**British Journal of Mathematics & Computer Science**
**10(1): 1-16, 2015, Article no.BJMCS.18616**
***ISSN: 2231-0851***

**SCIENCEDOMAIN** *international*
www.sciencedomain.org

# Cluster Analysis as a Strategy of Grouping to Construct Goodness-of-Fit Tests when the Continuous Covariates Present in the Logistic Regression Model

**Jassim N. Hussain[1*] and Atheer J. Nassir[2]**

*[1]Department of Statistics, Faculty of Administration and Economics, University of Karbala, Karbala, Iraq.*
*[2]Department of Clinical Pharmacy, Faculty of Pharmacy, Pharmacy Building (A15), Compedown Campus, University of Sydney, Sydney, NSW, 2006, Australia.*

***Article Information***

DOI: 10.9734/BJMCS/2015/18616
Editor(s):
(1) Vyacheslav Pivovarchik, Department of Applied Mathematics and Computer Science, South-Ukrainian National Pedagogical University, Ukraine.
Reviewers:
(1) Katarzyna Rostek, Faculty of Management, Warsaw University of Technology, Poland.
(2) Anonymous, Universidade Federal de Itajuba, Brazil.
(3) Behnam Sharif, University of Calgary, Canada.
(4) Kikawa richard cliff, Dept of Maths and Stats, Tshwane University of Technology, South Africa.
Complete Peer review History: http://sciencedomain.org/review-history/9908

***Original Research Article***

*Received: 01 May 2015*
*Accepted: 06 June 2015*
*Published: 20 June 2015*

## Abstract

When continuous covariates are present, classical Pearson and deviance goodness-of-fit tests to assess logistic model fit break down. Many goodness-of-fit (GOF) tests such as Hosmer–Lemeshow tests can be used in these situations. Meanwhile, it is simple to perform and widely used, it does not have desirable power in many cases and provides no further information on the source of any detectable lack-of-fit. We propose a new strategy of grouping based on a very general partitional clustering in the covariate space to construct two goodness-of-fit test statistics. Many simulation studies are implemented and clinical data set is analyzed to examine the performance of the proposed strategy of grouping and the developed GOF test statistics. The results show that the proposed strategy of grouping and GOF test statistics based on it has a potential for use in practice as a recommended strategy of grouping and as GOF test statistics to assess the adequacy of the logistic regression model.

*Keywords: Continuous covariates; cluster analysis; goodness-of-fit test; logistic regression; strategy of grouping.*

**Corresponding author: j_nassir2000@yahoo.com;*

# 1 Introduction

Nowadays, logistic regression model (LRM) is part of the standard empirical research and it is commonly employed in several disciplines including medical research, health research, and social research [1]. Overall goodness-of-fit (GOF) test for the LRM is considered as the principal activity in the modeling process. GOF test reflects whether the predicted values are an accurate representation of the observed values. GOF refers also to the adequacy of the fitted model. It is defined as an evaluation of how well model predicted outcomes agree with the observed data [2]. GOF test, on the other hand, widely referred to as lack-of-fit, because it is measuring how far the model is from the data; more than how much the model is good [3]. Omitted predictors, a misspecified form of the predictor, or an inappropriate link function can all result in poor fitting.

Assessing GOF for the LRM is widely studied and many strategies of grouping and GOF test statistics based on them have been proposed in the last three decades. Two strategies of grouping were proposed by Hosmer and Lemeshow (HL) [4-5], deciles of risk and the fixed cut-off points respectively, to group the range of the estimated dependent variable. In order to construct GOF test, the ranked estimated probabilities are grouped into $K$ groups according to either deciles of risk or prespecified fixed cut-off points. The test statistic is calculated by comparing the observed frequency ($O_k$) to the average predicted frequency ($E_k$) in the $k^{th}$ group, $k = 1, 2, \ldots, K$, via the familiar form of the statistic:

$$\hat{C}_{HL} = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} \quad (1)$$

HL's GOF test statistic is widely used due to its following properties: **(*a*)** it is naturally attractive and easy to compute; **(*b*)** it has sound support from simulation studies; **(*c*)** it is widely available in computer packages; and **(*d*)** In addition to these properties, lack of a better approach also contributes to its popularity. However, it has the following limitations [6-8]: **(*a*)** its limiting distribution has not been carefully derived; **(*b*)** it is a conservative test and has low power to detect specific types of lack-of-fit (such as nonlinearity in an explanatory variable); **(*c*)** it is highly dependent on how the observations are grouped; **(*d*)** if too few groups are used to calculate the statistic (for instance, five or fewer groups), it will almost always indicate that the model fits the data; and **(*e*)** when the HL GOF statistic indicates a lack of fit, it may be difficult to identify what types of subjects are not modeled well.

Many other researchers proposed different strategies based on HL's strategies of grouping such as [9] proposed a strategy of grouping based on assigning a score to the dependent variable categories. A chi-squared type GOF test was proposed by Zhang [10] for LRM using the fixed cut-off points strategy of grouping and the case-control data. The deciles of risk strategy of grouping was used by [11] to construct the GOF test when the correlated or grouped binary data are analyzed by using the logistic generalized estimating equations (GEE) model. Also, the fixed cut-off points strategy of grouping was adapted by [12] to propose a data-driven strategy for grouping data and a new chi-square type GOF test statistic based on case-control data for testing LRM by adapting the Zhang [10] test.

Tsiatis [13] proposed different approach to partition the multidimensional space of covariates into $K$ distinct groups, instead of grouping the observations by their predicted outcomes. An additive group effect for each group is added to the model to measure grouping lack-of-fit. A score statistic is used to test that all of the $K$ grouping effects are zero. Tsiatis' procedure is as follows: **(*a*)** the space of covariates matrix $(X_1, X_2, \ldots, X_P)'$ is partitioned into $K$ distinct groups in $P$-dimensional space denoted by $C_1, C_2, \ldots, C_K$. The indicator functions $I^{(k)}$ $(k = 1, 2, \ldots, K)$ are defined by $I^{(k)} = 1$ if $(X_1, X_2, \ldots, X_P)' \in C_k$ and $I^{(k)} = 0$ otherwise; **(*b*)** the model considered is:

$$ln(\pi_i/(1-\pi_i)) = \boldsymbol{\beta}'\boldsymbol{X}_i + \boldsymbol{\gamma}'\boldsymbol{I}_i \quad (2)$$

where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_P)$, $\boldsymbol{X}_i' = (1, x_{1i}, \ldots, x_{Pi})$, $\boldsymbol{I}_i' = (I_i^{(1)}, \ldots, I_i^{(K)})$, and $\boldsymbol{\gamma}' = (\gamma_1, \gamma_2, \ldots, \gamma_K)$.

Note that $\boldsymbol{\beta}'\boldsymbol{X}_i$ models all the original covariates and $\boldsymbol{\gamma}'\boldsymbol{I}_i$ models the regional shifts; **(*c*)** a score statistic is then constructed to test that $H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_K = 0$. Tsiatis's approach is conceptually elegant, but its lack a general rule for how to partition the covariate space, especially when continuous covariates are present. How to choose the number of distinct groups *K* has also remained largely unstudied. Many researchers proposed other strategies based on the idea of Tsiatis's strategy of grouping such as [14] proposed strategy of grouping based on categories of the covariates and constructed GOF test statistics based on this strategy to assess the adequacy of the multinomial LRM.

The deficiencies of the above strategies and GOF tests based on them motivated Pulkstenis, and Robinson [2] to propose a two-stage modification of the HL's strategy of grouping where in the first stage all dependents are sorted by model-based estimated probabilities within each unique covariate pattern, as defined by only the categorical covariates, and then in the second stage creating two sub categories within each covariate pattern, essentially splitting the category in two cells based on the median of estimated probabilities $\hat{\pi}_i$ within each of the *k* row. This additional stratification basically doubles the number of covariate patterns to incorporate information related to all continuous covariates in the model. Model-based expected counts $E_{khj}$ are computed exactly as before, and the proposed test statistics are a Pearson chi-square test given by:

$$X_{PR}^2 = \sum_{k=1}^{K} \sum_{h=1}^{2} \sum_{j=1}^{2} \frac{\left(O_{khj} - E_{khj}\right)^2}{E_{khj}} \tag{3}$$

and the deviance which is calculated by comparing the expected and the observed counts in the resulting contingency table as:

$$D_{PR}^2 = 2 \sum_{k=1}^{K} \sum_{h=1}^{2} \sum_{j=1}^{2} O_{khj} log\left(\frac{O_{khj}}{E_{khj}}\right) \tag{4}$$

where *k* denotes covariate patterns, *h* denotes the sub stratification due to ordering by fitted probabilities, and *j* denotes the categories of the dependent variable. The degrees of freedom for these statistics are given by $2K - p - 2$, where $2K$ refers to the number of rows in the new stratification splitting each row of contingency table and *p* is the number of categorical variables in the model. These GOF test statistics may be more powerful than the HL GOF test in some situations due to the fact that the structure of individual covariate patterns is kept intact rather than collapsed. The main limitations of these GOF test statistics are not recommended by authors when only continuous variables are modeled and HL GOF test would be preferable, or when only categorical variables are modeled and the standard Pearson or deviance chi-square would be appropriate.

All the reviewed works above are based on the subjective grouping to either the space of the estimated response variable or the space of the covariates. Recently, Xie et al. [15] proposed to use hierarchical clustering analysis (specifically Ward's Method) to group the space of the covariates into clusters of similarity. This has the advantage of identifying groups in which the observations are similarly profiled with respect to their covariate values. They assumed that $(x_{i1}, x_{i2}, \cdots, x_{iP})$ be the set of *P* covariate values for the $i^{th}$ observation, $i = 1,2, \cdots, n$. They proposed to use Ward's method of clustering to partition the space of the covariates into *K* groups, denoted by $C_1, C_2, \cdots C_K$, and they defined an indicator function for the $k^{th}$ group by $I^{(k)} = 1$ if $(x_{i1}, \cdots, x_{iP})' \in C_k$, and 0 otherwise. Consequently, they proposed two GOF tests to assess the adequacy of LRM based on this strategy of grouping. The first GOF test is a Pearson chi-square statistic ( $X_{P^*}^2$ ) similar to the HL GOF test to assess the lack-of-fit of LRM, given by:

$$X_{P^*}^2 = \sum_{k=1}^{K} (O_k - n_k \bar{\pi}_k)^2 / n_k \bar{\pi}_k (1 - \bar{\pi}_k) \tag{5}$$

where $n_k$ is the number of observations in $C_k$, $O_k$ is the observed number of events/successes in $C_k$, and $\bar{\pi}_k = \sum_{g=1}^{g_k} m_g \hat{\pi}_g / n_k$ is the average estimated probability in $C_k$, which has $g_k$ covariate patterns with $m_g$ observations in the $g^{th}$ covariate pattern. For abbreviation the second proposed GOF test is exactly same as

Tsiatis's GOF test which is discussed above. They considered the distribution of the proposed GOF test closes to chi-square distribution with $df = K - (P/2) - 1$ based on several assumptions about the true distribution of the observations and the parameter estimators.

Xie, et al. [15] pointed out that: **(*a*)** both GOF tests for decisions on model fit in their study are applicable to a wide range of model scenarios (data settings) when continuous covariates are present. **(*b*)** for all the simulation scenarios, the GOF tests show reasonably well-controlled type I error rate. **(*c*)** both GOF tests demonstrate at least equal, if not much higher power in detecting missing quadratic term or missing interaction term, as compared to currently widely used HL GOF test. **(*d*)** when both continuous and categorical covariates are present, the two GOF tests along with the HL GOF test outperform the Pulkstenis and Robinson tests in detecting missing quadratic term. But the Pulkstenis and Robinson tests appear to outperform the two proposed tests and the HL GOF test in detecting missing interaction term. **(*e*)** these comparisons provide an example in practice that both GOF tests demonstrate better properties.

In spite of these good properties of the strategy of grouping and GOF tests proposed by Xie, et al. [15] they have some disadvantages such as: **(*a*)** hierarchical clustering has a general disadvantage since it contains no condition for reallocation of elements which may have poorly classified at an early stage in the analysis [16]. **(*b*)** the output of this method of clustering necessarily represent hierarchical relationship among the elements, thus this method of clustering does not appropriate to handling large data sets. **(*c*)** when both continuous and categorical covariates are present, the proposed strategy of Xie, et al. [15] treated the dichotomous and ordinal covariates as interval data. **(*d*)** as it was known that the Ward's method is an automatic method in clustering the observation started with $n$ clusters and end with one cluster or in reverse direction, Xie et al. [15] did not show how are they used the number of clusters $K$. **(*e*)** they recommended more studies for determining the degrees of freedom for the approximated Pearson chi-square statistic and for improving the clustering process by choosing another clustering method better than the method is used.

All these strategies of grouping and GOF test statistics based on them have their limitations. These limitations of currently available strategies of grouping and GOF test statistics based on them are motivated us to propose a new strategy of grouping based on partitional clustering and constructing two GOF test statistics based on the proposed strategy to assess the adequacy of LRM. Consequently, the main objectives of this study are proposing a new strategy of grouping based on partitional cluster analysis, constructing two GOF test statistics of chi-square type (clustering chi-square $X_c^2$ and clustering chi-square deviance $D_c^2$) based on the proposed strategy of grouping to assess the adequacy of the LRM and evaluating the performance of the proposed strategy of grouping and the developed GOF tests.

The paper is organized as follows. Proposed strategy of grouping and GOF test statistics are given in Section 2. Thereafter in Section 3, we present some results of analyzing the simulation studies to evaluate the performance of the proposed strategy of grouping and the developed GOF test statistics. An application using real data set from the survival times of patients with prostate cancer who are randomly allocated to a treatment with an estrogen is presented in Section 4. Brief conclusions are presented in the last Section.

# 2 Proposed Strategy of Grouping

The proposed strategy of grouping is based on cluster analysis. The term "Cluster Analysis" consists of two main branches of clustering; Hierarchical and Partitional. Partitional clustering involves partitioning a given set of quantitative individuals or elements into a number of distinct groups, called clusters [17]. Also, [18] defined Partitional clustering as the process of organizing the quantitative elements in a dataset into clusters or groups so that the elements within the same cluster have a high degree of similarity, while the elements belonging to different clusters have a high degree of dissimilarity according to some defined similarity criteria. We construct GOF statistics based on the strategy of grouping combine the ideas of the partitional clustering to partition the quantitative covariates spaces in the first stage and to partition the categories of the observed dependent variable space in the second stage of the proposed strategy of grouping. The proposed strategy of grouping combined the ideas of partitional cluster technique ($K$-Means) [19] to partition the

spaces of the continuous covariates and the categories of the dependent variable to partition each group into two subgroups. The proposed strategy of grouping based on the mixed partitioning to divide the range of the covariates and the range of the observed and the estimated dependent variables into clusters of similarity. This enables us to avoid the subjective partitioning disadvantages, and has the advantage of identifying clusters which have the same observations with respect to their covariate and dependent variable values.

Hence, suppose **X** is a matrix of $n \times P$ covariates and **Y** is a vector of binary dependent variable in the data set with two categories (0 or 1). $X_j = \left(x_{1j}, x_{2j}, \cdots, x_{nj}\right)'$, $j = 1, 2, \cdots, P$ and $i = 1, 2, \cdots, n$ is the $j^{th}$ covariate in the matrix. The proposed strategy of grouping attempts to partition the elements of the covariates in the matrix **X** of the covariates into $K$ distinct clusters denoted by $C_1, C_2, \cdots, C_K$ as follows:

## 2.1 Select the Number of Clusters

The first step is selecting the number of clusters (groups) because the proposed strategy of clustering is affected by the fundamental and unresolved problem in cluster analysis. The problem is how many clusters are present in a given data set? This problem was widely studied. Several methods are proposed to identify the number of clusters statistically from a given data set which includes methods such as the gap statistic [20]; a simulated annealing clustering based method [21]; cluster isolation criterion [22]; cluster stability [23] and Rand's statistic [24]. All these methods are based on prior knowledge about the data without any constraint on this number; this means the number of clusters can be any number. Thus, there is no completely satisfactory method for determining the number of clusters for any type of cluster analysis.

These methods are not suitable to use in the strategy of grouping because (***a***) choosing any number of clusters is irrelevant in constructing the GOF tests and (***b***) there are some constraints on the number of clusters (groups) when the chi-square type GOF tests are constructed. For example [25] stated that the smallest number of clusters $K$ must be greater than or equal to 6 to ensure that 80% of the expected frequencies are greater than 5. Meanwhile [26] did not specify any number of clusters but instead he stated that *"the important point is that K should be larger when n is large. But it is not recommended that one use a very large value of K, and a choice in the range 5-15 seems right"*. Therefore, we will compare the results of multiple runs with different $K$ number of clusters in the range 5-15 and choose the best one according to the above criterion (expected frequencies) and other two criteria (CC and $R^2$) will discuss later.

## 2.2 Partition the Space of Covariates

In the second step we partition the space of covariates into $K$ distinct clusters $C_1, C_2, \cdots, C_K$ using the idea of $K$-means cluster analysis [17-27]. The clustering is done on the basis of distance measures and according to the following basic requirements of the clustering process [28]:

$$\left.\begin{array}{ll} a)\ C_k \neq \emptyset, & k = 1,2, \cdots K \\ b)\ C_k \cap C_l = \emptyset, & k, l = 1,2, \ldots, K \text{ and } k \neq l \\ C)\ \boldsymbol{X} = \bigcup_{k=1}^{K} C_k & \end{array}\right\} \quad (6)$$

then the association function $I(\mathbf{x}_i, C_k)$ is defined as follows:

$$I(\mathbf{x}_i, C_k) = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ is allocated to } k^{th} \text{ cluster} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where $\mathbf{x}_i$ is the $i^{th}$ vector of covariates and $C_k$ is the $k^{th}$ cluster. Then, the clustered covariates in this stage are used to fit the appropriate LRM to estimate the probabilities $\hat{\pi}_i$, for more details of how to estimate the probabilities see [25,29].

### 2.3 Cross-classify the Observed Dependent Variable $Y_i$ and the Estimated Probabilities $\hat{\pi}_i$

In the last step of the strategy of clustering, a cross-classification table for the observed dependent variable $Y_i$ and the estimated probabilities $\hat{\pi}_i$ is created by partitioning each cluster into two parts based on the categories of the observed dependent variable $Y_i$ (0 and 1). Each observation then belongs to one of $2K$ distinct clusters as in Table 1.

Three criteria are used to assess the efficiency of the proposed strategy of grouping. The first criterion is the frequency in each cluster (group) where [25] stated that it must be insure that 80% of the expected frequencies are greater than 5. The second criterion is the Clustering Criterion (CC) which is equal to the mean of the squared error or the mean of the distance measurement between the elements of data set and their cluster centers given by:

$$CC = \frac{1}{K}\sum_{k=1}^{K}\sum_{\mathbf{x}_i \in C_k} I(\mathbf{x}_i, C_k)\|\mathbf{x}_i - \boldsymbol{c}_k\|^2 \tag{8}$$

where $\|.\|$ is the Euclidean norm, $\mathbf{x}_i$ is the $i^{th}$ vector of covariates, $C_k$ is the $k^{th}$ cluster, $\boldsymbol{c}_k$ is the centriod of the $k^{th}$ cluster $I(\mathbf{x}_i, C_k)$ is as defined in Eq. (7) and $K$ is the number of clusters (groups). The small value of the $CC$ means the clustering process was efficient. The third criterion is the dissimilarity criterion. The clustering process aims to produce disjoint or dissimilar clusters. Consequently, the third criterion is used to measure the dissimilarity between clusters is [20].

$$R^2 = (TSS - SSE)/TSS \tag{9}$$

where $TSS = \sum_j \sum_i \left(x_{ij} - \bar{c}_j\right)^2$ is the total sum squares which are resulting from comparing $x_{ij}$ the individual observations for each variable against $\bar{c}_j = \sum_k \tilde{c}_{kj}/K$ the grand mean for the $j^{th}$ variable and $SSE = \sum_k \sum_{i \in R_k} \sum_j \left(x_{ikj} - \tilde{c}_{kj}\right)^2$ is the sum squared error resulting from comparing $x_{ikj}$ the $i^{th}$ element in the $j^{th}$ variable of the $k^{th}$ group against the $k^{th}$ group mean $\tilde{c}_{kj}$ for the $j^{th}$ variable. The value of $R^2$ is interpreted as a measurement of dissimilarity between clusters. Therefore, we have a clustering process gives disjoint or dissimilar clusters when we have large value of $R^2$.

## 3 Construct the Proposed GOF Test Statistics

The proposed strategy of clustering is an automatic process that often yields clusters (groups) with unequal sizes, where $n_1 \neq n_2 \neq n_3 \neq \cdots \neq n_K$; here $n_k$ is the number of observations in the $k^{th}$ cluster. Then, the assumption of the groups with equal size is avoided. Consequently, the importance of large clusters must be considered by calculating the proportional weight of each cluster and using this weight to calculate the weighted observed and weighted estimated frequencies for each cell in Table 1. Therefore, the weighted observed frequency $O_{gk}$ of $g^{th}$ category of dependent variable in $k^{th}$ cluster is calculated as:

$$O_{gk} = w_k \sum_{i=1}^{n_{gk}} Z_{igk} Y_{igk} \tag{10}$$

where $Y_{igk}$ is the $i^{th}$ observation of the dependent variable in the $g^{th}$ category and in $k^{th}$ cluster; $n_{gk}$ is the number of the observations into $g^{th}$ category of dependent variable and $k^{th}$ cluster; $Z_{igk} = 1$ if $Y_i$ belong to $\mathrm{g^{th}}$ category in $\mathrm{k^{th}}$ cluster and 0 otherwise and $w_k = n_k/n$ is the proportional weight of $k^{th}$ cluster.

Meanwhile, the weighted expected frequency $E_{gk}$ of $g^{th}$ category in $k^{th}$ cluster is calculated as follows:

$$E_{gk} = w_k \sum_{i=1}^{n_{gk}} Z_{igk} \hat{\pi}_{igk} \tag{11}$$

where $Z_{igk} = 1$ if $\hat{\pi}_i$ belong to $g^{th}$ category in $k^{th}$ cluster and 0 otherwise, $\hat{\pi}_{igk}$ is the $i^{th}$ observation of the estimated probability in the $g^{th}$ category and in $k^{th}$ cluster, and $n_{gk}$ is the number of observations of the estimated probabilities in the $g^{th}$ category of the dependent variable and in $k^{th}$ cluster. After that, a cross-classification table consists of the weighted observed frequencies $O_{gk}$ which is calculated as in Eq. (10) and the weighted expected frequencies $E_{gk}$ which is calculated as in Eq. (11) is created as in Table 1.

**Table 1. Cross-classify the observed and the estimated dependent variables**

| Clusters categories of $Y_i$ | $C_1$ | $C_2$ | … | $C_K$ |
|---|---|---|---|---|
| 0 | $O_{01}$ | $O_{02}$ | … | $O_{0K}$ |
| | $E_{01}$ | $E_{02}$ | … | $E_{0K}$ |
| 1 | $O_{11}$ | $O_{12}$ | … | $O_{1K}$ |
| | $E_{11}$ | $E_{12}$ | … | $E_{1K}$ |

The chi-square type GOF test statistics are constructed by comparing the weighted observed and the weighted expected frequencies in $g^{th}$ category of dependent variable in $k^{th}$ cluster. Hence, the first proposed GOF test statistic is the clustered chi-square $X_c^2$ which is calculated as follows:

$$X_c^2 = \sum_{k=1}^{K} \sum_{g=1}^{2} \left(O_{gk} - E_{gk}\right)^2 / E_{gk} \tag{12}$$

and the second proposed GOF test statistic is the clustered deviance $D_c^2$ is given:

$$D_c^2 = 2 \sum_{k=1}^{K} \sum_{g=1}^{2} O_{gk}\, log\left(O_{gk}/E_{gk}\right) \tag{13}$$

The proposed GOF test statistics $X_c^2$ and $D_c^2$ are designed to allow us to assess the GOF of the LRM. They differ from the existing GOF test statistics in their construction based on efficient strategy of grouping and weighted observed and expected frequencies for each cluster. Hence, the criteria discussed above are used to investigate the performance of the proposed strategy of grouping as in simulation study 1, and the other simulation studies are used to investigate the performance of the proposed GOF test statistics. Also, clinical dataset is analyzed to investigate the performance of the proposed GOF tests as shown in the next Sections.

# 4 Simulation Results and Discussion

Extensive simulation studies are conducted to assess the performance of the proposed strategy of grouping and to investigate the properties of the proposed GOF tests. The objectives are to assess the adequacy of the proposed strategy of grouping to give disjoint groups, to assess the power of the GOF test statistics to detect a variety of departures from the LRM and to compare the performance of the proposed GOF test with some of existing GOF test (when appropriate). Both of the proposed GOF test statistics are based on the chi-square approach with additional stratification; it follows that both tests should follow asymptotically chi-square distribution with $2K - P - 2$ degrees of freedom, where $K$ is the number of clusters (groups) and $P$ is the number of parameters in the model.

In all simulation studies, a given model is assumed for both the true underlying linear predictor and the joint distribution of the covariates. Random sample of $n$ vectors of uncorrelated covariates in each replication was generated, and $\mathbf{x}_i'\boldsymbol{\beta}$ computed for each independent sampling unit. The response $Y$ was determined by comparing the underlying model $\pi(\mathbf{x}_i) = exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)/\left(1 + exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)$ to a uniform random variable $\left(u_i \sim U(0,1)\right)$, assigning $y_i = 1$ if $u_i \leq \pi(\mathbf{x}_i)$ and 0 otherwise, $i = 1, 2, \ldots, n$. This process gives a matrix $\mathbf{X}$ of $n \times P$ covariates and Y is a vector of binary dependent variable in the data set with two categories (0 or 1). The proposed strategy of grouping attempts to partition the elements of the covariates in the matrix $\mathbf{X}$ into $K$ distinct clusters denoted by $C_1, C_2, \cdots, C_K$ in the first stage and in the second stage the categories of the dependent variable are used to partition the observations of each cluster into two sub clusters. Therefore, the

following simulation studies are implemented under the null hypothesis that the fitted model is the correct model to provide the above objectives.

## 4.1 Assess the Performance of the Proposed Strategy of Grouping

The idea of this simulation study is taken from [6] with some modification. Two data settings are considered to examine the performance of the proposed strategy of grouping when the following fitted LRMs are the correct model:

Model (1)

$$logit\big(\pi(\mathbf{x}_i)\big) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \tag{14}$$

Model (2)

$$logit\big(\pi(\mathbf{x}_i)\big) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \tag{15}$$

These models represent an increasing in complexity according to the number of covariates and differ in the distributions. The first model in Eq. (14) consists of only continuous covariates $X_1 \sim U(-6,6)$ and $X_2 \sim N(0,2)$. The second model in Eq. (15) consists of discrete covariate $X_3 \sim Poisson(6)$ in addition to the covariates in the first model to evaluate the effect of the discrete covariate on the strategy of clustering. Therefore, these models represent the expected cases that may occur in practice.

The distribution of $\pi(\mathbf{x}_i)$ is considered as a transformation of the distribution of $\mathbf{x_i}$. For example the Uniform distribution $U(-6,6)$ produces a symmetric distribution with mostly small or large probabilities; while the covariate has a highly skewed right distribution such as the $\chi^2(4)$ distribution results mostly small but a few large probabilities. Other choices for the distribution produce a more uniform distribution of probabilities [6]. The $\boldsymbol{\beta}$ values are chosen in these models according to the following properties [29]: **(*a*)** the signs of $\boldsymbol{\beta}$ determine whether $\pi(\mathbf{x}_i)$ is increasing or decreasing as $\mathbf{x}_i$ increases. **(*b*)** When $\boldsymbol{\beta} \to \mathbf{0}$ the curve of $\pi(\mathbf{x}_i)$ is flatten to a horizontal straight line. **(*c*)** When $\boldsymbol{\beta} = \mathbf{0}$ $Y$ is independent of $\boldsymbol{X}$. **(*d*)** for quantitative $\boldsymbol{X}$ when $\boldsymbol{\beta} > \mathbf{0}$ the curve of $\pi(\mathbf{x_i})$ have the shape of the cumulative distribution function $(cdf)$ of the logistic distribution.

According to these considerations we choose the distributions of the covariates. Also, we choose the beta coefficients of the models according to these characteristics. Consequently, the beta coefficients are chosen as $\beta_0 = 0,\ \beta_1 = 0.8,\ \beta_2 = 0.75,$ and $\beta_3 = 0.60$ to give approximately same importance for the covariates. In this simulation study, the performance of the proposed strategy of grouping is assessed. Therefore, the first step is generating sample of size $n = 200$ values of the covariates according to their properties in both data settings. The proposed strategy of grouping is used to partition the covariates space into $K$ clusters in the first step. The results in Table 2 show the values of the criteria (frequency of each cluster, cluster criterion CC and dissimilarity criterion $R^2$) for $K = 5, 10$ or 15 in both models.

These criteria will be used to evaluate the performance of the proposed strategy of clustering. All these criteria show that the value of *CC* decreases when the number of clusters $K$ increases, on the other hand, the value of $R^2$ increases when the number of clusters $K$ increases. These results also show that the number of clusters $K = 15$ is inappropriate to be used in construction of GOF test of chi-square type because it gives more than one frequency less than 5 in spite of the low value of *CC* and a high value of $R^2$.

At the same time these results show that the number of clusters $K = 10$ shows a potential use where all the three criteria agree that this number is the best to be used in constructing a chi-square type GOF tests. In brief, these results (large frequency in each group, small value of CC and large value of $R^2$) show that the proposed strategy of clustering has adequate efficiency to give separate clusters (groups) regardless of the number of clusters or the number of covariates.

## 4.2 Assess the Performance of the Proposed GOF Tests ($X_c^2$ and $D_c^2$)

The performance of the proposed GOF tests ($X_c^2$ and $D_c^2$) is assessed in the second step. Extensive simulation studies were conducted to assess the power of the proposed GOF tests ($X_c^2$ and $D_c^2$) to detect the particular types of departure from LRMs under a relatively wide variety of data settings. The power of a statistical test is the probability that it will correctly lead to the rejection of a false null hypothesis [30], and also, he defined the statistical power as the ability of a test to detect an effect, if the effect actually exists. Conventionally a test with a power greater than 80% is considered statistically powerful [31]. The detection power of the proposed GOF test statistics are evaluated by simulating data from several LRMs, and then the models are fitted with purposely exclude terms.

**Table 2. The results of conducting the strategy of clustering on both models and $n = 200$**

| **Model 1:** $logit(\pi(\mathbf{x}_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $*C_k$= **K** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | CC | $R^2$ |
| **5** | 42 | 43 | 26 | 35 | 54 | | | | | | | | | | | 1.08 | 0.87 |
| **10** | 21 | 14 | 21 | 11 | 30 | 16 | 18 | 27 | 17 | 25 | | | | | | 0.76 | 0.94 |
| **15** | 18 | 8 | 17 | **04** | 14 | **03** | 14 | 22 | 22 | 15 | 10 | 6 | 13 | 23 | 11 | 0.59 | 0.96 |
| **Model 2:** $logit(\pi(\mathbf{x}_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$ | | | | | | | | | | | | | | | | | |
| **5** | 44 | 23 | 35 | 52 | 46 | | | | | | | | | | | 1.38 | 0.75 |
| **10** | 20 | 21 | 21 | 23 | 07 | 24 | 21 | 05 | 24 | 34 | | | | | | 1.12 | 0.85 |
| **15** | 14 | 17 | 29 | 12 | 19 | 12 | 15 | 17 | **01** | 19 | 06 | 12 | 12 | 10 | 05 | 0.98 | 0.89 |

*$*C_k$ is the clusters and the number in each cell represent the frequency of each cluster.*
*** K=Number of clusters*

### 4.2.1 The detection power of omitting the quadratic term

In the first data setting, we use the idea of the data setting that was modified by Xie et al. [15] from the one used in Hosmer et al. [6] to investigate the effect of deleting a quadratic term in a continuous covariate on the performance of the proposed GOF tests. The model has two continuous covariates ($X_1$ and $X_2$) and the quadratic term of $X_1$ as follows:

$$logit(\pi(\mathbf{x}_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} \tag{16}$$

where $X_2 \sim U(-3,3)$, we used three different distributions for $X_1$: $X_1 \sim U(-1,1)$, $X_1 \sim N(0,4)$ and $X_1 \sim Beta(4,2)$. The values of the coefficients of the model are defined as in the above modified study. These values are chosen as: $\beta_0 = -3.2324$, $\beta_1 = 0.5583$, $\beta_2 = 0.5002$ and $\beta_3 = 1$. In this simulation study we assess the effect of deleting the quadratic term $X_1^2$ from the model on the performance of the new GOF test statistics and compare the performance of the proposed GOF tests with the performance of and HL GOF test ($\hat{C}_{HL}$) and the GOF tests proposed by Xie et al. [15] using these reduced models. Table 3 shows the test size (type I error rates) and the percent of times each of the tests rejected the hypothesis of fit (test power) at the $\alpha = 0.05$ level using sample sizes $n = 200$ and $n = 500$ with 500 replications.

The results in Table 3 show that the estimated Type I error rates for the proposed GOF tests ($X_c^2$ and $D_c^2$) appear to be on the conventional side for the supposed level $\alpha = 0.05$ of the test in both sample sizes and for all the distributions. These results also show that HL GOF test does not control the Type I error in both sample sizes and in most distributions. In this case, HL GOF test has an inflated Type I error rates in both sample sizes and most distributions. Furthermore, these results show that the estimated Type I error rates for the Pearson chi-square type statistic $X_{P^*}^2$ which is proposed by Xie et al. [15] appears to be on the conventional side for the supposed level $\alpha = 0.05$ of the test. Meanwhile the score test statistic $T^*$ which is proposed by Xie et al. [15] has inflated Type I error rates in both sample sizes and most distributions.

According to these estimated Type I error rates, it is important to make detection power comparisons between the proposed GOF tests and the other tests. Consequently, Table 3 shows the results for different distributions. For all these distributions, the detection powers of the two proposed GOF tests are higher than the detection power of HL GOF test. This is in addition to the lack of control on the Type I error of this test. Meanwhile, the detection powers of the proposed GOF tests are similar to the detection powers of chi-square test proposed by Xie et al. [15] but higher than the detection powers of the score test $T^*$, in spite of the detection powers for $T^*$ test when the distribution is uniform in both sample sizes are questionable according to their Type I error rates.

**Table 3. Type I error rates and detection powers when the quadratic term is omitting**

| **Original model [a]: $logit(\pi(x)) = -3.2324 + 0.5583X_1 + 0.5002X_1^2 + X_2$** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fitted Model: $logit(\pi(x)) = -3.2324 + 0.5583X_1 + X_2$** | | | | | | | | | | | | |
| | **Type I error rates (Test size)** | | | | | | **Test power (%)** | | | | | |
| | **n=200** | | | **n=500** | | | **n=200** | | | **n=500** | | |
| $X_1$ ~[b] | *U* | *N* | *Beta* | *U* | *N* | *Beta* | *U* | *N* | *Beta* | *U* | *N* | *Beta* |
| $X_c^{2\,c}$ | **0.01** | **0.02** | **0.01** | **0.01** | **0.01** | **0.01** | **100** | **99.9** | **100** | **100** | **100** | **100** |
| $D_c^{2\,c}$ | **0.01** | **0.01** | **0.01** | **0.01** | **0.01** | **0.01** | **100** | **100** | **100** | **100** | **100** | **100** |
| $\hat{C}_{HL}$ [d] | 0.07 | 0.06 | 0.06 | 0.05 | 0.05 | 0.06 | 11.7 | 34.3 | 16.7 | 18.6 | 66.3 | 21.6 |
| $X_{P^*}^{2}$ [d] | 0.04 | 0.02 | 0.04 | 0.02 | 0.02 | 0.03 | 85.7 | 99.9 | 29.6 | 99.9 | 99.9 | 58.2 |
| $T^{*}$ [d] | 0.07 | 0.04 | 0.07 | 0.06 | 0.05 | 0.06 | **89.3** | 100 | 34.5 | **100** | 99.9 | 64.3 |

[a] *Parameters are chosen in a way that the quadratic term effect has high impact in the prediction function (Xie et al. [15])*

[b] *U is uniform distribution; N is normal distribution; Beta is Beta(4,2) distribution.*

[c] *$X_c^2$ is the proposed clustered chi-square GOF test statistic, $D_c^2$ is the proposed clustered chi-square deviance.*

[d] *$X_{P^*}^2$ is the Pearson chi-square GOF test statistic and $T^*$ is the score statistic both GOF tests proposed by Xie et al. [15] $\hat{C}_{HL}$ is the Hosmer and Lemeshow test statistic, their values in the table calculated by Xie et al. [15]*

In general, the detection powers of the proposed GOF tests are somewhat higher than that of HL GOF test and somewhat similar to the detection powers of Xie et al. [15] GOF tests. The increase in the sample size and the skewness in the distribution of the deleted quadratic terms do not affect the detection powers of the proposed GOF tests but the detection powers of other GOF tests are increased as sample size increases. The skewness in the distribution of the deleted quadratic term decreases the detection powers of other GOF tests and inflated the Type I error rates. The proposed GOF tests maintain higher or at least equal power as the other GOF tests.

#### 4.2.2 The detection power of omitting the interaction between dichotomous and continuous variables

The idea of the second model is taken from [2] to investigate the performance of the new GOF test statistics and the ability to detect the omission of the interaction between dichotomous and continuous variables. The model is given by:

$$g(X,d) = \beta_0 + \beta_1 X + \beta_2 d + \beta_3 Xd \tag{17}$$

where $X \sim U(-3;\ 3)$ and $d$ is Bernoulli with $p = 0.5$. Parameters were chosen at specified values $\beta_0 = 0.10$, $\beta_1 = 0.10$, $\beta_2 = 0.20$ and $\beta_3 = 0.20 + m$ where $m = 0.10;\ 0.30;\ 0.50$ or $0.70$. These values of $m$ allow a series of models that increase in the strength of the interaction. The dependent variable $Y$ is simulated by comparing an independently simulated covariate $W \sim U(0,1)$ with the true logistic probabilities $\pi(x,d) = Exp\big(g(X,d)\big)/\Big(1 + Exp\big(g(X,d)\big)\Big)$ which is estimated from fitting the LRMs in Eq. (17) and the rule: $Y = 1$ if $W \le \pi(x,d)$ and $Y = 0$ otherwise. According to these properties four LRMs differ in the strength of the interaction are fitted with omitting the interaction term to investigate the performance of the proposed GOF test statistics $X_c^2$ and $D_c^2$. The performance of the proposed GOF test statistics is compared to the HL GOF test $\hat{C}_{HL}$ and Pulkstenis-Robinson GOF test statistics ( $X_{PR}^2$ and $D_{PR}^2$ ) using these reduced models. The data is generated with the above model and then the model is fitted with omitting only the

interaction term. A total of 500 iterations were performed for sample sizes $n = 100$ and $n = 500$ subjects and the detection powers are provided in Table 4.

These results indicate that all GOF test statistics have detection powers that increase with both sample size and the strength of the interaction. Also, for this simulation study, the proposed GOF test statistics $X_c^2$ and $D_c^2$ have somewhat higher detection powers than the other GOF test statistics.

**Table 4. Simulated powers of detection from fitting model in Eq. (17) with omitting interaction term $Xd$**

| **Original model: $g(X,d) = \beta_0 + \beta_1 X + \beta_2 d + \beta_3 Xd$** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Fitted model: $g(X,d) = \beta_0 + \beta_1 X + \beta_2 d$** | | | | | | | | |
| **Sample size** | **$n = 100$** | | | | **$n = 500$** | | | |
| Interaction coefficients | | | **$\beta_3 =$** | | | | **$\beta_3 =$** | |
| Statistics | *0.3* | *0.5* | *0.7* | *0.9* | *0.3* | *0.5* | *0.7* | *0.9* |
| $\chi_P^2$ | 5.8 | 4.9 | 5.0 | 3.2 | 4.8 | 5.4 | 4.0 | 1.8 |
| $\hat{C}_{HL}$ | 4.3 | 3.9 | 3.0 | 5.2 | 5.8 | 4.2 | 6.0 | 6.8 |
| $\chi_{PR}^2$ | 6.2 | 15.4 | 23.8 | 40.4 | 9.0 | 29.2 | 65.5 | 96.2 |
| $D_{PR}^2$ | 9.2 | 16.2 | 22.8 | 37.4 | 9.4 | 28.6 | 64.6 | 95.6 |
| $X_C^2$ | **18.6** | **20.4** | **28.2** | **30.0** | **52.6** | **67.2** | **72.8** | **80.4** |
| $D_c^2$ | **12.2** | **23.4** | **24.6** | **26.6** | **49.2** | **63.4** | **62.0** | **65.6** |

#### 4.2.3 The detection power of omitting the main effect term

In addition to the ability to detect the omission of the interaction term, it is of interest to evaluate the power to detect the omission of the main effect term. This was done by generating data from the following model [2] and then the model is fitted after excluding $\beta_2 X_2$:

$$logit\pi(\mathbf{x}_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 Z \tag{18}$$

In this model the regression coefficients are chosen as $\beta_0 = -0.25$, $\beta_1 = 0.75$, $\beta_2 = -2.75$, $\beta_3 = 0.5$, $\beta_4 = 0.25$, $X_i \sim U(-6, 6)$ and $Z \sim U(-2, 2)$. The dependent variable $Y$ is simulated as in the above simulation studies. A total of 500 replications were performed for sample sizes of $n = 200$; $n = 400$ and $n = 800$ elements and the detection powers are provided in Table 5.

**Table 5. Powers of detection from fitting the model in Eq. (18) after omitting the main effect $X_2$**

| **Original model: $logit\pi(\mathbf{x}_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 Z$** | | | |
|---|---|---|---|
| **Fitted model: $logit\pi(\mathbf{x}_i) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 Z$** | | | |
| **Statistics** | **$n = 200$** | **$n = 400$** | **$n = 8\ 00$** |
| $\hat{C}_{HL}$ | 4.6% | 3.4% | 3.4% |
| $X_{PR}^2$ | 4.8% | 6.6% | 5.6% |
| $D_{PR}^2$ | 5.2% | 7.0% | 5.8% |
| $X_c^2$ | **7.8%** | **10.2%** | **15.7%** |
| $D_c^2$ | **6.2%** | **9.8%** | **12.3%** |

The results given in Table 5 indicate that none of these GOF tests are particularly powerful to detect this kind of misspecification in the model.

#### 4.2.4 The detection power of a misspecification link function

Lastly, we examined the ability to detect a misspecification of the link function by generating data from the following complimentary log-log model and incorrectly fitting a logit model [2]:

$$log\big(-log(1-\pi)\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z \quad (19)$$

where $\beta_0 = -1.0$, $\beta_1 = 0.5$, $\beta_2 = -0.5$, $\beta_3 = 0.10$, $X_i \sim U(-6,6)$ and $Z \sim U(0,15)$. A total of 500 replications were performed for sample sizes of $n = 200$; $n = 400$ and $n = 800$ elements and the power of detection from fitting the model in Eq. (19) with a misspecification link function are provided in Table 6.

**Table 6. Simulated power of detection from fitting the model in Eq. (19) with a misspecification link function**

| **Original model: $log\big(-log(1-\pi)\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z$** | | | |
|---|---|---|---|
| **Fitted model: $logit(\pi_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z$** | | | |
| **Statistics** | **$n = 200$** | **$n = 400$** | **$n = 8\ 00$** |
| $\widehat{C}_{HL}$ | 4.6% | 3.4% | 3.4% |
| $X^2_{PR}$ | 4.8% | 6.6% | 5.6% |
| $D^2_{PR}$ | 5.2% | 7.0% | 5.8% |
| $X^2_c$ | 7.8% | 10.2% | 15.7% |
| $D^2_c$ | 6.2% | 9.8% | 12.3% |

The results given in Table 6 again indicate a fairly low power to detect this type of lack-of-fit in the model.

# 5 Clinical Data Set

As an example in real situation, the survival times of 501 patients with prostate cancer who are randomly allocated to a treatment with an estrogen are considered. This data set was considered by several authors such as [32-33] and more recently by [34]. In this data set, the value of the outcome variable status classifies the cause of death as 1 = cancer (the event of interest) and 0 = other. The patients are considered treated if they received at least 1 mg of estrogen daily and the other covariates as described in Table 7. This data set will be used to assess the performance of the proposed strategy of grouping based on partitional clustering in the first step and in the second step the performance of the developed GOF tests is assessed as in the following subsections.

**Table 7. Description of the risk factors in the clinical data set**

| **Variable name** | **Description** | **Code** |
|---|---|---|
| Pat-ID | Patient number | |
| Status (Y) | The cause of death | 1= cancer<br>0= others |
| Treat ($X_1$) | treatment estrogen (mg) | 0.0 (Placebo)<br>0.2<br>1.0<br>5.0 |
| Ftime ($X_2$) | follow-up time (month) | |
| Age ($X_3$) | age ( year) | |
| Wtind ($X_4$) | weight index=wt(kg)-ht(cm)+200 | |
| SBP ($X_5$) | Systolic Blood Pressure/10 | |
| Shem ($X_6$) | Serum Hemoglobin (g/100ml) | |
| Size ($X_7$) | Size of Primary Tumor (cm^2) | |
| SPAP ($X_8$) | Serum Prostatic Acid Phosphates | |
| HCD ($X_9$) | History of Cardiovascular Disease | 0= has not<br>1=has |

## 5.1 Assess the Performance of the Proposed Strategy of Clustering

We used the proposed strategy of grouping to partition the space of the continuous covariates in this data set for different number of clusters $K = 5,\ 10$ or $15$ and assess the performance of this strategy based on the criteria discussed above. The results in Table 8 present the estimated values of these criteria.

**Table 8. The criteria of assessing the performance of the proposed strategy of clustering in clinical data set**

| $C_k$ K | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | CC | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 58 | 169 | 27 | 129 | 118 | | | | | | | | | | | 0.79 | 0.33 |
| 10 | 7 | 72 | 85 | 90 | 73 | 61 | 44 | 16 | 43 | 10 | | | | | | 0.71 | 0.50 |
| 15 | 35 | 30 | 35 | 5 | 17 | 24 | 7 | 87 | **1** | 37 | 44 | 86 | **1** | 37 | 55 | 0.64 | 0.56 |

The results in Table 8 show the values of the criteria which are used to assess the performance of the proposed strategy of grouping in this data set. All these results show that the values of CC decreased and the values of $R^2$ increased when the number of clusters K increased. Also all these results show that the number of clusters $K = 15$ is not suitable to be used in constructing GOF tests of chi-square type because it gives more than one frequency less than 5 in spite of it gives a smaller value of CC and higher value of $R^2$. At the same time these results show that the number of clusters $K = 10$ shows a potential use where all these criteria agree that this number is the best to be used in constructing GOF tests of chi-square type. In brief these results show that the proposed strategy of clustering has adequate efficiency to give separate clusters (groups) with a moderate number of clusters.

## 5.2 Assess the performance of the proposed GOF tests

The performance of the proposed GOF tests ($X_c^2$ and $D_c^2$) is assessed in the second step. In order to do this, the outcome variable status ($y$) classified the cause of death as 1 = cancer (the event of interest) and 0 = others. The objective of this study is to model this outcome as a function of various covariates. Thus, two LRMs are candidates to associate this outcome variable with the different covariates in this clinical data set as follows:

Model 1

$$logit\big(\pi(\mathbf{x}_i)\big) = \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i3} + \beta_3 x_{i4} + \beta_4 x_{i5} + \beta_5 x_{i6} + \beta_6 x_{i7} + \beta_7 x_{i8} \quad (20)$$

Model 2

$$logit\big(\pi(\mathbf{x}_i)\big) = \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i3} + \beta_3 x_{i4} + \beta_4 x_{i5} + \beta_5 x_{i6} + \beta_6 x_{i7} + \beta_7 x_{i8} + \beta_8 x_{i9} + \beta_9 x_{i1} \quad (21)$$

where $\pi(\mathbf{x}_i) = Pr(y_i = 1|\mathbf{x}_i)$ and the covariates $\mathbf{X}$ are described as in Table 8. Model 1 represents the relationship between outcome variable and the continuous covariates in the data set. Model 2 represents the relationship between the outcome variable and the combined covariates (continuous and categorical). Since Model 1 is nested in Model 2, the likelihood ratio test was used to compare the two models and the result showed that Model 1 is a significant departure from Model 2 ($p - value = 0.0223$). Table 9 represents the decisions of the proposed GOF tests along with the HL GOF test to assess these two models.

The results in Table 9 show that while the HLGOF test accepts the two models, both of the proposed GOF tests reject Model 1, but none of these GOF tests finds evidence to reject Model 2. These decisions to reject Model 1 are also supported by the likelihood ratio test, which indicates that Model 1 is nested in Model 2. Also, these results show, based on clinical considerations and the original analysis of this data set, that Model 2 with all the covariates is an adequate model to associate the outcome variable status with the covariates in this clinical data set.

**Table 9. Decisions of the three GOF tests on the three LRMs at $\alpha = 0.05$ in the clinical data set**

| GOF tests | Decisions on different models | |
|---|---|---|
| | **Model 1** | **Model 2** |
| $X_c^2$ | Reject | Not reject |
| $D_c^2$ | Reject | Not reject |
| $\hat{C}_{HL}$ | Not reject | Not reject |

# 6 Conclusions

The results from implementing the simulation studies and analyzing the clinical data set assist to conclude that the proposed strategy of grouping has adequate efficiency in giving separate clusters. It is more flexible in determining the number of clusters. It is appropriate to construct the GOF test statistics of chi-square type. It helps to avoid the problems with the similarity measurements in the categorical data and it does not need to convert the continuous covariates, so that the information is not lost.

On the other hand, two GOF test statistics $X_c^2$ and $D_c^2$ are constructed based on the proposed strategy of grouping. The results from conducting the simulation studies and analyzing the clinical data set assist to conclude that the proposed GOF test statistics have a wide range of detection to detect any departure in the fitting model from the true model. They have adequate power of detection higher than the existing GOF test statistics for different factors such as sample size and the departure from the true model. They have a high power of detection when the simple LRM (with one covariate) or multiple LRM (with more than one covariate) are used to fit the data set and they have adequate performance when the sample size is large.

In this paper, the problems of partitioning the space of the covariates and developing GOF tests are studied for the binary LRM. Thus, for improvement and future works the strategy of clustering and the GOF test statistics based on it can be adapted to apply to the other logistic regression models, such as ordinal or multinomial LRMs. Also, another area of research is adapting this strategy of grouping and the GOF test statistics for use in the other GLMs such as Poisson regression model and log-linear model. The proposed strategy is based on partitioning the range of continuous covariates using $K$-means and multiple comparisons to determine the appropriate number of clusters $K$. Further study may result in a better or optimal rule in determining $K$ and choice of clustering method. Another future study, a partitional clustering process may be developed to cluster the combined data set (continuous and categorical) covariates at the same time, when both continuous and categorical covariates are present. Adapting the strategy of clustering to partition the data set with categorical covariates only is also, another area of future work.

# Competing Interests

Authors have declared that no competing interests exist.

# References

[1] Mc Cullagh P, Nelder JA. Generalized linear models. Chapman and Hall; 1989.

[2] Pulkstenis E, Robinson TJ. Two goodness-of-fit tests for logistic regression models with continuous covariates. Statistics in Medicine. 2002;21:79-93.

[3] Archer KJ, Lemeshow S, Hosmer DW. Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. Computational Statistics & Data Analysis. 2007;51:4450–4464.

[4] Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Communication in Statistics: Part a-Theory and Methods. 1980;9(10):1043–1069.

[5] Hosmer DW, Lemeshow S. Applied logistic regression. John Wiley & Sons, Inc.; 1989.

[6] Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in Medicine. 1997;16:965-980.

[7] Pigeon JG, Heyse JF. A cautionary note about assessing the fit of logistic regression models. Journal of Applied Statistics. 1999;26(7):847–853.

[8] Kuss O. Global goodness-of-fit tests in logistic regression with sparse data. Statistics in Medicine. 2002;21:3789–3801.

[9] Lipstiz SR, Fitzmaurice GM, Molenberghs G. Goodness-of-fit tests for ordinal response regression models. Applied Statistics. 1996;45(2):175-190.

[10] Zhang B. A chi-squared goodness-of-fit test for logistic regression models based on case-control data. Biometrika. 1999;86( 3):531–539.

[11] Evans S, Li L. A comparison of goodness of fit tests for the logistic GEE model. Statistics in Medicine. 2005;24:1245–1261.

[12] Deng X, Wan S, Zhang B. An improved goodness-of-fit test for logistic regression models based on case-control data by random partition. Communications in Statistics-Simulation and Computation. 2009;38:233–243.

[13] Tsiatis AA. A note on a goodness-of-fit test for the logistic regression model. Biometrika. 1980;67(1): 250–251.

[14] Fagerland MW, Hosmer DW, Bofin AM. Multinomial goodness-of-fit tests for logistic regression models. Statistics in Medicine. 2008;27(21):4238-4253.

[15] Xie X, Pendergast J, Clarke W. Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. Computational Statistics & Data Analysis. 2008;52:2703–2713.

[16] Everitt BS. Cluster analysis. 2nd ed. Halsted Press: A Division of John Wiley & Sons, Inc; 1980.

[17] Redmond SJ, Heneghan CA. Method for initializing the K-means clustering algorithm using *kd*-trees. Pattern Recognition Letters. 2007;28:965–973.

[18] Lei M, He P, Li Z. An improved K-means algorithm for clustering categorical data. Journal of Communication and Computer. 2006;3(8):20-24.

[19] Everitt BS, Landau S, Leese M. Cluster analysis. 4th ed. Arnold a Member of the Holder Headline Group; 2001.

[20] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001;63(2):411–423.

[21] Lukashin AV, Fuchs R. Analysis of temporal gene expression profiles: Clustering by simulated annealing and determining the optimal number of clusters. Bioinformatics. 2001;17(5):405–414.

[22] Fred ALN, Leitão JMN. A new cluster isolation criterion based on dissimilarity increments. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2003;25(8):944–958.

[23] Bertrand P, Bel Mufti G. Loevinger's measures of rule quality for assessing cluster stability. Computational Statistics & Data Analysis. 2006;50(4):992–1015.

[24] Chae SS, DuBien JL, Warde WD. A method of predicting the number of clusters using Rand's statistic. Computational Statistics & Data Analysis. 2006;50(12):3531-3546.

[25] Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. John Wiley & Sons, Inc; 2000.

[26] DasGupta A. Asymptotic theory of statistics and probability. Springer Science and Business Media LLC; 2008.

[27] Lei M, He P, Li Z. An improved K-means algorithm for clustering categorical data. Journal of Communication and Computer. 2006;3(8):20-24.

[28] Bagirov AM. Modified global k-means algorithm for minimum sum-of-squares clustering problems. Pattern Recognition. 2008;41:3192-3199.

[29] Agresti A. Categorical data analysis. John Wiley & Sons, Inc; 2002.

[30] Greene WH. Econometric analysis. 4th ed. Upper Saddle River, NJ: Prentice Hall; 2000.

[31] Mazen AM, Graf LA, Kellogg CE, Hemmasi M. Statistical power in contemporary management research. The Academy of Management Journal. 1987;30(2):369-380.

[32] Kay R. Treatment effects in competing-risks analysis of prostate cancer data. Biometrics. 1986;42:203–211.

[33] Lunn M, McNeil D. Applying Cox regression to competing risks. Biometrics.1995;51:524–532.

[34] Coviello V, Boggess M. Cumulative incidence estimation in the presence of competing risks. The Stata Journal. 2004;4(2):103–112.



***Peer-review history:***
*The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)*
*http://sciencedomain.org/review-history/9908*