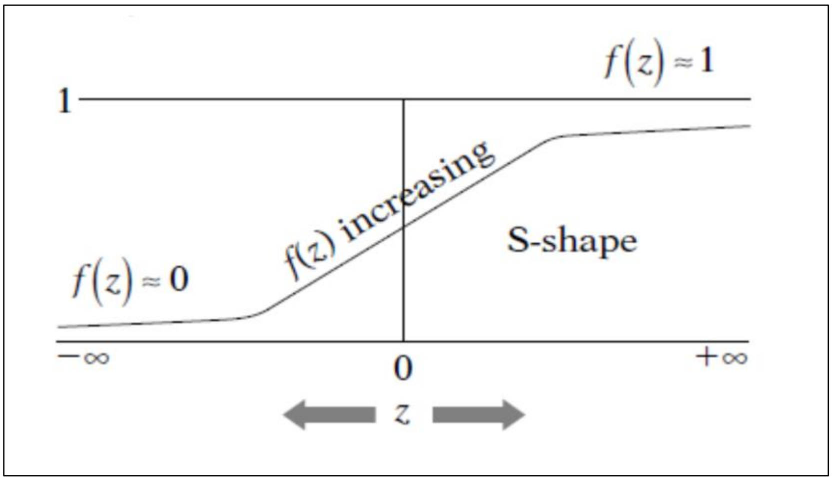


Figure 1: Curve of the logistic function

The logistic model is very popular due to the logistic function $f(z)$ that ranges between 0 and 1. The model is formed to illustrate a probability, that is always a number between 0 and 1. Therefore, it is unlikely to gain a risk estimate either above 1 or below 0 for the logistic model. For other possible models, this is not always the same case. Thus, whenever a probability is evaluated, the logistic model is usually the first choice. As illustrated in Figure (1), when we begin at $z = -\infty$ and shift to the right, then as z rises, the value of $f(z)$ hovers close to zero for a while, then begins to rise greatly to 1, and eventually levels of around 1 as z boosts to $+\infty$. The outcome is an elongated, S-shaped figure (Kleinbaum and Klein) [2].

Assume a generalized linear model like this formula **[3]**

$$y_i = x_i' \beta + \varepsilon_i \qquad (2.2)$$

Where,

$x_i' = [1, x_{i1}, \dots\dots, x_{ip}]$ Data matrix, $i = 1, 2, \dots, p$ and

$p =$ number of independent variables

$\beta = [\beta_0, \beta_{1,} \dots\dots, \beta_p]$ Coefficient vectors

$\varepsilon_i =$ The error terms

And y_i is the response variable that picks the value 0 or 1. Suppose, the variable y_i could be a Bernoulli random variable together with probability distribution as takes after:

$$p(y_i = 1) = \pi_i$$

$$p(y_i = 0) = 1 - \pi_i$$

Presently since $E(\varepsilon_i) = 0$, $E(y_i) = \pi_i$, this suggests that

$$E(y_i) = x_i' \beta = \pi_i \tag{2.3}$$

In case the answer is binary, the error expression ε_i can have two values,

$$\varepsilon_i = \begin{cases} 1 - (x_i' \beta) \, , when \;\; y_i = 1 \\ \quad -x_i' \beta \quad\;\; , when \;\; y_i = 0 \end{cases} \tag{2.4}$$

The logit response function has the form:

$$\pi_i = \frac{1}{1+e^{-\left(x_i' \beta\right)}} \tag{2.5}$$

The binary regression logistic model determines the likelihood of the selected answer on the basis of the values of the explanatory variables. The main difficulty with the linear probability model is that there are limits to probabilities at 0 and 1. However, linear functions are inherently limitless. The fix is to shift the possibilities so that they are no longer limited. The probability is converted into odds by lowering the upper limit and the natural odds logarithm. Therefore, configuring the outcome to match a linear function of the separate variables requires a logit or a binary response model [4]. All of these demonstrate details that are appeared here in [3,5,6].

$$logit = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = x_i'\beta \qquad (2.6)$$

Look that after accepting e to both sides of Equation (2.6), we get

$$odd = \frac{\pi_i}{1-\pi_i} = e^{x_i'\beta} \qquad (2.7)$$

Assume a test of n independent observations of the group $(x_{i1}, \dots, x_{ip}, y_i), i = 1, 2, \dots., n$, where y_i indicates the value of a dichotomous result variable (which is coded as 0 or 1, speaking to the nonappearance or the appearance of the characteristic, separately) and x_{ip} is the value of the i^{th} $(p = 1, 2, \dots P)$ independent variables for the i^{th} subject. The equation can be expressed as [1].

$$\pi_i(y_i = 1) = \frac{e^{(\beta_0 + x_{i1}\beta_1 + \cdots\dots + x_{ip}\beta_p)}}{1 + e^{(\beta_0 + x_{i1}\beta_1 + \cdots\dots + x_{ip}\beta_p)}}$$

$$= \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \qquad (2.8)$$

Where x_i is the i^{th} row of $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$ which is an $n \times (p+1)$ matrix of values for independent variables. The regression coefficients are a vector of the $\beta = (\beta_{0,}\beta_1, \dots, \beta_p)$.

The response variable y_i is the Bernoulli random variable in logistic regression, and the probability mass function of y_i is [7]:

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad , \; i = 1\,,2\,,\ldots,n \tag{2.9}$$

2.3 Likelihood Function and Maximum Likelihood Estimation (MLE)

2.3.1 Likelihood Function

The likelihood function is similar in the figure to the probability density function, except the function parameters are reversed: the likelihood function communicates the values of β with respect to well-known, constant values for y. Hence, the likelihood function for controlling the data can be given as **[7]**:

$$L(\beta|y) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{2.10}$$

2.3.2 Maximum Likelihood Estimates (MLE):

The estimate of logistical models in anomalous data is one of the most strong questions to cite **[8]**, The ML (Maximum Likelihood) technique is the most common method used to estimate the logistic regression model parameter **[7].**The (MLE) is the way in which the parameters (β) are estimated to maximize the likelihood function in Equation (2.10). The basic point (maxima or minima) of a function happen when the first derivative equals 0. Nonetheless, the critical point is maximum in case the second derivative assessment at that point is less than zero (See a great calculus content, like **Spivak**). Hence, we get the result of (MLE) from calculating primary and secondary derivatives of the likelihood function. To pick the derivative of Equation (2.10) with respect to β. After modifying conditions, we can write the maximized equation as follows:

$$L(\beta|y) = \prod_{i=1}^{n} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} (1-\pi_i) \quad (2.11)$$

Let $$\frac{\pi_i}{1-\pi_i} = e^{\sum_{p=0}^{P} x_{ip}\beta_p} \quad (2.12)$$

Which, after some algebraic from Equation (2.12) π_i becomes

$$\pi_i = \frac{e^{\sum_{p=0}^{P} x_{ip}\beta_p}}{1+e^{\sum_{p=0}^{P} x_{ip}\beta_p}} \quad (2.13)$$

Replacing Equation (2.12) and Equation (2.13) for the first and second term on Equation (2.11) respectively, the result will be:

$$L(\beta|y) = \prod_{i=1}^{n} \left(e^{\sum_{p=0}^{P} x_{ip}\beta_p}\right)^{y_i} \left(1 - \frac{e^{\sum_{p=0}^{P} x_{ip}\beta_p}}{1+e^{\sum_{p=0}^{P} x_{ip}\beta_p}}\right) \quad (2.14)$$

Equation (2.14) can be composed as:

$$= \prod_{i=1}^{n} \left(e^{y_i \sum_{p=0}^{P} x_{ip}\beta_p}\right) \left(1 + e^{\sum_{p=0}^{P} x_{ip}\beta_p}\right)^{-1} \quad (2.15)$$

Usually, the part of the likelihood function to maximize. The ordinary method in connection with estimating β is called the (ML) method which takes the values of β that maximize the above likelihood function. Mathematically, it is much simpler to maximize the log-likelihood function which is defined as:

$$logL(\beta|y) = l(\beta) = \sum_{i=1}^{n} log\left(e^{y_i \sum_{p=0}^{P} x_{ip}\beta_p}\right) + log\left(1 + e^{\sum_{p=0}^{P} x_{ip}\beta_p}\right)^{-1}$$

$$= \sum_{i=1}^{n} y_i log\left(e^{\sum_{p=0}^{P} x_{ip}\beta_p}\right) - \log(1 + e^{\sum_{p=0}^{P} x_{ip}\beta_p})$$

$$= \sum_{i=1}^{n} y_i\left(\sum_{p=0}^{P} x_{ip}\beta_p\right) - \log(1 + e^{\sum_{p=0}^{P} x_{ip}\beta_p}) \qquad (2.16)$$

To discover the critical points from the log likelihood function, we have to design the initial derivative with regard to every β on par with zero agreeing with equation (2.16).

$$\frac{\partial}{\partial\beta_p}\sum_{p=0}^{P} x_{ip}\beta_p = x_{ip} \qquad (2.17)$$

$$\frac{\partial l(\beta)}{\partial\beta_p} = \sum_{i=1}^{n}\left[y_i x_{ip} - \frac{1}{1+e^{\sum_{p=0}^{P} x_{ip}\beta_p}}\frac{\partial}{\partial\beta_p}\left(1 + e^{\sum_{p=0}^{P} x_{ip}\beta_p}\right)\right]$$

$$= \sum_{i=1}^{n}\left[y_i x_{ip} - \frac{e^{\sum_{p=0}^{P} x_{ip}\beta_p}}{1+e^{\sum_{p=0}^{P} x_{ip}\beta_p}} x_{ip}\right]$$

$$= \sum_{i=1}^{n}\left[y_i x_{ip} - \pi_i x_{ip}\right]$$

$$= \sum_{i=1}^{n} x_{ip}(y_i - \pi_i) \qquad (2.18)$$

The (MLE) with regard to β can be established by checking every $P + 1$ equations in Equation (2.18) which is to zero to find every β_p .

If the matrix of the second partial derivative is non-positive, at that moment the critical point will be maximum. The generalized form of the matrix of second partial derivative with respect to $\beta_{p'}$ for Equation (2.18) is:

$$\frac{\partial^2 l(\beta)}{\partial\beta_p\partial\beta_{p'}} = \frac{\partial}{\partial\beta_{p'}}\sum_{i=1}^{n} x_{ip}(y_i - \pi_i)$$

$$= \frac{\partial}{\partial\beta_{p'}}\sum_{i=1}^{n}\left[y_i x_{ip} - \pi_i x_{ip}\right]$$

$$= -\sum_{i=1}^{n} x_{ip}\ \frac{\partial}{\partial\beta_{p'}}\left(\frac{e^{\sum_{p=0}^{P} x_{ip}\beta_p}}{1+e^{\sum_{p=0}^{P} x_{ip}\beta_p}}\right)$$

$$= -\sum_{i=1}^{n} x_{ip}\pi_i(1-\pi_i)\, x_{ip'} \qquad (2.19)$$

Newton's method is the foremost known method for finding out the systems of non-linear equations, and it is renowned as the Newton Raphson method as well **[7]**

2.4 Newton-Raphson method:

Setting the equation equal to zero in equation (2.18) results in a series of nonlinear equations of $P+1$ each of the unknown variables $P+1$. The vector with elements β_p is the solution to the system. After proving that the second partial derivative matrix is less than zero, it is the global maximum instead of the local maximum. Thuse, we may conclude that this vector includes estimates of parameters with the highest probability of occurrence of the observed data. On the other hand, the system of nonlinear equations is difficult to solve, and the answer cannot be deduced algebraically, as is the case with linear equations. An iterative method should be used to approximate the answer numerically. Newton's technique is probably the most widely used solving method for systems of nonlinear equations.

Newton’s approach starts with an expectation for the solution and then employes the initial two terms of Taylor polynomial which were assessed at the start quess to get a new estimate which is nearer to the solution. This operation continued until the rapprochement of the genuine result. We see that the first n terms in the Taylor series for f at point $x = x_0$ is

$$\sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!}(x-x_0)^j \qquad (2.20)$$

Assuming that all of f is first n derivatives at x_0 exist. The equation of a tangent line is also the premier degree of Taylor polynomial for f at the point $(x_0, f(x_0))$. For the next approximation of the root to be found where $f(x) = 0$, we will use the point $(x, 0)$ which is cutting the tangent line with the x-axis. The first step of Newton's method is to require the first degree Taylor polynomial as a guess for f, which we need to set a break-even with to zero:

$$f(x_0) + f'(x_0) \times (x - x_0) = 0 \qquad (2.21)$$

After some algebraic to find x, we get:

$$x = x_0 - \frac{f(x_0)}{f'(x_0)} \qquad (2.22)$$

This present value of x will be the next root approximation. We make $x_1 = x$ and continue to produce $x_2, x_3, \ldots,$ until the successive approximations have converged.

It is not difficult to generalize Newton's mechanism to a system of equations. In this case, those in Equation (2.18) are the equations whose roots we want to find a solution to the first derivatives of the equation of log-likelihood. In fact, as equation (2.18) is a system of P + 1 we want to get its roots at the same time, the use of matrix notation to express every step of the Newton-Raphson is more convenient. The equation (2.18) can be written as $l'(\beta)$. For each β_p, Let the vector of initial approximations be represented by $\beta^{(t-1)}$, then the first step of Newton-Raphson could be shown as:

$$\beta^{(t)} = \beta^{(t-1)} + \left[-l"\left(\beta^{(t-1)}\right)\right]^{-1} \cdot l'\left(\beta^{(t-1)}\right) \qquad (2.23)$$

Let α be a length n column vector with elements $\alpha_i = \pi_i$. Notice that it is also possible to write each element of α as $\mu_i = E(y_i)$. Using the multiplication of matrices, we can demonstrate that:

$$l'(\beta) = X'(y - \alpha) \tag{2.24}$$

Is a length of $P + 1$ column vector with the element $\frac{\partial l(\beta)}{\partial \beta_p}$, as derived from the equation (2.18). Now let V be a square matrix of order n, with $\pi_i(1 - \pi_i)$ elements on the diagonal and all others be zeros elsewhere. Again, we can check that by using matrix multiplication

$$l"(\beta) = -X'VX \tag{2.25}$$

Which is a square matrix $(P + 1) \times (P + 1)$ with the elements $\frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_{p'}}$. The equation (2.23) can now be written down as:

$$\hat{\beta}^{(t)} = \hat{\beta}^{(t-1)} + [X'VX]^{-1} \cdot X'(y - \alpha) \tag{2.26}$$

Where $V = diag\big(\hat{\pi}_i(1 - \hat{\pi}_i)\big)$ which was newly estimated by using $\hat{\beta}^{(t-1)}$. By expressing previous iterative updates with a given algebraic operation, $\hat{\beta}_{ML}$ in equation (2.27) is obtained at convergence, leading to an iteratively reweighted least square solution (IRLS):

$$\hat{\beta}_{ML} = [X'VX]^{-1} \cdot X'VZ, \tag{2.27}$$

Where, on the basis of the outcome in the $(t - 1)$th iterative, $V = diag\big(\hat{\pi}_i(1 - \hat{\pi}_i)\big)$ with $\hat{\pi}_i$ and $Z = X\beta^{(t-1)} + V^{-1}(y - \alpha)$.

2.5 Confusion Matrix

The confusion matrix is an instrument which effortlessly and potently states performance of classifier and it has the benefit of being simple to illustrate the finding. The performance of any models and algorithm can be evaluated using confusion matrix. The row of the confusion matrix indicate the values of the predictive class, while the columns indicate the actual class values. Each cell exemplifies one of the probable mixture of actuality and predictive. True positives (TP), false positives (FP), false negatives (FN), and true false (TF) are the four types of the 2x2 confusion matrix **[9].** The excellent model will only have values on the diameter, with all other cells being zeros, while the poor model will be distributed alike across all cells. The measure of how bad a model is can be shown by the error matrix. A misclassified pattern can be determined by each cells value **[10]**

Table 1: Confusion matrix

Confusion matrix		Actual Values	
		P	N
Predicted Values	P	True Positives	False Positives
	N	False Negatives	True Negatives

So as to summarize the outcome of confusion matrix, the accuracy, sensitivity, precision, and specificity methods can be used.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.28}$$

The accuracy is gained by dividing the exactly expected number (TP+TN) by all samples collectively, and is embodied by (2.28). With all methods that have been mentioned above so as to summarize the outcome of confusion matrix, precision and sensitivity are the most prevalent methods as seen in (2.29) and (2.30) in turn.

$$Precision = \frac{TP}{TP+FP} \qquad (2.29)$$

Precision is a positive predictive value that determines how many of the predicted positive samples (TP+FP) are genuine positives (TP). When attempting to decrease the amount of false positives (FP), precision is utilized as a performance measure.

$$Sensitivity = \frac{TP}{TP+FN} \qquad (2.30)$$

The number of total positive samples (TP+FN) categorized as positive classes (TP) is measured by sensitivity.

$$Specificity = \frac{TN}{TN+FP} \qquad (2.31)$$

The number of total negative samples (TN+FP) categorized as negative classes (TN) is measured by specificity.

2.6 Area under the ROC curve

To categorize a test result as positive, sensitivity and specificity rely on a single cut point **[15]**. The area under the ROC (Receiver Operating Characteristic) curve provides a more detailed assessment of categorization accuracy. This curve, which comes from signal detection theory, illustrates how the receiver handles signal presence when there is noise. For a full range of feasible cut points, it shows the likelihood of detection genuine signal (sensitivity) and false signal (1-specificity).

$$AUC = \int_{-\infty}^{\infty} TPR(C)FPR'(C)\, dc \tag{2.32}$$

TPR(C) and FPR' (C) represent the true positive rate (sensitivity), as well as false positive rate (1-specificity) for a given cutoff or threshold value, in turn. The AUC is computed as the region under the ROC curve that traces the TPR at different thresholds according to FPR. **[11], [12].**

2.7 Mean Square Error (MSE) for MLE:

The function β given by $E(\hat{\beta} - \beta)^2$ is the mean square error (MSE) of the estimator β for parameter $\hat{\beta}$, and this is referred to as $MSE_{\widehat{\beta}}$.This is also known as estimators risk function, with the quadratic loss function called $(\hat{\beta} - \beta)^2$. In comparison with other distance scales, MSE has at least two advantages: First, it can be analytically monitored and, secondly, the interpretation of it is **[13]**

$$MSE_{\widehat{\beta}} = E(\hat{\beta} - \beta)^2 = Var(\hat{\beta}) + (E(\hat{\beta}) + \beta)^2$$

$$= Var(\hat{\beta}) + (Bias\ of \hat{\beta}\ )^2 \tag{2.33}$$

The discrepancy between the expected value of $\hat{\beta}$ and the real value of β is the bias of an estimator $\hat{\beta}$ of a parameter β; that is, $Bias(\hat{\beta}) = E(\hat{\beta}) - \beta$. The unbiased estimator is called an estimator whose bias is identically equal to 0 and satisfies $E(\hat{\beta}) = \beta$ for all β. Therefore, the mean square error (MSE) has two parts, one measures the estimator uncertainty (precision), and the other measures its bias. The combined variance and bias of an estimator that has good mean square error (MSE) properties is minimal. We need to find estimators that control both variance and bias to discover an estimator with strong mean square error properties.

For an unbiased $\hat{\beta}$ estimator, we have:

$$MSE_{\widehat{\beta}} = E\left(\hat{\beta} - \beta\right)^2 = Var\left(\hat{\beta}\right) \qquad (2.34)$$

According to **[14]** asymptotically $\hat{\beta}_{ML}$ is commonly distributed, and the asymptotic variance-covariance matrix is equal to the inverse of the Fisher information matrix, that is computed by

$$Var\left(\hat{\beta}_{ML}\right) = [X'VX]^{-1} \qquad (2.35)$$

And since $\hat{\beta}_{ML}$ is unbiased estimator of β, MSE of $\hat{\beta}_{ML}$ is obtained as

$$MSE\left(\hat{\beta}_{ML}\right) = tr\left(Var\left(\hat{\beta}_{ML}\right)\right) + E\left(\hat{\beta}_{ML} - \beta\right)' E\left(\hat{\beta}_{ML} - \beta\right)$$

$$= tr((X^T VX)^{-1}) = \sum_{j=1}^{p} \frac{1}{d_j} \qquad (2.36)$$

Where the j^{th} eigenvalue of $X'VX$ is d_j, and the trace operator is tr.

2.8 Model Assumptions of Logistic Regression:

It has to satisfy the assumption of logistic regression so as to satisfy the validity of the model. The general assumptions used in logistic regression analysis are as follows:

1- A logistical function of the explanatory variables is the conditional probabilities.

2- No major variables will be omitted.

3- No irrelative variables are included.

4- Independence of errors.

5- The observation of the variables are independent.

6- The explanatory variables do not depend linearly on each of them.

7- The error of the model is distributed binomially.

8- There should be no outlier and influential point in the dataset

Evaluation and validation of model performance should be an additional and critical step after adapting the logistic regression model and before reaching any conclusion based on fit. Evaluating the results of logistic regression with the influential outlier observations in the data set is the subject of this thesis.

CHAPTER 3

THE INFLUENTIAL OBSERVATIONS IN THE LOGISTIC REGRESSION

3.1 Introduction

The logistics regression model has been commonly used in a range of fields and in recent years has increased dramatically. His success raises the need for diagnostic instruments to assess the model's suitability. "Diagnosis has been an important element of logistical stagnation in recent years," said Hosmer and Lemeshow [15]. An outlier observation is a data point with large residual that does not follow general trend of the observations. Data points that significantly affect the estimations in model are called as influential observations [16]. Such findings are identified and their effects on the binary logistics model are studied. In this part, we study the form and the detection methods of outliers and influential observations in LR.

3.2. Outlier and Influential Observations in Logistic Regression Model

The logistic regression model has been commonly used in a range of fields and in recent years has increased dramatically. Diagnostic tools for model adequacy determination is important part of regression modeling process [15]. There are required conditions or assumptions that must be met in this process. Because the distortion form and the required assumptions form are unreliable and misleading, this model needs to be applied. Therefore checking the model assumption is important. A significant assumption in regression model is that the constructed model is appropriate for all observations in the data. But, it is quite difficult to obtain a dataset in which all the assumptions are met. Furthermore, logistic model is not robust model. It means that even one single unusual observation is sufficient to affect the modeling process in a bad way. These observations can cause bad effect to the estimates obtained from the model. It is obvious that not all observations in the data have same role when constructing the model. Data points that substantially affect the estimations in model are called as influential observations. An outlier observation is a data point with large residual that does not follow general trend of the observations, while influential observations are data points which possess an effect on any part of the model results (parameter estimates, model adequacy and model assumption). [17, 18, 19, 20, 21, 22, 23, 24]. Hawkins' definition [19] perfectly catches the essence and spirit of the word: "An outlier is an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism".

The presence of influential observations would possibly result in distorted analysis and ambiguous results [25, 26, 27], and therefore, in the interpretation of the outcome, it is essential to be sensitive to influential observations and take account of them. The points of an outlier are extremely closely related to influential observations. Influential observations are accounted for as points that have a noticeably higher influence on the computed values of various estimations, either separately or in combination with other observations (coefficients, standard errors, t-values etc.) [28]. Outliers and influencing observations may arise during logistic regression as misclassification between binary (0, 1) answers. Points on the incorrect side of the hyperplane/classifier are referred to as misclassification [29]. It can happen when there is a significant change in the predictor (explanatory) variables, which causes a deviation in the response (labels).

Diagnostic measures are particular amounts calculated from data in order to detect 0nfluential points that may be used to eliminate or fix these influential points. Changes in the regression model can be caused by the absence of each observation, but removing the influential observations will create major changes in the model [30].

3.3 Literature Review and Problem Statement

A residual vector and projection matrix [8] are the fundamental building blocks of the logistic regression to evaluate outlying and influential points. The approach originally used for linear regression and logistic regression is also used to expand this thinking and use linear regression approximations as pointed out in citation [8]. In the logistic regression model, several traditional diagnostic approaches are available to determine a single influential observation, which are DFBETAS statistics [64], DFFITS [24] Likelihood displacement statistics [41], Cook`s Distance (CD) [36].

Many of the approaches focused on the idea that the single observation is excluded from the data set to explore any improvements to regression coefficients with respect to the regression fitting. In truth, not all of these approaches can particularly be utilized to detect multiple influential outlying observations. Due to the fundamental issues of masking and swamping, the results of these techniques are usually disappointing [31]. When one influential observation is not determined as an influential observation, the masking problem arises because multiple influential observations cover each other effectively. Reversely, when natural observations are mistakenly determined as an influential observation, a swamping problem arises. So it is known that it is important to use a simultaneous case method or multiple case methods instead of the techniques mentioned previously for the accurate detection of influential observations, as they are more efficient in preventing these problems as well as determining the optimal set of influential observations. Some authors introduce some ways to detect multiple influential observations. These are Burcin Coskun & O. Alpu [65] who proposed two novel multiple influential observation diagnostic measures (Generalized Cook Distance based on Generalized Standardized Pearson Residuals- GCD.GSPR and Modified Cook Distance- mCD*) for the model known as logistic regression model, A.A.M.Nurunnabi and A.H.M. Rahmatullah Imon and M. Nasser [66] proposed a new measure for the determination of multiple influential observations in logistic regression on the basis of a generalized version of DFFITS. A.A.M. Nurunnabi, M. Nasser & A.H.M.R. Imon [67] introduced a resilient influence distance that has the capability of stating multiple Influential observations. A. H. M. Rahmatullah Imon a & Ali S. Hadi [33] developed a generalized version of standardized Pearson residuals (GSPR) on the grounds of group deletion and then proposed a technique to determining multiple outliers. Simultaneous methods can tackle these problems by concurrently looking for ideal solutions in the search space. The simplest solution of multiple case influence observation detection systems can be generated by taking all potential influence observation combinations. This means that all potential permutations for data set observations are put together in two sub-sets which are influential and non-influent, and that a choice is then taken, depending upon which the best combination is formed. Although it is nearly difficult to do this practically, it needs so many potential subsets and combinations with substantial computation. Thus, it may be viewed as a combinatorial optimization problem to examine the ideal group of influencing observations.

In order to solve optimization issues with a high-dimensional search space, exact optimization algorithms are unable to give a suitable solution. Exhaustive search is not practicable in these cases because the search space expands steadily with the problem size. Classical approximation techniques of optimization such as greedy algorithms provide several hypotheses for resolving issues. Sometimes, in every situation, it is difficult to validate these assumptions, therefore, meta-heuristic algorithms, nowadays, have been widely developed to address optimization problems, which require little or no assumptions about a problem and it is possible to do a search across wide areas for possible solutions. Because of their capacity to explore the world and use local resources, population-based meta-heuristic algorithms are ideal for searches across the globe [32]. This is so to present a new technique for identifying the best collection of observations with a strong effect on the partial likelihood function and therefore parameter estimations, and also the model's predictive abilities in a Logistic regression model on the grounds of the meta-heuristic algorithm. Therefore, in this research, the nature-inspired population-based Binary Particle Swarm Optimization (BPSO) method used by Kennedy and Eberhart [51] was used for the multiple case analysis approach to represent the optimal collection of influential observations utilizing the objective function, which is designed to prevent possible masking and swamping difficulties in the Logistic Model by the use of the likelihood displacement statistic. This is done in order to strengthen the logistic regression estimate with the identification and elimination of influence. In this analysis, the aim of using the BPSO is that it has a structure which is basic, easy to use, quick and inexpensive, with few adjusting parameters and a global search strategy that is less dependent on the starting point [32], [68]. Thus, without the need for onerous computations, the optimum collection of influential observations would be determined simultaneously. BPSO is used as an experimentation method for this aim. With the aid of numerous simulation trials and actual data sets the suggested system efficacy was checked.

3.4 Detection Methods of Influential Observations in Logistic Regression

3.4.1 Residual and leverage

The residual i^{th} is accounted for as the difference between the observed value and the fitted value in linear regression$(y_i - \hat{y}_i)$ [33]. To stress the fitted values in logistic regression for each covariate pattern, we record the fitted value of the i^{th} covariate pattern as $\hat{y}_i = \hat{\pi}_i$ for that covariate pattern. As a result, the i^{th} residual is calculated as follows:

$$\hat{\epsilon}_i = y_i - \hat{\pi}_i, \qquad i = 1,2,\ldots,n \tag{3.1}$$

The hat matrix is of critical importance for the study's linear regression. The values for dropping the outcome variable into the covariate space are provided by this matrix. The residuals linear regression $\left(Y - \hat{Y}\right)$ is also described by the hat matrix so that this forms a variety of studies. Pregibon [8] considered a linear approach to the fitted values that create a hat matrix for logistical regression by using the linear regression of the weighted least squares.,

$$H = V^{\frac{1}{2}}X(X'VX)^{-1}X'V^{\frac{1}{2}} \tag{3.2}$$

Where V is the $n \times n$ diagonal matrix with the general variable $v_i = \hat{\pi}_i(1 - \hat{\pi}_i)$. The diagonal components of the hat matrix are considered the leverage values of linear regression. To signal by the quantification h_i we indicate the i^{th} diagonal part of matrix H described in (3.2). It is clear to demonstrate this

$$h_i = \hat{\pi}_i(1 - \hat{\pi}_i)x_i'(X'VX)^{-1}x_i \tag{3.3}$$

Where, $x_i' = \left[1, x_{1i}, x_{2i}, \ldots, x_{pi}\right]$ is a $1 \times k$ vector of i^{th} case observations.

The residuals calculate the magnitude of ill-fitted factor/covariate patterns in logistic regression. Thus, the suspicious outliers are the observations that have a significant residual. At this point, however, we have a normal question: How large is this? The residual described in (3.1) is not qualified, so they do not contribute easily to outliers detection. Let us now add several scaled versions of the above residues which are widely used for the diagnosis of outliers.

The Pearson residuals are chi-square elements from Pearson and these can be put to use to classify patterns that are unacceptable. The big assumption in the linear regression is that the variance of the error doesn't depend on conditional mean$E(y_i|x_i) = \hat{\pi}_i$. Nonetheless, we have Bernoulli errors in logistic regressions, there for the error variance is a function of the conditional mean, i.e.

$$Var\left(y_i \middle| x_i\right) = v_i = \hat{\pi}_i(1 - \hat{\pi}_i) \qquad (3.4)$$

For i^{th} factor/covariant pattern, the Pearson residual described is given by

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i}}, \qquad i = 1, 2, \dots, n \qquad (3.5)$$

An observation is referred to as an outlier if the appropriate Pearson residual is over a number c in absolute terms. because residuals from Pearson have been scaled, a rational option maybe 3 for c (refer to [34]), which is consistent with the normal theory's 3-distance law. But we often experience that too many observations are identified as outliers by the cut-off value 3. Chen and Liu [35] can then be accompanied by c as an appropriately selected constant between 3 and 5.

If we apply the Pregibon [8] linear regression-like approximation of the residual for i^{th}, we see that

$$\hat{\epsilon}_i = y_i - \hat{\pi}_i \approx (1 - h_i)y_i \qquad (3.6)$$

As a result, we can calculate the residual variance as

$$V(\hat{\in}_i) = v_i(1 - h_i) \tag{3.7}$$

As a result, the Pearson residues do not have an equivalent variable of 1. That is why we must use the standard Pearson residues, which are denoted by:

$$r_{si} = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i(1-h_i)}}\ , \qquad i = 1,2,\dots,n \tag{3.8}$$

if $|r_{si}| > c$, The i^{th} observation can be labeled an outlier

3.4.2 Cook's Distance

The Cook's distance statistics suggested by Cook [36] measures the Euclidean distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$ which is parameter estimation when i^{th} observation isn't in the data set anymore. The distance of i^{th} Cook is defined in the case of a logistic regression model [30] as:

$$\mathrm{CD}_i = \frac{(\hat{\beta}^{(-i)} - \hat{\beta})'(X'VX)(\hat{\beta}^{(-i)} - \hat{\beta})}{k\hat{\sigma}^2}, i = 1,2,\dots,n, \tag{3.9}$$

Where $\hat{\beta}^{(-i)}$ with the i^{th} observation removed is the estimated parameter of β. Use linear approximations including those proposed in **[8]**, Equation (3.9) can be indicated as:

$$\mathrm{CD}_i \approx \frac{1}{k} r_{si}^2 \left(\frac{h_{ii}}{1-h_{ii}}\right) \tag{3.10}$$

Where r_{si} is the residual of i^{th} standard Pearson defined as:

$$r_{si} = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i(1-h_i)}}\ , \qquad i = 1,2,\dots,n \tag{3.11}$$

Where the leverage value of i^{th} is h_{ii}, which is actually the i^{th} diagonal element of the matrix of leverage

$$H = V^{\frac{1}{2}}X(X'VX)^{-1}X'V^{\frac{1}{2}} \tag{3.12}$$

And V is a matrix diagonal, with the v_i element diagonal defined as follows:

$$V(y_i|x_i) = v_i = \hat{\pi}_i(1 - \hat{\pi}_i) \tag{3.13}$$

An observation is considered influential if the result of Cook's distance is larger than 1. (see Ref.**[37]**)

3.4.3 Difference of Fits (DFFITS)

Belsley, Kuh and Welsch also suggested DFFITS in 1980 and it is based on $\hat{y}_i$ and $\hat{y}_i^{(-i)}$ discrepancy**.** The difference of fits (DFFITS) was first mentioned in Ref. [24], it is characterized by:

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_i^{(-i)}}{\hat{\sigma}^{(-i)}\sqrt{h_{ii}}}, \quad i = 1,2,\ldots,n \tag{3.14}$$

With the i^{th} observations removed, $\hat{y}^{(-i)}$ represents the fitted response and $\hat{\sigma}^{(-i)}$ represents the estimated standard error. DFFITS can be denoted by the remaining Pearson criteria and leverage values as:

$$\text{DFFITS}_i = r_{si}\sqrt{\frac{h_{ii}}{(1-h_{ii})}\,\frac{v_i}{v_i^{(-i)}}} \tag{ 3.15 }$$

If the observation values DFFITS is bigger than $\sqrt{\frac{k}{n}}$, where c is a properly selected constant of 2 to 3 or more **[38, 39, 40]**, then it is termed as an observation.

3.4.4 Difference in Beta (DFBETAS)

The one-step difference between the MLE of the parameter vector and the MLE of the parameter vector without the i^{th} observation is defined as DFBETA and used to measure the influence of the i^{th}observation. A Fisher scoring step is assumed in this one-step and is calculated as follows:

$$\beta - \hat{\beta}^{(i)} \approx DFBETA_i = (X'VX)^{-1}X'_i V_i^{\frac{1}{2}}(1-h_{ii})^{-\frac{1}{2}} r_{si} \qquad (3.16)$$

Where h_i is the leverage and r_{si} is the standardized Pearson residual

DFBETAS is the standardized DFBETA statistics for evaluating the impact of i^{th} observation on the j^{th} regression parameter which is known as DFBETA for the j^{th} parameter divided by it is estimated standard deviation. The standard deviation is calculated using the data in this case.

$$DFBETAS_{j(i)} = \frac{DFBETA_{j(i)}}{\hat{\sigma}(\beta_j)} \qquad (3.17)$$

Value of $DFBETAS_{j(i)}$ greater than two would certainly indicate a major impact from a single point [64]

3.4.5 Likelihood Displacement (LDᵢ) Statistic

The likelihood of displacement, introduced by Cook (1986), is one useful and general technique for comparing $\hat{\beta}$ and $\hat{\beta}^{(-i)}$.

Let β be a $p \times 1$ vector for unknown parameters and $\hat{\beta}$ be the maximum likelihood (ML) estimate of β obtained from a sample of size n. The influence of the i^{th} observation on the parameter estimate can be assessed by studying the difference between β and $\hat{\beta}^{(-i)}$, where $\hat{\beta}^{(-i)}$ denotes the ML estimate of β obtained from the sample of size $n-1$ excluding the i^{th} observation. Likelihood displacement, which is defined as [41]

$$\mathrm{LD}_i(\beta) = 2\left[lnL(\hat{\beta}) - lnL(\hat{\beta}^{(-i)})\right] \qquad (3.18)$$

Where $L(\hat{\beta})$ is the logistic regression partial likelihood function, $\hat{\beta}$ is a vector estimation parameter from the entire data set and $\hat{\beta}^{(-i)}$ is the parameter estimate set obtained when the observation i is removed. Since conventional instruments frequently provide specious outcomes because of masking and swamping difficulties, it is more appropriate to assess the likelihood displacement by eliminating a group of notes rather than eliminating them one by one. In this study, the fitness function of the PSO system is divided by the "number of observations detected as influencers". The goal of this approach is to locate observations that maximize the fitness function.

CHAPTER 4

A PROPOSED TECHNIQUE FOR PARTICLE SWARM OPTIMIZATION TO DETERMINE THE INFLUENTIAL OBSERVATIONS IN A LOGISTIC REGRESSION MODEL

4.1 Introduction

Problems of optimization are common in different disciplines. Sometimes, the objective function or model's actual and practical nature restrictions can make these difficulties exceedingly complicated. Derivative-based approaches, as summarized in [42], [43], have traditionally been used in optimization procedures. These methods are dependable and have been demonstrated to be efficient in a variety of optimization issues. However, these methods may encounter problems including becoming stuck at local minimums, increasing computing complexity, and being unsuitable for certain types of objective functions. This resulted in the necessity to design a new class of techniques to solve such deficiencies. Heuristic optimization approaches are rapidly evolving tools that can overcome the majority of the drawbacks associated with derivative-based methods.

In recent years, many algorithms for obtaining solution's subsets that aren't quite perfect have been presented. Colony optimization (ACO), Genetic Algorithms (GA), as well as particle swarm optimization (PSO) in addition to some other algorithms are examples of Algorithms that are meant to mimic evolutionary process. PSO is a type of evolutionary algorithm that is predicated on swarm intelligence and is relatively new. In comparison to other EA, PSO is less costly and can converge faster. [44]

Particle swarm optimization (PSO) was initially referred to as a new heuristic approach by Kennedy and Eberhart [45], [46]. The original aim of their study was to use arithmetic to encourage the social behavior of fish schools and bird flocks and. As their study proceeded, they found out that with some alterations, their social behavior model can as well be used as an effective optimizer. In the original PSO version, only non-linear issues were dealt with for continuous improvement. However, numerous improvements in PSO development have strengthened their ability to deal with a wide range of difficult engineering and scientific issue optimization. Recent progress in these fields is summarized in [47] and [48].

Several PSO algorithm variations have been proposed, but the most often used is Shi and Eberhart's [49] (Gbest model) global version of PSO, in which the entire population is treated as a single neighborhood during the optimization process. One of its most attractive features is the simplicity of the PSO method, as it only uses two model equations [50]. Each particle has a coordinated position (x_i), and velocity (v_i) is a feasible solution in PSO, using 2 vectors. The two vectors associated with each particle in the search space of N-dimension are $X_i = [x_{i1}, x_{i2}, \dots, x_{in}]$ and $V_i = [v_{i1}, v_{i2}, \dots, v_{in}]$. A swarm is made up of a group of particles (or potential solutions) that move (fly) around the viable solution space in search of the best solution. On the basis of its best investigation, better swarm experience, and its prior speed vector, each particle adjusts its position based on the following model:

$$v_{dn}^{t+1} = wv_{dn}^{t} + \overbrace{c_1 r_1 (p_{dn} - x_{dn})}^{cognitive\ component} + \overbrace{c_2 r_2 \left(p_{gn} - x_{dn}\right)}^{social\ component} \tag{4.1}$$

$$x_{dn}^{t+1} = x_{dn}^{t} + v_{dn}^{t+1} \tag{4.2}$$

In the case when c_1 and c_2 consist of two positive constants, r_1 and r_2 are numbers produced at random for [0,1], w is the weight of inertia, p_{dn} is the best position particle i achieved on the basis of its own experience. While, p_{gn} depends on the total swarm experience which is optimum particle position and t is the iteration index.

Kennedy and Eberhart [51] suggested binary particle swarm optimization to address problems involving combinatorial optimization in binary space (BPSO).

4.2 Original Particle Swarm Optimization (PSO) Algorithm

One of the most successful strategies developed is the optimization of the particulate swarm (PSO). PSO was created by Kennedy and Eberhart [52] as a meta-heuristic optimization method based on the existence of the population. This technique based on swarm intelligence was influenced by means of social conduct of flocking birds, that is dependent on their prior knowledge when in search of food or companion. Each possible solution symbolizes a particle in this stochastic technique. In the PSO method, the population of the particle is said to swarm. Particles are put at random in the problem's search space. In PSO, each particle searches for random locations and velocities and is modified to spot the most effective solutions for each iteration. Every iteration updates the particle locations to a predetermined quality criterion named the objective function. The objective function in the PSO algorithm distinguishes between the particle and the food (or mat) [53]. At each stage, apart from the swarms expertise and personal knowledge, each particle's speed is refined and upgraded based on inertia. Each particle's experience is kept by its most effective location (Pbest). Swarm's knowledge is kept by the swarm's global best place (Gbest). PSO's unique function is that it explores multiple points in diverse regions at the same time for solution space to identify a globally optimal solution.

PSO is an evolutionary method based on population, with numerous significant benefits over other methods of optimization including [54]

- Unlike many traditional approaches, it is a derivative-free algorithm.
- It may be mixed and matched with other optimization methods to create hybrid tools.
- They are less affected by the convexity and continuity of objective functions.
- Unlike many other competing evolutionary methods, they have less parameters to modify.
- It is capable of escaping local minima.
- Simple mathematics and logic operations make it simple to implement and program.
- It can address objective functions having a stochastic character, such as when one of the optimization variables is represented as random.
- To start the iteration process, it doesn't need a premium starting solution.

PSO is a well-known optimizer that has been frequently employed to find a solution for optimization problems. This has made it interesting to develop the performance of the algorithm and theoretical research. Some research on the performance of PSO in topological systems and in parameter studies has also been undertaken [55]

To improve performance in addressing multimodal problems, Kennedy and Mendes presented a ring topological structure PSO (LPSO) [56] and a Von Neumann topological structure PSO (VPSO) [57]. In addition, Liang and Suganthan [58] proposed the Dynamic Multi-Swarm PSO (DMS-PSO) to develop topological structure in a dynamic fashion. Hybrid PSO with different evolutionary paradigms is another current research area in PSO. A PSO selection process similar to GA has been presented by Angeline [59]. In addition, GA and PSO hybridisation was applied [60] to the ongoing construction of artificial neural networks. Beheshti et al. presented further research, based upon a more advanced PSO and Newtonian motion-legs, and also on Median-oriented Particle Swarm Optimization (MPSO) [61].

The use of PSO is contingent on the shape of the problem and structure, that is the problem domain. Initially, PSO was used in continuous space to overcome optimization problems. There were, however, several discrete (or binary) problems with optimization. Therefore, to find solution for the problems of combinatorial optimization in binary space, the binary version of PSO is optimized.

4.3 Binary Particle Swarm Optimization (BPSO) Algorithm

The original PSO method could only be employed to find solutions to issues related to continuous real-valued solution elements. Kennedy and Eberhart [62], who originally introduced the original PS0, devised a version of the PSO method for addressing issues with binary values, like combinational optimization issues. Binary particle swarm optimization is the name of the improved algorithm (BPSO). BPSO has an unusual characteristic in that it utilizes the identical velocity as PSO but replaces the position with the following selection of roulette wheel selection [63]. Each particle's location, $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ in the BPSO approach are stated by binary values in which, $x_{in} \in \{0,1\}$. The distinction between PSO and BPSO in the concept of velocities in continuous space is in the probability that the appropriate element in a particle will be given to the value 1, which defines the velocities of the individual $V_i = (v_{i1}, v_{i2}, \dots, v_{in})$ as defined.

Following are the BPSO algorithm steps:

Step1. Initial X_i and V_i randomly generated positions for each i^{th} particle are calculated as

$$X_i = (x_{i1}, x_{i2}, \dots, x_{in}) \tag{4.3}$$

$$V_i = (v_{i1}, v_{i2}, \dots, v_{in}) \tag{4.4}$$

Where, $x_{in} \in \{0,1\}$, v_{in} is the velocity of the i^{th} particle in the n^{th} dimention. In the swarm, the particles number and position is defined by j and n, for $i = 1,2,\dots,j$ respectively. the velocity of a particle is also confined to $v_{in} \in [v_{min}, v_{max}]$. v_{in} is set to v_{max} when it is greater than v_{max} and if v_{in} is smaller than v_{min}, then v_{in} is set to v_{min}. $v_{min} = -v_{max}$, normally.

Step2. All particle fitness values in the swarm are determined based on the objective function.

Step3. The particles of Pbest$_i$ and Gbest are calculated according to fitness values as in equations (4.5) and (4.6).

$$Pbest_i = \left(p_{i1},\ p_{i2},\ \dots,\ p_{in}\right) \tag{4.5}$$

$$Gbest = \left(p_{g1},\ p_{g2},\ \dots,\ p_{gn}\right) \tag{4.6}$$

Where, Pbest$_i$ and Gbest are the vectors indicating the best locations of the i^{th} particle to date and the best particle with the most reliable fitness value found in the entire swarm.

Step4. Using equation (4.7) and equation (4.8) are needed respectively, to update velocity and position.

$$v_{dn}^{t+1} = w \times v_{dn}^{t} + \overbrace{c_1 \times r_1 \times (p_{dn} - x_{dn})}^{cognitive\ component} + \overbrace{c_2 \times r_2 \times \left(p_{gn} - x_{dn}\right)}^{social\ component} \tag{4.7}$$

$$x_{dn}^{t+1} = \begin{cases} 1 & if \quad r_3 < sigm(v_{dn}^{t+1}) \\ 0 & \qquad otherwise \end{cases} \tag{4.8}$$

Where $sigm(v_{dn}^{t+1})$ is the function of sigmoid limiting transformation, in the interval $[0,1]$, r_1 , r_2 and r_3 represent evenly dispersed numbers at random. c_1 and c_2 are two parameters that respectively Stand for coefficients of social and cognitive manner, whereas w represents the parameter of inertia and t is the present number of iteration.

Step5. Steps 2, 3 and 4 are being used repeatedly till the specified number of iterations has been met.

4.4 The BPSO based Suggested Method for Determining the Influential Observations in Logistic Model

4.4.1 Proposed Method

As mentioned before, single case deletion methods have many drawbacks. When several influential observations have been made, they show misleading results in particular. Considering observations as influential and non-influential, detection of several observations of influential nature in the Logistic Regression Model may be called a problem of combinatorial optimization nature due to issues related to the masking and swamping. The power of BPSOs is because the technique is organized and can quickly solve a large number of problems through combination optimization, even the ones that are hard to address utilizing other methods. Given these features, it is considered that the BPSO solution may be optimal for solving this question of optimization. However, certain key elements should be taken into consideration so as to employ the benefits of the BPSO algorithm and to apply it correctly:

4.4.1.1 Building of Particles

The $X_i = (x_{i1}, x_{i2}, \dots, x_{id}, \dots, x_{in})$ positions of each particle are evaluated as binary values, 0 or 1. In fact, each particle has its own value as mock variable that represents $1 \times n$ vector in which n can be described as observations number belonging to dataset. $x_{id} = 0$ symbolizes the influential mock and $x_{id} = 1$ represents the non-influential mock for the d^{th} observation of the i^{th} particle for each $d = 1, 2, \dots, n$. We have allocated "0" for possible influential observations accepted by the proposed method dependant on BPSO. Thus, the results of these established observations will rest or delete from the model.

4.4.1.2 Definition of Objective Function

The choice of the objective function for simultaneous identification of influencing observations is the most essential aspect of the BPSO. The objective function ought to be appropriate for the purpose of optimization problem. In this analysis, an LD-based objective function of Likelihood Displacement was used to define influential observations in the logistics model.

$$\mathrm{LD}_{0_s} = \frac{2[lnL(\hat{\beta}) - lnL(\hat{\beta}_{(-0_s)})]}{m+1} \tag{4.9}$$

Where, 0_s, is identified as possible observations in particles (or position) and m stands for the number for possible influential observations of influential nature (that is; the 0 s number in particles). As well as this, $\hat{\beta}$ is the vector of the estimates of the parameters gained from the complete dataset ($x_{id} = 1$ for each $d = 1,2,\ldots,n$) and $\hat{\beta}_{(-0_s)}$ represent the vector's estimates of parameters gained when removing the observations m observations identified in the particle as potential influential observations. The statistics of likelihood deisplacement, LDi identified by Cook [37] can be an efficient criterion to classify influential observations in the Logistic model; nonetheless, identical to the downsides of other single-case deletion approaches, it is evident that it is affected by swamping and masking issues as it detects one-by-one influential observations. It'd also be better fitting too to measure the likelihood of displacement by removing a collection of observations instead of singularly removing them.

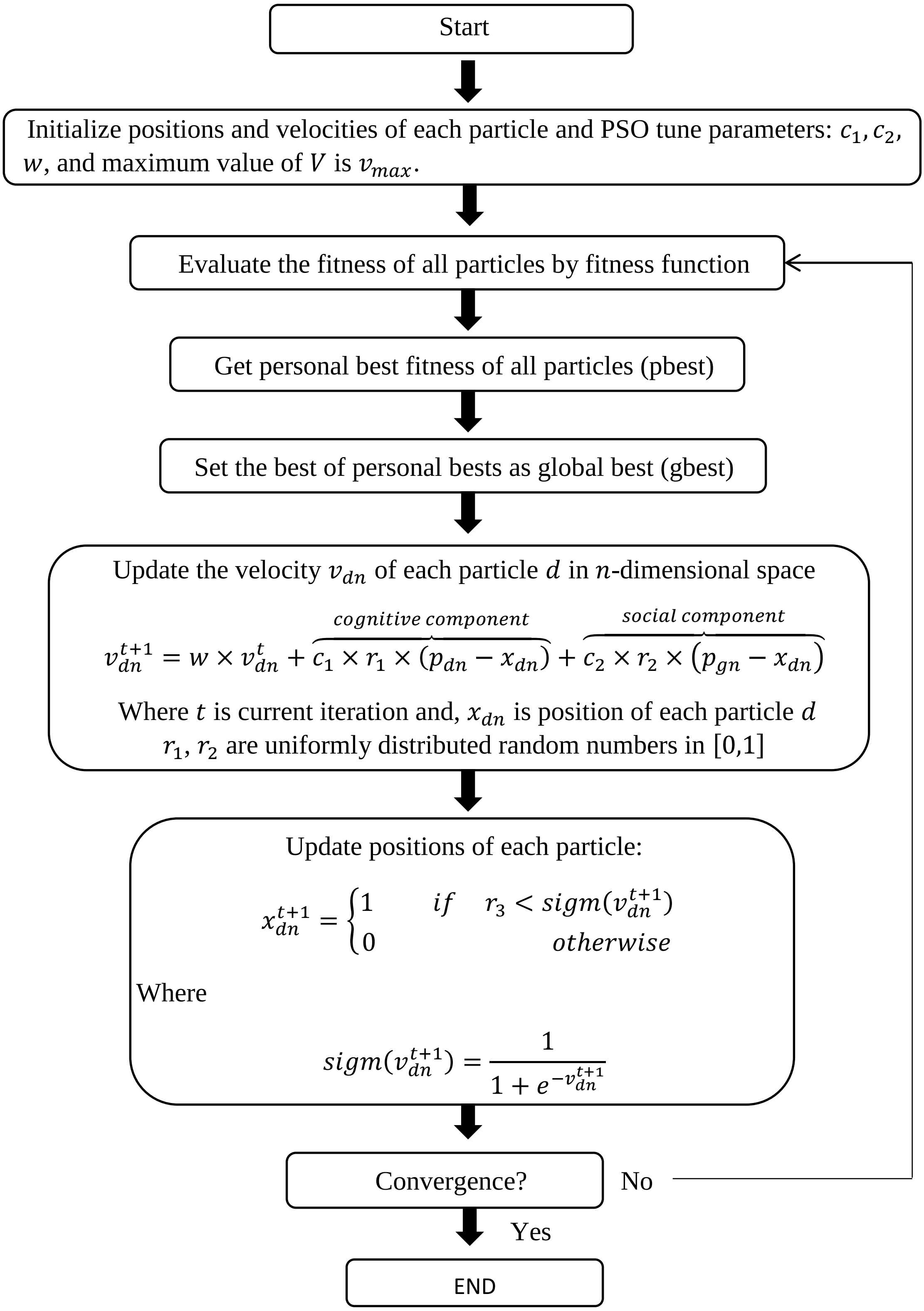

Start
Initialize positions and velocities of each particle and PSO tune parameters: c_1, c_2, w, and maximum value of V is v_{max}.
Evaluate the fitness of all particles by fitness function
Get personal best fitness of all particles (pbest)
Set the best of personal bests as global best (gbest)
Update the velocity v_{dn} of each particle d in n-dimensional space
$v_{dn}^{t+1} = w \times v_{dn}^{t} + \overbrace{c_1 \times r_1 \times (p_{dn} - x_{dn})}^{cognitive\ component} + \overbrace{c_2 \times r_2 \times (p_{gn} - x_{dn})}^{social\ component}$
Where t is current iteration and, x_{dn} is position of each particle d
r_1, r_2 are uniformly distributed random numbers in [0,1]
Update positions of each particle:
$x_{dn}^{t+1} = \begin{cases} 1 & if \quad r_3 < sigm(v_{dn}^{t+1}) \\ 0 & otherwise \end{cases}$
Where
$sigm(v_{dn}^{t+1}) = \frac{1}{1 + e^{-v_{dn}^{t+1}}}$
Convergence?
No
Yes
END

Figure 2: Flow chart of BPSO

In the present analysis, the division by the "number of observations determined as influential data points" of these statistics measured for a group of observations is called an objective function in the BPSO method to avoid masking and swamping impacts. With such a method the goal would be finding a collection of observation which optimize the objective LD (0 s) function for the limit. In other words, since the collection of observations determined as powerful observations by BPSO is excluded, the aim would be spotting the collection of observations which possess the largest influence on partial likelihood algorithms.

For the observations group for LD_{0_s}, there are several potential combinations and it takes burdensome calculation to decide all possible combination groups.

BPSO has consequently been utilized as an algorithm to detect the best collection of observations which impact partial probability function to eliminate this difficult calculation and to solve masking and swamping issues. By way of contrast, the BPSO-based theoretical technique would be aimed at excluding all points of data and that's not the case where it is able to assess important data points if the probability displacement statistics collected were not split by (m+1) for a set of observations. since our objective is to maximize and each data point appears to change the situation in likelihood displacement, the determination of non-influential data points as effective will be avoided and the right recognition maintaining important data points are to be given by splitting the likelihood displacement determined by removing a set of observations by (m+1). It is noted that since the group of 0s is the optimum collection of observations known as observations of influential type, the suggested objective function would obtain the maximum value. At the same time, masking and swamping results will not greatly impact it as they concurrently look for the optimum collection of insightful observations in the realm of search.

4.4.1.3 Definition of PSO Parameters

The maximum number of iterations, number of particles, coefficients of acceleration (c_1 and c_2), weight of inertia (w), and maximum velocity (v_{max}) are all tuning parameters that can enhance the PSO's inclusive efficiency. Clerc and Kennedy's analysis (2002) found that better convergence can be accomplished by deciding c_1, c_2 and w based on what is seen below:

$$\begin{cases} w = \frac{1}{\alpha-1+\sqrt{\alpha^2-2\alpha}} \\ c_1 = c_2 = \alpha w \end{cases} \tag{4.10}$$

Where $\alpha > 2$ and where that is to say, The tuning parameters are then dependent on the existence of the issue space. To put it another way, these parameters do not have a particular value that can be generalized to all optimization problems. The varying performance of the algorithm will result in changes to these parameters. Using the equations results as a quid (4.10) and the ones of an error and trial procedure within the analysis of simulation, the suggested BPSO-based approach fits very nicely together with the below group of parameters of tune type in Table 1, which provides the best success in determining the optimum set of influential observations.

Table 2: Tune parameters in the approach suggested in the logistic regression model to determine observations of influential nature on the basis of BPSO.

Tune parameters	Values checked in the BPSO method
Weight inertia (w)	0.9
Coefficients of acceleration, (c_1, c_2)	$c_1 = c_2 = 2$
v_{max}	4
Particulate number	the same as the number of particle positions, i.e. the number of observations in the dataset (n)
Number of iterations to the limit	100 for $\mathrm{n} = 100, 150$ for $n = 150, 200$ for $n = 200$

4.5 Simulation Design

In this part, a simulation study of Monte Carlo under different contamination rates and sample sizes has been formulated to state the performance of BPSO based proposed method and to compare this method with the traditional diagnostic techniques; likelihood displacement (LD), Cook's Distance (CD), DFBETAs. Simulation study has been performed for 2 independent variables. The regression coefficients values in the model are fixed to be $\beta_0 = 1,\ \beta_1 = 2,\ \beta_2 = 2$. The independent variables values (X) in the logistic regression model were generated from the N(0,1) normal distribution and the error terms were derived from the logistic distribution $\varepsilon i \sim \Lambda(0,1)$ with different sample sizes of 100, 150 and 200 in the data sets for the simulation. Response variable values were generated by obtaining probability values

$$\pi(x) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \qquad (4.11)$$

and satisfying the following equation [69], [70]

$$y_i = \begin{cases} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i < 0\ , & 0 \\ \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i \geq 0\ , & 1 \end{cases} \qquad i = 1,2,\ldots,n \qquad (4.12)$$

We have contaminated each simulated dataset at 4%, 6% and 10% rates for each sample size. It is ensured that the value set to be contaminated according to the contamination rates is randomly generated in a different order in each generated sample. The covariates to be contaminated were determined as $+4\sigma$ for mildly unusual observations and $+7\sigma$ for extreme unusual observations. The y values correlating to the contaminated observations were altered from 1 to 0 in the simulation conducted for influential observations [66]. Simulation study were repeated 100 times. The parameters have been estimated through eliminating the observations' collection identified for the suggested BPSO-based system; LD, CD, DFBETAs and for the authentic logistic model through the fitness of all points of data (including observations of influential type). For each simulation scenario, we repeated data creation and model fitting 100 times, recording the mean of each assessment criterion. These are recorded as mean square errors of parameter estimates (MSE), mean of sensitivity (Sens), area under the ROC curve (AUC) and Nagelkerke's R^2. As well as this, the mean of masking percentage as a proportion of undetected real influential observations (MP) and the mean of swamping percentage as a rare of non-influential observations detected as influential observations (SP) have also been recorded [66], [71].

CHAPTER 5

RESULT AND CONCLUSION

5.1 Result of Simulation

On simulated datasets created employing the methods outlined in the preceding section, the suggested BPSO-based technique for identifying influential observations in the logistic regression model was compared against standard diagnostic approaches based on specified assessment criteria.

Table 3

The masking (MP) and swamping (SP) proportions of the proposed BPSO-based technique and standard diagnostic methods among 100 simulations for different contamination rates (r_{cont}) and sample size (n).

n	r_{cont}	MP and SP	BPSO-based approach	CD	LD	DFBETAs
100	0.04	MP	0.073	0.749	0.578	0.333
		SP	0.000	0.000	0.015	0.015
	0.06	MP	0.124	0.746	0.652	0.451
		SP	0.000	0.000	0.025	0.008
	0.1	MP	0.101	0.828	0.725	0.497
		SP	0.003	0.000	0.056	0.005
150	0.04	MP	0.096	0.762	0.664	0.396
		SP	0.000	0.000	0.025	0.015
	0.06	MP	0.105	0.900	0.752	0.485
		SP	0.000	0.000	0.017	0.006
	0.1	MP	0.196	0.852	0.814	0.521
		SP	0.002	0.000	0.020	0.021
200	0.04	MP	0.102	0.732	0.523	0.415
		SP	0.000	0.001	0.019	0.012
	0.06	MP	0.099	0.930	0.721	0.489
		SP	0.000	0.010	0.100	0.004
	0.1	MP	0.138	0.861	0.852	0.638
		SP	0.007	0.008	0.109	0.007

After the application of the BPSO-based approach to the generated artificial data sets for the determination of influential observations, we found that the masking probability was in general close to a far extent to 0 and the swamping probability was on the verge of zero in all scenarios, demonstrating that the suggested identification procedure has decreased masking and swamping proportions, as seen in Table 3. Despite the fact that the SP for the CD method was zero in almost all situations, the MP for the CD method was the highest in every simulated dataset scenario. We know that masking is a more significant issue than swamping in outlier identification (Zhang et al. 2016). In addition, as shown in Table 3 in each simulated scenario, the LD and DFBETAs methods did not have a low masking and swamping proportions (especially masking proportion) like the one of BPSO-based strategy. In order to highlight the effect of influential observations on the model outcomes through different assessment criteria, we first developed a logistic regression model using all observations on simulated datasets. And then, to demonstrate the impact of the observations found by the suggested BPSO-based technique and standard diagnostic procedures, we built a logistic regression model by eliminating the observations determined by the suggested BPSO-based technique and standard diagnostic techniques (results seen in Table 4 for MSE, sensitivity, AUC and R^2).

We can see that the MSE in each simulation scenario for the original model constructed with all observations was the highest, but MSE was the lowest when we built a logistic regression model by eliminating the data detected in each scenario using the BPSO-based technique. After the model was created with no observed data using standard diagnostic methods, MSE decreased marginally, but not as much as the proposed strategy. Each simulation scenario containing all data points, it is clear that the other evaluation criteria Nagelkerke's R^2 for the original model were the smallest. But when we built the logistic regression model without the observations detected by BPSO based proposed technique, R^2 was the highest. It was discovered that R^2 for the model that is constructed by omitting observations detected by standard diagnostic methods, has improved based on the original model along with all the data, although not by the same amount as the suggested technique.

Table 4: MSE, Sensitivity, AUC and Nagelkerke's R^2 which are evaluated for the Logistic regression model by eliminating the set of data points that CD, LD, DFBETAs, BPSO-based technique determined, respectively and for the original logistic regression model by constructing with all observations for differnet contamination rates (r_{cont}) and sample size (n).

n	**r_{cont}**	**Criteria**	**BPSO-based**	**Original**	**LD**	**CD**	**DFBETAs**
100	0.04	MSE	0.789	9.122	3.712	5.633	2.544
		Sens	0.854	0.541	0.785	0.741	0.784
		AUC	0.899	0.652	0.754	0.702	0.801
		R^2	0.785	0.362	0.578	0.498	0.611
	0.06	MSE	0.965	12.986	4.325	6.234	3.698
		Sens	0.841	0.448	0.698	0.654	0.722
		AUC	0.836	0.487	0.685	0.590	0.748
		R^2	0.754	0.301	0.531	0.444	0.621
	0.1	MSE	1.100	14.784	6.578	7.451	3.998
		Sens	0.803	0.450	0.657	0.607	0.706
		AUC	0.805	0.424	0.647	0.555	0.699
		R^2	0.719	0.287	0.506	0.524	0.587
150	0.04	MSE	0.955	12.641	4.639	7.744	3.512
		Sens	0.849	0.503	0.695	0.681	0.722
		AUC	0.854	0.539	0.648	0.632	0.784
		R^2	0.773	0.349	0.581	0.457	0.625
	0.06	MSE	1.237	16.360	6.455	8.784	5.178
		Sens	0.820	0.400	0.632	0.601	0.694
		AUC	0.815	0.396	0.619	0.517	0.687
		R^2	0.715	0.287	0.465	0.406	0.578
	0.1	MSE	1.321	17.455	7.779	9.632	4.897
		Sens	0.779	0.319	0.549	0.530	0.631
		AUC	0.746	0.380	0.584	0.432	0.617
		R^2	0.703	0.251	0.451	0.420	0.497
200	0.04	MSE	1.103	15.655	6.471	9.854	5.735
		Sens	0.806	0.473	0.608	0.574	0.685
		AUC	0.812	0.487	0.533	0.547	0.709
		R^2	0.779	0.305	0.520	0.403	0.575
	0.06	MSE	1.741	18.458	8.743	10.850	5.897
		Sens	0.782	0.354	0.584	0.521	0.634
		AUC	0.766	0.329	0.599	0.472	0.614
		R^2	0.673	0.247	0.425	0.357	0.509
	0.1	MSE	2.201	20.744	9.872	11.667	5.478
		Sens	0.715	0.257	0.461	0.489	0.574
		AUC	0.707	0.307	0.540	0.382	0.524
		R^2	0.695	0.204	0.405	0.365	0.443

Moreover, the sensitivity and AUC of the original model together with all observation points in each scenario simulated were lowest, but is increased better than other diagnostic techniques when we built the logistic regression model by omitting the observation determined by the BPSO-based strategy in each scenario.

All of these findings show that the suggested BPSO-based technique works well and is quite suitable alternative technique for determining influential observations, according to each evaluation measure. To put it another way, the proposed technique offers a reliable and efficient approach to this problem. In addition, these results show that the suggested data generation strategy for data pollution is valid as well in the simulation study, because all the negative impacts of the results affecting the model were revealed. The reason is the model results were rather poor based on each evaluation criterion when the logistic regression model was created using all the observations in the simulated data sets. Furthermore, when examining the results of evaluation measures by BPSO-based and conventional techniques for each sample size and contamination rate, we discovered that additional modification in sample size and contamination rate disproportionately affected the performance of each conventional diagnostic methodology. Despite this, the BPSO-based technique was not influenced by the change in contamination rates and sample size, but rather achieved good results in the presence of high contamination and sample size.

5.2 Conclusion

Logistic regression model is commonly employed regression tool to infer relationship between binary response variable and independent variables using probability scores in many field of science. However, to obtain accurate results from the model, the models rely on some assumptions. One of these assumptions is that there should be no influential observations in the dataset. This is so because logistic model is not robust model and one single unusual data point is sufficient to affect model results unduly and cause misleading results from the model. Thus, correct determination of influential data points group is a quite significant step in the modeling duration. In logistic regression model, there are many diagnostic techniques for determination of influential observations. But these standard techniques are founded on single-step techniques and results acquired from these techniques are frequently misleading because of masking and swamping problems. Thus, using standard diagnostic methods is not useful for the determination of the optimal group of influential observations as a result of these problems.

Simultaneous techniques may solve masking and swamping problems because these techniques investigate optimal way simultaneously in search area. This work's main contribution is the idea for a novel simultaneous strategy to determine the optimal group of influential observations in the logistic regression model, based on BPSO. As shown in simulation studies, this proposed technique determines the best group of observations with a strong influence on both the partial likelihood function as well as parameter estimates, in addition to the predictive capabilities of the model. The suggested method has convincingly outperformed traditional techniques under different evaluation criteria in different simulated settings (MP, SP, MSE, R^2, sensitivity, AUC). Compared to standard methods, this strategy clearly eliminates issues of masking and swamping problems. In addition, in each simulation case, the BPSO-based technique gives the lowest MSE value and the highest values of R^2, sensitivity and AUC. The model estimates were not significantly altered if the logistic regression model was obtained by removing the observations indicated by these standard diagnostic techniques. It is very likely that the effects of masking and swamping caused this to happen. The logistic regression has nevertheless become more resilient when the logistic models were designed by the elimination of the data indicated using the suggested BPSO-based technique. The suggested BPSO-based technique for identifying important data in a logistic regression model was shown to be effective in a simulation studies. It's also worth noting that the recommended data creation approach for the data set simulated with influencing observations is correct, as we've seen all of the model's unwanted impacts. This is because the model outcomes were comparatively bad according to every evaluation criterion when the logistic regression model was created by all observations of the simulated data set. Overall, we feel that the proposed method for identifying influencing data in a logistic regression model will be of great use as evidenced by the simulation results.

REFERENCES

1) Hosmer, Lemeshow & Sturdivant, Applied logistic regression 2013
2) Kleinbaum and Klein (2002), logistic regression
3) D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 3rd Edition, John Wiley and Sons, Inc., USA, 2002.
4) David W. Hosmer & Stanley Lemeshow , Second Edition, Applied Logistic Regression.
5) A. Afifi, V. A. Clark and M. Susanne, *Computer-Aided Multivariate Analysis*, 4th Edition, Chapman & Hall/CRC, London, New York, 2004.
6) E. B. Atitwa, Socio-Economic Determination of Low Birth Weight in Kenya: An Application of Logistic Regression Model, *American Journal of Theoretical and Applied Statistics*, 4 (2015), no. 6, 438-445.
7) A. C. Scott, Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation, (2002).
8) D. Pregibon, Logistic regression diagnostics, Ann. Stat. 9 (1981), pp. 977–986.
9) S. V. Stehman, “Selecting and interpreting measures of thematic classification accuracy,” Remote Sensing of Environment, vol. 62, no. 1, pp. 77-89, 1997.
10) D. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation,” Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2011.
11) Cortes, C., and M. Mohri. 2004. AUC optimization vs. error rate minimization. In Advances in neural information processing systems, 313–20, England: The MIT Press.
12) Fawcett, T. 2006. An introduction to roc analysis. Pattern Recognition Letters 27 (8):861–74. doi: 10.1016/j.patrec.2005.10.010.
13) Songfeng Zheng (2017), Survival Analysis: A Modified Kaplan-Meir Estimator.
14) Deniz Inan and Nuriye Sancar (2020) particle swarm optimization based ridge logistic estimator, communications in Statistics – simulation and computation
15) Hosmer and Lemeshow (2000), applied logistic regression, second edition
16) Imon and Hadi, 2008

17) E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based Outlier: Algorithms and Applications," The International Journal of Very Large Databases, Vol. 8(3-4), 2000, pp. 237 – 253.

18) D. Pregibon, "Logistic Regression Diagnostics," Annals of Statistics, Vol. 9, 1981, pp. 977 – 986.

19) D. M. Hawkins, Identification of Outliers, London: Chapman and Hall, 1980.

20) P. J. Rousseeuw, and A. M. Leroy, Robust Regression and Outlier Detection. New York: John Wiley and Sons, 1987.

21) M. Breunig, H. P. Kriegel, R. Ng, and J. L. O. F. Sander, "Identifying Density-based Local Outliers," Proc. of the ACM SIGMOD, International Conference on Management of Data, New York: ACM Press, 2000, pp. 93 – 104.

22) C. C. Aggarwal, and P. S. Yu, "Outlier Detection for High Dimensional Data," Proc. of the 2001 ACM SIGMOD International Conference on Management of Data, ACM SIGMOD Record, Vol. 30(2), 2001, pp. 37 – 46.

23) V. J. Hodges, and J. Austin, "A Survey of Outlier Detection Methodologies," Artificial Intelligence Review, Vol. 22, 2004, pp. 85 – 126.

24) S. Sotoodeh, "Outlier Detection in Laser Scanner Point Clouds," Proc. of the IAPRS, Dresden, Vol. XXXVI/5, 2006, pp. 297 – 301.

25) AJ leone, M.Minutti-Meza, "Influential Observations and Inference in Accounting Research" 2019

26) COOK, R. D. (1979). Influential observations in linear regression. J. Amer. Statist. Assoc. 74 169–174. MR0529533

27) DRAPER, N. R. and SMITH, H. (1998). Applied Regression Analysis, 3rd ed. Wiley, New York. MR1614335

28) D. A. Belsley, E. Kuh, and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: John Wiley and Sons, 1980.

29) B. Scholkopf, and A. J. Smola. Learning with Kernels, 2002, MIT press, Cambridge, Massachusetts.

30) Nuriye Sancar & Deniz Inan (2020) Identification of influential observations based on binary particle swarm optimization in the cox PH model, Communications in Statistics - Simulation and Computation, 49:3, 567-590

31) G. Deliorman & D. Inan (2020): Binary particle swarm optimization as a detection tool for influential subsets in linear regression

32) Zahra Beheshti, Siti Mariyam Hj. Shamsuddin; a review of population-based meta-heuristic algorithm, march 2013

33) A. H. M. Rahmatullah Imon a & Ali S. Hadi b (2008), identification of multiple outliers in logistic regression

34) T.P. Ryan, Modern Regression Methods,Wiley, New York, 1997.

35) Chen and Liu(1993)

36) R.D. Cook, Detection of influential observations in linear regression, Technometrics 19 (1977), pp. 15–18.

37) R.D. Cook, Assessment of local influence, J. Roy. Stat. Soc., Ser-B 48 (1986), pp. 131–169.

38) D.J. Finney, The estimation from individual records of the relationship between dose and quantile response, Biometrika 34 (1947), pp. 320–334.

39) A.S. Hadi, A new measure of overall potential influence in linear regression, Comput. Stat. Data Anal. 14 (1992), pp. 1–27.

40) A.H.M.R. Imon, Identifying multiple influential observations in linear regression, J. Appl. Stat. 32 (2005), pp. 929–946.

41) WY Poon, YS Poon, (2002), total behavior of likelihood displacement

42) J. A. Momoh, R. Adapa, and M. E. El-Hawary, “A review of selected optimal power flow literature to 1993. I. Nonlinear and quadratic programming approaches,” IEEE Trans. Power Syst., vol. 14, no. 1, pp. 96–104, 1999.

43) J. Echer and M.Kupferschmid, Introduction to Operations Research. New York: Wiley, 1988.

44) Omar Saber Qasim and Zakariya Yahya Algamal, (2018), Feature selection using particle swarm optimization-based logistic regression model

45) J. Kennedy and R. Eberhart, “Particle swarm optimization,” in Proc. IEEE Int.

Conf. Neural Netw., 1995, vol. 4, pp. 1942–1948.

46) R. Eberhart and J. Kennedy, “A new optimizer using particle swarm theory,” in Proc. 6th Int. Symp. Micro Machine Human Science, 1995, pp. 39–43.

47) H. Xiaohui, S. Yuhui, and R. Eberhart, “Recent advances in particle swarm,” in Proc. Congr. Evol. Comput., 2004, vol. 1, pp. 90–97.

48) R. C. Eberhart and Y. Shi, “Guest editorial,” IEEE Trans. Evol. Comput. (Special Issue on Particle Swarm Optimization), vol. 8, no. 3, pp. 201–203, Jun. 2004.

49) Y. Shi and R. Eberhart, “A modified particle swarm optimizer,” in Proc. IEEE World Congr. Comput. Intell., 1998, pp. 69–73.

50) M. R. AlRashidi and M. E. El-Hawary, (2008), a survey of particle swarm optimization applications in electric power systems

51) J. Kennedy and R.C. Eberhart, A discrete binary version of the particle swarm algorithm, in 1997 IEEE international conference on systems, man, and cybernetics. Computational cybernetics and simulation, Vol. 5. IEEE, 1997, pp. 4104–4108.

52) R. Eberhart and J. Kennedy, Particle swarm optimization, in Proceedings of the IEEE international conference on neural networks, Vol. 4. Citeseer, 1995, pp. 1942–1948.

53) Bozorg-Haddad, O., M. Solgi, H. A. Lo~A. 2017. Meta-heuristic and evolutionary algorithms for engineering optimization, vol. 294. Hoboken, USA: John Wiley & Sons.

54) M. R. AlRashidi, Student Member, IEEE, and M. E. El-Hawary, Fellow, IEEE, (2008), A Survey of Particle Swarm Optimization Applications in Electric Power Systems

55) Zahra Beheshti, Siti Mariyam Hj. Shamsuddin, (2013), a review of population-based meta-heuristic algorithm

56) Kennedy, J., Mendes, R., “Population structure and particle swarm performance”, Proceedings of IEEE Congress on Evolutionary Computation, (2002), pp. 1671–1676.

57) Kennedy, J., Mendes, R., “Neighborhood topologies in fully informed and best-of-neighborhood particle swarms”, IEEE Transactions on Systems, Man, and

Cybernetics Part-C, Vol. 36, No. 4, (2006), pp. 515-519.

58) Liang, J. J., Suganthan, P. N., “Dynamic multi-swarm particle swarm optimizer”, Proceedings of Swarm Intelligence Symposium, (2005), pp. 124–129.

59) Angeline, P. J., “Using selection to improve particle swarm optimization”, Proceedings of the 1998 IEEE International Conference on Evolutionary Computation, (1998), pp. 84–89.

60) Juang, C. F., “A hybrid of genetic algorithm and particle swarm optimization for recurrent network design”, IEEE Transactions on Systems, Man, and Cybernetics Part-C, Vol. 34, No. 2, (2004), pp. 997–1006.

61) ZahraBeheshti, Siti Mariyam Hj.Shamsuddin and Shafaatunnur Hasan (2013), Median-oriented Particle Swarm Optimization

62) J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," Systems, Man, and Cybernetics, 1997.'Computational Cybernetics and Simulation'., 1997 IEEE International Conference on, vol. 5, 1997.

63) Kenji TANAKA, Takio KURITA, Tohru KAWABE (2007), Selection of Import Vectors via Binary Particle Swarm Optimization and Cross-Validation for Kernel Logistic Regression

64) Menard, Scott W. (2002). Applied Logistic Regression (2nd ed.).

65) Burcin Coskun & O. Alpu (2019), Diagnostics of multiple group influential observations for logistic regression models

66) A.A.M.Nurunnabi and A.H.M. Rahmatullah Imon and M. Nasser 2010, Identification of multiple influential observations in logistic regression

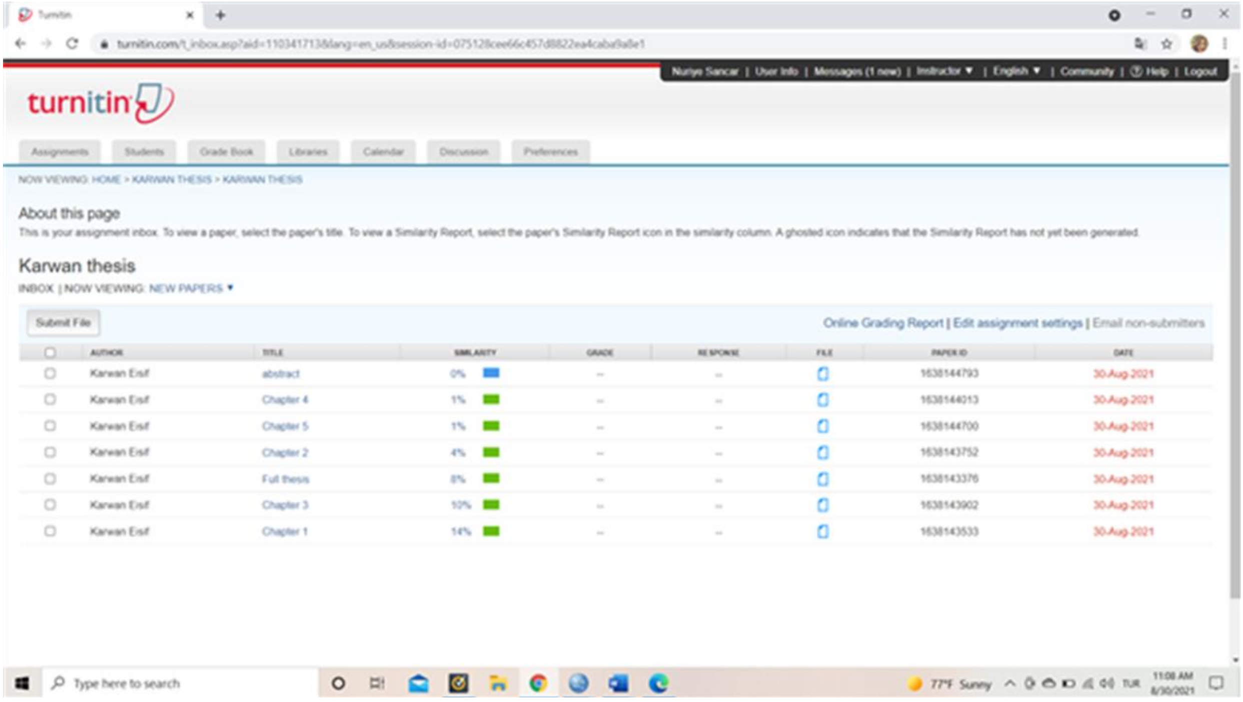
67) A.A.M. Nurunnabi, M. Nasser & A.H.M.R. Imon 2016, Identification and classification of multiple outliers, high leverage points and influential observations in linear regression.

68) Calazan et al. 2014, A hardware accelerator for particle swarm optimization.

69) Norazan vd., 2012, Determining the Significant Factors Affecting the Physical and Mental Components of Academicians Using Robust Linear Regression Models

70) Kordzakhia, Mishra ve Reiersolmoen, 2001

71) Norazan, M., Sanızah, A., Habshah, M., 2012, Identifying bad leverage points in logistic regression model based on robust deviance components, 62-67.


Turnitin
turnitin.com/t_inbox.asp?aid=110341713&lang=en_us&session-id=075128cee66c457d8822ea4caba9a8e1
Nuriye Sarıcar | User Info | Messages (1 new) | Instructor ▼ | English ▼ | Community | ⓘ Help | Logout
turnitin
Assignments
Students
Grade Book
Libraries
Calendar
Discussion
Preferences
Type here to search
77°F Sunny
TUR
11:08 AM
8/30/2021

NOW VIEWING: HOME > KARWAN THESIS > KARWAN THESIS

About this page

This is your assignment inbox. To view a paper, select the paper's title. To view a Similarity Report, select the paper's Similarity Report icon in the similarity column. A ghosted icon indicates that the Similarity Report has not yet been generated.

Karwan thesis

INBOX | NOW VIEWING: NEW PAPERS ▾

Submit File

Online Grading Report | Edit assignment settings | Email non-submitters

□	AUTHOR	TITLE	SIMILARITY	GRADE	RESPONSE	FILE	PAPER ID	DATE
□	Karwan Eisf	abstract	0%	--	--	▯	1638144793	30-Aug-2021
□	Karwan Eisf	Chapter 4	1%	--	--	▯	1638144013	30-Aug-2021
□	Karwan Eisf	Chapter 5	1%	--	--	▯	1638144700	30-Aug-2021
□	Karwan Eisf	Chapter 2	4%	--	--	▯	1638143752	30-Aug-2021
□	Karwan Eisf	Full thesis	8%	--	--	▯	1638143376	30-Aug-2021
□	Karwan Eisf	Chapter 3	10%	--	--	▯	1638143902	30-Aug-2021
□	Karwan Eisf	Chapter 1	14%	--	--	▯	1638143533	30-Aug-2021