AL- Rafidain
University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain
University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain
University College for
Sciences

The Use Penalized Quasi Likelihood (PQL) to Estimate Multilevel Binary Logistic Model to Identify the Factors of Anemia

Zainab M. Redha — zainab.reda@s.uokerbala.edu.iq

Nabaa A. Mohsen — nabaa.abbas@s.uokerbala.edu.iq

Department of Statistics - Faculty of Administration and Economic - Kerbala University, Kerbala, Iraq

Mehdi W. Naserallah

mehdi.wahab@uokerbala.edu.iq

Department of Statistics - Faculty of Administration and Economic - Kerbala University, Kerbala, Iraq

Article Information

Article History:

Received: December, 20, 2022

Accepted: March, 3, 2023

Available Online: December, 31, 2023

Keywords:

Binary logistic regression analysis, Multilevel Binary logistic regression analysis, penalized quasi maximum Likelihood, regression, Hierarchical data

Abstract

Multilevel logistic regression analysis is used in many fields, including health, medical, geography, social and educational, where the researchers were interested in this analysis to identify and study the nature of the relationship between the behavior of the vocabulary or units of study, and the social, environmental and economic variables in different environments in which they live and belong, and in such hierarchical data. In this research, multilevel data was used, Data were taken on anemia in children, as the dependent variable represents anemia infections (infected - uninfected) from four hospitals affiliated to the Babylon Health Department, namely (Al-Hah General Teaching Hospital - Imam Al-Sadiq Hospital (PBUH) - Marjan General Teaching Hospital - Babel Women's and Children's Hospital) which represents the third level in the analysis and according to the type of lobby (public - private), which represents the second level. (50) cases of anemia were taken from the general ward from the general ward and (20) cases from the private ward, and from Imam al-Sadiq Hospital (P) Peace, (75) sick cases were taken from the general ward and (20) sick cases from the private ward. From Marjan General Teaching Hospital, (30) sick cases were taken from the private ward and (20) sick cases from the general ward, and (40) sick cases were taken from the general ward and (25) sick cases from the general ward, and these sick cases represent the level Third, in the multi-level analysis, so that the total number of disease cases is (290) cases, and the independent variables that can affect anemia were taken (sex, age, weight, occupation, marital status, smoking, academic achievement, place of residence, infection with other diseases, blood pressure). That is, the first level contains (10) independent variables. The logistic analysis of the multilevel binary was carried out using the NCSS 2022 program, using the method of possibility semi-penalty. has been reached Significance of the multi-level binary logistic regression model, as the p-value = 0.0015, which is less than the level of significance 1%, and achieved a high odds ratio of (0.7767). The variables (age - occupation - smoking - blood pressure) are not significant and the variables (weight - water source - place of residence - other diseases - wealth index) are significant.

Correspondence:

Zainab M. Redha

zainab.reda@s.uokerbala.edu.iq

https://doi.org/10.55562/jrucs.v54i1.617

1. Introduction

Anemia is a disorder that occurs when your blood produces fewer healthy red blood cells than is typical. Your body does not receive enough oxygen-rich blood if you have anemia. You may feel exhausted or weak due to a lack of oxygen. Additionally, you can get headaches, nausea, or shortness of breath. About 3 million Americans have anemia. Mild anemia is a common and treatable condition that can develop in anyone. It may come about suddenly or over time, and may be caused by your diet, medicines you take, or another medical condition. Anemia can also be chronic, meaning it lasts a long time and may never go away completely. Some anemia forms are hereditary. Iron-deficiency anemia is the most prevalent kind of anemia. Anemia is more common in some persons, particularly in pregnant and menstruating women. Additionally, those who take particular medications or treatments, do not get enough iron, do not get certain vitamins, and are at increased risk. Anemia may potentially indicate a more serious ailment, such as gastrointestinal bleeding, infection-related inflammation, renal illness, cancer, or autoimmune disorders. Anemia is diagnosed by your doctor based on your medical history, a physical examination, and test findings. Depending on the kind and severity of your anemia, you may need treatment. You could require iron supplements, vitamins, or medications that stimulate the production of additional red blood cells for some kinds of mild to severe anemia. To prevent anemia in the future, your doctor may also suggest healthy eating changes. The fact that the observations take on a hierarchical structure makes multi-level models one of the most popular models for application and data analysis. Multi-level models may be classified into three categories (random section model, random slope model, random section model and random slope). In this papers the aim was to use the multi-level binary logistic regression model and estimate its dependency using Penalized Quasi Likelihood (PQL) method to determine the most important factors affecting the incidence of anemia, since the usual regression models calculate the random error for one source of data, but when there are several levels of extracted data, we get several random lights that come From every level thought out.

2. Multilevel Data[1]

Before getting into the specifics of the multi-level analysis, it is important to comprehend what a level is since, from a statistical perspective, it differs greatly from the idea of a known variable. Characteristics are different from another set of vocabulary, so that it can be said that it is a unique statistical community (unique with characteristics) whose features can be identified and accessed by itself, but if there is a second level (subsequent) to this first level, the (next) or second level must be this. However, in the event of a third level of data, it must be one of the bigger units that must contain the items, observations, or units in the second level. They are larger units that include the items, observations, or partial units in the previous (first) level, and in the case of level four of the data, it must be one of the biggest units that must include items, observations, or units from level three (whether or not they overlap at each level), under the condition that the groups that include those units are a random sample from the collection with a lot of these units, and that each level of the vocabulary has a set of characteristics that can be expressed as random variables, whether qualitative or quantitative, as a result, it is the lowest level, whereas the top level is the final. For instance, if we have a community (students), they are a first level (lower level) because they share certain characteristics that set them apart. Colleges are a second level and include the lower level (students), universities are a third level and include colleges (third level), and governorates are a fourth level that includes universities (higher level). The sample taken from these communities may be thought of as a multistage sample when the structure and nature of the data acquired from a given community are in a hierarchical manner., reducing the guarantee, increasing the efficiency of the results, can be taken into account in such samples. It gives a good description of the community variables, these variables are called multilevel data, but there are many difficulties in reaching results about these variables, as the data depends on each other, and there is a state of correlation between the observations due to the multiplicity of levels within the hierarchical data. In this case, the use of single-level statistical models is not effective, and therefore we need a complex statistical model that accommodates the hierarchical shape of the data to reach good inferences about the community.

3. Nested and not Nested Data[2]

It is that each of the vocabulary of any level belongs to one unit or vocabulary of one and only one of the vocabulary of the other level, so that each unit or vocabulary of the other level belongs to one and only one of the vocabulary of the next level and so on. For example, students belong to colleges, and these colleges belong to one university, and these universities are located in one governorate. And the non-overlapping data, the units of the first level can belong to more than one of the units of the other level, for example, students in a particular area belong to several schools, and students in one of the middle schools enroll in different colleges or different universities.

4. Multilevel Analysis [3][4]

It is one of the types of statistical analysis imposed by the nature of the data used in the study, and it is the analysis in which the members of the study population belong to different groups of groups grouped with a smaller number of larger units, and the latter belong in different numbers to a smaller number of larger units and so on. In the one-level analysis, there are no groups higher than that level, and in such a case the explanatory variables affect the dependent variable independently of each other, and as a result a single random error is generated in the regression model specified for that relationship, which is supposed to have a normal distribution with an average Zero and a certain variance, and this type of analysis (single-level) in which the influence of the independent variables is direct on the dependent variable (even if it was affected by other independent variables), but this effect is unacceptable if it is increased, and the independent variable most closely related to other independent variables must be deleted. In the case of more than one level of data, a group of first-level units all follow one of the second-level units, and therefore they are all affected by the characteristics of that group. In other words, the second-level units have the same effect on all the variables of the items to which they belong. As a result, the specific explanatory variables are determined. For the units of the first level based on the characteristics of that unit itself from the units of the second level with the existence of a random border between the units of the first level, and accordingly, in the case of more than one unit in the second level, each of them has a different effect from one unit to another, and thus a difference occurs between the vocabulary of the first level As a result of that dependence, in addition to the discrepancy that exists between those vocabulary, and therefore the dependent variable that measures a certain phenomenon in the vocabulary of the first level is affected by two types of variation, each of them must be identified alone, and this is the basis of the concept of multilevel analysis, and when there are variables from higher levels, it will multiply Random limits that must be estimated in order to determine the impact of all variables at all levels on the response variable that measures a phenomenon. In multi-level analysis we must have at least 20 units of units at the highest level, and this implies that the number of units at the lower level is at least twice that number and this applies to the lower level as well.

5. Binary Logistic Regression [4][5]

The aim of linear regression is to fit a straight line to a number of points that minimize the sum of squares of the remainder. , that is, containing a straight line to a number of points that reduce the sum of the squares of the remainder. Regression models are used for several purposes, describing and analyzing the relationship between variables, predicting and selecting variables. But when the dependent variable in the regression model is of the descriptive (qualitative) type, it takes two values in numeric form (0,1) such as (1 success, 0 failure), (1 disease cured, 0 non-cured disease), (1 arrival, Non-arrival (0), ...etc., it is called the Binary Logistic Regression (BLR) model, and when the dependent variable takes more than two values, it is called the Multiple Logistic Regression (MLR) model, in logistic regression the goal is not to estimate the model parameters (meaning measuring the change caused by the independent variables in the dependent variable), but the goal is to measure the probability of occurrence or non-occurrence of the phenomenon under study, and the binary logistic regression will be the subject of our study in this thesis, but at multiple levels.

Logistic regression is based mainly on the assumption that the dependent variable is a binary variable that follows the Bernoulli distribution. It takes the value (1) with the probability of P (the probability of the response occurring) and the value (0) with the probability of q = 1-P (the probability that the response will not occur). As we know, that in linear regression whose independent variables and the dependent variable take continuous values, the model that connects the variables is:

$$\mathrm{Y} = \mathrm{b}_0 + \mathrm{b}_1\mathrm{X}_\mathrm{i} + \mathrm{e} \qquad (1)$$

Since Y is a variable that represents a continuous variable, and the average observed (real) Y values are E(Y/X) and $\mathrm{e} = \mathrm{Y} - \hat{\mathrm{Y}}$, then equation (1) can be written as follows:

$$\mathrm{E(Y/X)} = \mathrm{b}_0 + \mathrm{b}_1\mathrm{X}_\mathrm{i} \qquad (2)$$

And that in regression, as it is known, the right side of these models takes values from (-∞) to (+∞), but when the dependent variable is binary, it takes values of zero or one, then linear regression is not appropriate because $\mathrm{E(Y/X)} = \mathrm{P(Y = 1)} = \infty$.

Because the value of the right side is confined between zero and one. Thus, the model is not applicable from the point of view of regression, and to solve this problem, the natural logarithm is introduced into the dependent variable, and since $0 \leq \mathrm{P} \leq 1$, the ratio P / (1-P) is a positive amount confined between (∞,0) That is $0 \leq$ P / (1-P) ≤ 1. Therefore, the regression model can be written in the case of one independent variable as follows:

$$\ln\left(\frac{\mathrm{P}}{1-\mathrm{P}}\right) = \mathrm{b}_0 + \mathrm{b}_1\mathrm{X}_\mathrm{i} \qquad (3)$$

If we have more than one independent variable, the model will be as follows:

$$\ln(\mathrm{odds}) = \ln\left(\frac{\mathrm{P}}{1-\mathrm{P}}\right) = \mathrm{b}_0 + \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{b}_\mathrm{j}\mathrm{X}_{\mathrm{ij}};\ \mathrm{j} = 1,2,\ldots..\mathrm{k} \quad , \qquad \mathrm{i} = 1,2,\ldots..,\mathrm{n} \qquad (4)$$

And by taking the inverse of the natural logarithm (Exp) of the function (4), it can be written in the following form:

$$P = \frac{1}{1+\exp(-\mathrm{b}_0 + \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{b}_\mathrm{j}\mathrm{X}_{\mathrm{ij}})} \qquad (5)$$

This model is called the logistic regression model or the Logit model, and the transformation ln(P/(1-P)) is called the Logit transformation or the Logarithm Odds Ratio. And that the logistic function is a continuous function that takes values (0,1), where y approaches zero as the right side approaches (∞-) and y approaches (1) as the right side approaches (∞+), and the logistic function is the same when The right side is equal to 1. Therefore, the logistic regression model is a logarithmic transformation of linear regression by converting it into a logistic function, so it will follow the characteristics of the logistical distribution, which makes the probabilities confined between (1,0), hence its name as logistic regression.

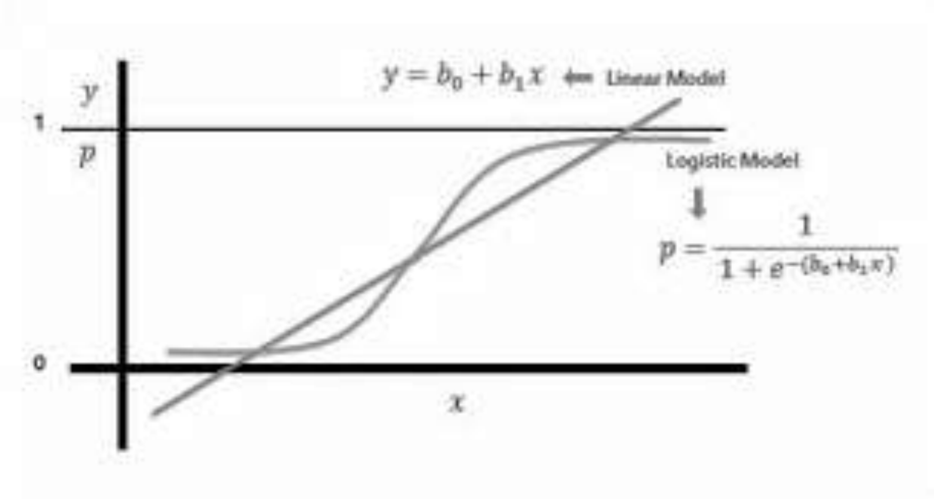

Figure (1): Linear and logistic regression curve

6. Multilevel Binary Logistic Regression [6][7]

It is also called the hierarchical logistic regression model or the logistic regression model with random effects, and it is a model used in the case of hierarchical data that consists of overlapping levels so that the variables at each level are affected by the other levels. The general objective of the multi-level logistic regression is to estimate the probabilities (likelihood) that we can suggest the occurrence of the event or the non-occurrence of the event with a certain probability, taking into account the dependency of the data .The multi-level logistic regression model is considered a natural extension of the single-level logistic regression model (a regression model with independent variables that directly affect the dependent variable) with the treatment of some or all of the parameters in the model to be random instead of being fixed in the single-level model)

single level model:

$$Logit\ p_{ij} = \beta o + \beta_1 x_i \tag{6}$$

Through this model (single level), we can determine the relationship between the dependent variable of the two-response and one of the independent variables in the form of a linear relationship, As the single-level model aims to estimate $p_{ij} = p_r(y_i = 1|x_i)$ where y_i is a dependent binary response variable , p_{ij}The probability that the observation i achieves the response (1) depending on the variable x_i (x_i has an effect on the probability of the occurrence or non-occurrence of the response) As the single-level model aims to estimate:

In a one-level analysis, there are no groups higher than that level. In such a case, the independent variables affect the dependent variable independently of each other. As a result, a single random error is generated, which is assumed to be a normal distribution with zero mean and constant variance. Usually, the goal of using multi-level logistic regression models is to examine hierarchical data and find the relationship between two or more variables, one of which is dependent and belongs to the category of binary descriptive variables that have a Bernoulli distribution, while the nature of the independent variables expected to have a relationship with that dependent variable varies through procedures. Estimating model parameters. The nature of that relationship between the dependent variable and the independent variables can be determined at different levels of the data. A two-level model can be written with one independent variable as follows:

$$Logit\ (p_{ij}) = ln\left[\frac{p_{ij}}{1-p_{ij}}\right] = \beta o + \beta_1 x_{ij} + u_j \qquad (combined\ model\ ) \tag{7}$$

Where $p_{ij} = p_r(y_{ij} = 1)$, u_j the random effect of the second level where $u_j \sim N(0, \hat{\sigma}_U)$, Model (7) can be adopted as the standard logistic model, provided that u_j and y_{ij} are independent

$$Logit\ (p_{ij}) = ln\left[\frac{p_{ij}}{1-p_{ij}}\right] = \beta_{oj} + \beta_1 x_{ij} \qquad (level\ 1\ model\ ) \tag{8}$$

$$\beta_{oj} = \beta_o + u_j \qquad \text{(level 2 model)} \tag{9}$$

The conditional density function for group j is similar to that used in logistic regression and is written in the following formula:

$$f(y_j|x_j, u_j) = \prod_{i=1}^{nj} \frac{\exp\left[y_{i(\beta_o+\beta_1 x_{ij}+u_j)}\right]}{1+\exp(\beta_o+\beta_1 x_{ij}+u_j} \tag{10}$$

Where y_j responses for set j, x_j is the independent variable of set j.

$$f(y_j|x_j) = \int f(y_j|x_{j,}u_j)g(u_j)du_j \tag{11}$$

A three-level model can be written with one independent variable that has a fixed effect and a random effect:

$$logit(p_{ijk}) = ln\left[\frac{p_{ijk}}{1-p_{ijk}}\right] = \beta_o + \beta_1 x_{ijk} + u_{1jk}x_{ijk} + u_{ok} + u_{ojk}\ \ (combined\ model) \tag{12}$$

Where i First level guide j Second level directory K guide level three u_{ojk}, u_{ok}The random effect coefficient associated with the independent variable x_{ijk}

The model can be explained as follows:

$$logit(p_{ij}) = \ln\left[\frac{p_{ij}}{1-p_{ij}}\right] = \beta_{ojk} + \beta_{1j}x_{ij}\ (level\ 1model)$$

$$\beta_{ojk} = \beta_{ok} + u_{ojk}\ \ (level\ 2\ model) \qquad (13)$$

$$\beta_{1j} = \beta_1 + u_{1j}\qquad (level\ 2\ model)$$

$$\beta_{ok} = \beta_o + u_{ok}\qquad (level\ 3\ model)$$

where for the model that we will be interested in studying, it is the multi-level (two-level) binary logistic regression model with a random slope and a fixed limit.

$$logit(p_{ij}) = \beta_{oj} + \beta_{1j}x_{ij}$$

$$\beta_{oj} = y_{oo} + y_{o1}z_j + u_{oj} \qquad (14)$$

$$\beta_{1j} = y_{1o} + y_{11}z_j + u_{oj}$$

$$logit\ (p_{ij}) = (y_{oo} + y_{o1}z_j + u_{oj}) + (y_{1o} + y_{11}z_j + u_{1j})x_{ij}$$

$$= y_{oo} + y_{o1}z_j + u_{oj} + y_{10}x_{ij} + y_{11}z_jx_{ij} + u_{1j}x_{ij}$$

By rearranging the equation (14),

$$logit\ (p_{ij}) = y_{oo} + y_{1o}x_{ij} + y_{o1}z_j + y_{11}z_jx_{ij} + u_{oj} + u_1x_{ij} \qquad (15)$$

$y_{11}z_jx_{ij}$represents the important part of the equation, which expresses the specific part of the internal interaction resulting from the intersection of the first and second levels. Where i=1,2,3…n (First Level Vocabulary Guide), j=1,2,3…N (Second Level Vocabulary Guide) In this model, it is assumed that the number of vocabulary in the first level is equal for all groups, meaning that $n_j = n$ $\forall_j = 1,2 \ldots . N$. x_{ij} the value of the independent variable for item i in group j (the independent variable in the first level), z_j the value of the independent variable for j (the independent variable of the second level) , y_{oo} is the mean of the categorical one, y_{1o} is the mean slope, y_{o1} regression group coefficient level (clustering) , y_{11} regression coefficient interaction across . When $y_{11} = 0$ and we will put a constraint that the mean of random effects is equal to zero Where $\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2{}_0 & \sigma_{01} \\ \sigma_{01} & \sigma^2{}_1 \end{bmatrix}\right)$

7. Penalized quasi-likelihood:[8][9][10]

Researchers Laird (1978) and Stiratelli (1984) suggested the quasi-penalty counting approach (pql), one of the iterative techniques, as a Bayes approximation. More recently, researchers (Schall 1991) and (Stiratelli 1984) have employed it (McGilchrist and Aisbett 1991). Contrarily, in this approach, the estimate process is carried out by repeatedly adjusting the various linear models, and the parameters and random effects are roughly inferred in the hierarchical models. When used on clustered binary data, pql has a tendency to underestimate components of variance with absolute fixed effects.

Suppose that unit i out of n units that affect the dependent variable (response variable) y_i consisting of levels X_i , Z_i associated with fixed and random effects.

And suppose that the random effects vector is b (q×1), then the expected value of $y_i$will have mean μ_i^b and variance $\emptyset a_i v(\mu_i^b)$ meaning that:

$$E(y_i/b) = \mu_i^b$$

$$Var(y_i/b) = \emptyset a_i v(\mu_i^b)$$

Where v(.) The specified variance function , a_i is a known constant, Ø Estimated variance parameter (dispersion parameter), which may be known or unknown. The conditional mean is related to y_i by means of a function relating $g(\mu_i^b)$ to the inverse of $h = g^{-1}$, $g(\mu_i^b)) = n_i^b = X_i^t\alpha + Z_i^tb$ and X_i , $Z_i$independent variables, α is a fixed effects vector p×1, b Random effects vector q×1, the responses vector $y_i = (y_1, \ldots\ldots, y_n)^t$, and that the arrays will contain the rows 〚$X_i{}^t$ و $Z_i{}^t$with X and Z, the conditional average achieves:

$$E(y_i/b) = h(X\alpha + Zb) \tag{16}$$

Assuming that b has a multivariate normal distribution with mean (0) and a covariance matrix D , where D=D(θ) which depends on the contrast compounds vector θ,

To estimate the value of α, we will use the quasi-likelihood function, which is written in the following form:

$$ql(\alpha,\theta) = \propto |D\ |^{-\frac{1}{2}} \int \exp[-\frac{1}{2\emptyset}\sum_{i=1}^{n} d_i(y_i; g(\mu_i^b)) - \frac{1}{2}b'D^{-1}b]\, db \tag{17}$$

Where $d_i(y_i;\ g(\mu_i^b) = -2\frac{y-u}{a_i v(u)}du$, and $d_i(y_i;\ g(\mu_i^b))$ is a measure of the deviation fit. Equation (17) can be written as follows:

$$c|D\ |^{-\frac{1}{2}} \int e^{-k(b)}\ db \tag{18}$$

Let k' and k" be partial derivatives of the first and second order with dimensions q ((q×q relative to b), and by ignoring the factorial constant c and taking the logarithm, we get:

$$ql(\alpha,\theta) \approx -\frac{1}{2}\log|D| - \frac{1}{2}\log|\ k''(\hat{b})\ | - k'(\hat{b}) \tag{19}$$

Where $\hat{b}$=$\hat{b}$ (α, θ) refer to solution , $k'(b) = -\sum_{i=1}^{n}\frac{(y_i-\mu_i^b)z_i}{\emptyset a_i v(\mu_i^b)g\prime(\mu_i^b)} + D^{-1}b = 0$

We differentiate again with respect to b:

$$k''(b) = -\sum_{i=1}^{n}\frac{z_i {z_i}^t}{\emptyset a_i v(\mu_i^b)\left[g'(\mu_i^b)\right]^2} + D^{-1} + R \tag{20}$$

$$\approx {z_i}^t wz + D^{-1}$$

D is the variance matrix, w is a diagonal matrix (n×n), which is defined as the frequency weights of the general linear model, R remaining range,

$$W=\{\emptyset a_i v(\mu_i^b)\left[g'(\mu_i^b)\right]^2\}^{-1}$$

$$R = -\sum_{i=1}^{n}(y_i - \mu_i^b)\, z_i \frac{\partial}{\partial b}[\frac{1}{\emptyset a_i v(\mu_i^b)g\prime(\mu_i^b)}] ,$$

$$E(R) = 0$$

By ignoring R and integrating equations from (19) to (20), we get:

$$ql(\alpha,\theta) \approx -\frac{1}{2}\log|I+Z^t WZD\ | - \frac{1}{2\emptyset}\sum_{i=1}^{n}(d_i\left(y_i;\mu_i^{\hat{b}}\right) - \frac{1}{2}\hat{b}^t D^{-1}\hat{b} \tag{21}$$

And that $\hat{b}$ is chosen to maximize the last two periods, and assuming that the weights of the repetition of the general linear model are of little change (or not all) as a function of the average, and we will ignore the first period and choose α to maximize the second period, and therefore $(\hat{\alpha},\hat{b})$=$(\hat{\alpha}(\theta),\hat{b}(\theta))$,

$$\hat{b}(\theta) = \hat{b}(\hat{\propto}\ (\theta)) ,$$

$$Pql \approx -\frac{1}{2\emptyset}\sum_{i=1}^{n}(d_i\left(y_i;\mu_i^{\hat{b}}\right) - \frac{1}{2}\hat{b}^t D^{-1}\hat{b}) \tag{22}$$

By differentiating with respect to α and b, we obtain the parameters of the mean:

$$\sum_{i=1}^{n}\frac{(y_i - \mu_i^b)X_i}{\emptyset a_i v(\mu_i^b)g'(\mu_i^b)} = 0 \tag{23}$$

$$\sum_{i=1}^{n}\frac{(y_i - \mu_i^b)Z_i}{\emptyset a_i v(\mu_i^b)g'(\mu_i^b)} = 0 \tag{24}$$

Fisher's technique was created by GREEN in 1987 to solve equations (23, 24) as an iterative weighted least squares problem (IWLS), which has a dependent variable and a weights matrix that is updated with each iteration. His approach was somewhat modified in this study. To take use of the similarities in Harville's naturalistic theory computations (1977) by defining vector Y as having components) $\mu_i^b(g^t)\ Y_i = n_i^b + (y_i - \mu_i^b)$ by using Fisher's method, a solution to equations (23) and (24) can be expressed as the iterative solution of the system:

$$\begin{bmatrix} X^tWX & X^tWXD \\ Z^tWX & I+Z^tWZD \end{bmatrix}\begin{bmatrix} \alpha \\ V \end{bmatrix} = \begin{bmatrix} X^tWY \\ Z^tWY \end{bmatrix} \qquad (25)$$

Where $b = Dv$

In 1977 Harville derived equation (25) as the best linear unbiased estimation for α and b:

$$Y=X\alpha+Zb+\varepsilon \qquad (26)$$

Where $\varepsilon \sim N(0, W^{-1})$ and $b \sim N(0, D)$, ε , b are independent

The estimation of α is obtained by the following formula:

$$(X^tV^{-1}X)\alpha = X^tV^{-1}Y \qquad (27)$$

$$\hat{\alpha} = (X^t\,V^{-1}X)^{-1}X^tV^{-1}Y \qquad (28)$$

Where $V=W^{-1} + ZDZ^t$

Thus, the estimate for b is as follows:

$$\hat{b} = D\hat{v} = DZ^tV^{-1}(Y - X\hat{\alpha}) \qquad (29)$$

8. Applied Side

Data were taken on anemia in children, as the dependent variable represents anemia infections (infected - uninfected) from four hospitals affiliated to the Babylon Health Department, namely (Al-Hah General Teaching Hospital - Imam Al-Sadiq Hospital (PBUH) - Marjan General Teaching Hospital - Babel Women's and Children's Hospital) which represents the third level in the analysis and according to the type of lobby (public - private), which represents the second level. (50) cases of anemia were taken from the general ward from the general ward and (20) cases from the private ward, and from Imam al-Sadiq Hospital (P) Peace, (75) sick cases were taken from the general ward and (20) sick cases from the private ward. From Marjan General Teaching Hospital, (30) sick cases were taken from the private ward and (20) sick cases from the general ward, and (40) sick cases were taken from the general ward and (25) sick cases from the general ward, and these sick cases represent the level Third, in the multi-level analysis, so that the total number of disease cases is (290) cases, and the independent variables that can affect anemia were taken (sex, age, weight, occupation, marital status, smoking, academic achievement, place of residence, infection with other diseases, blood pressure). That is, the first level contains (10) independent variables. The logistic analysis of the multilevel binary was carried out using the NCSS 2022 program, using the method of Penalized Quasi Likelihood (PQL).

Table (1): Multi level Binary Logistic regression results

Method	Random Intercept only model	Random Intercept with fixed effect model	Random Coefficient with fixed effect model
Log Likelihood	-345.8988	-22.96769	-123.5675
AIC	1323.897	788.5654	989.675
BIC	1879.887	235.8696	433.7782

Table (2): Random Intercept with fixed effect model results for Anemia

Variable		Odd ratio	Standard error	Statistics	P-value	(p_{ijk})
Sex (Male- Female)		1.46	1.88	0.21	0.98	0.1211
Age of patient	**10-14**	2.54	4.11	0.55	0.64	0.2231
	15-19	3.55	2.56	0.71	0.48	0.1245
	20-44	5.14	9.32	0.10	0.33	0.2114
	45-64	7.46	12.54	0.34	0.23	0.1175
	> 64	9.46	14.12	0.21	0.21	0.2068
Weight		0.22	0.67	2.11	0.0023	0.8684
Water source (Improved-non-improved		0.11	0.45	5.17	0.0011	0.9789
Occupation		13.55	11.90	0.14	0.19	0.0907
Marital status		6.55	10.55	0.22	0.21	0.0266
Smoking		7.43	13.45	0.31	0.36	0.0242

Residence (Urban-Rural)	0.78	0.44	2.03	0.0056	0.8878
Infection with other diseases	0.88	0.56	2.29	0.0043	0.8655
Blood pressure	20.32	16.55	0.005	0.8999	0.0011
Wealth index (poor-rich)	0.98	0.49	3.21	0.0011	0.9989
parameters estimates	Coefficient	Standard error	Z	P-value	Odd Ratio(Model)
fixed effect intercept (β_o)	-0.3356	1.8887	-0.343	0.0015	0.7767
Random Coefficient with fixed effect model (β_1)	0.5676	1.5754	0.641	0.0021	
Random fixed effect Var (u_{1jk})	0.4466	0.2242			
Random fixed effect Var (Lobby type)	0.3315	0.1134			
Intercept only Var (Hospitals)	0.6474	0.6675			

The estimated binary logistic regression equation is as follows:

$$logit(p_{ijk}) = ln\left[\frac{p_{ijk}}{1-p_{ijk}}\right] = -0.3356 + 0.5676x_{ijk} + 0.4466x_{ijk} + 0.3315 + 0.6474 \qquad (30)$$

Sig.: 0.0015 0.0021

9. Result discussion

Table (1) has three models of multi-level logistic regression, the first as the null model with a random error constant, the second with a random constant with fixed effects, and the third as random coefficients with a random constant, as the random constant model with fixed effects achieved the least comparison criteria between the models with the lowest odds ratio of (-22.96769) and less Akaiki information (AIC = 788.5654) and (BIC = 235.8696) compared to the rest of the models, and this model is more suitable for real data. Table (2) shows the significance of the multi-level binary logistic regression model, as the p-value = 0.0015, which is less than the level of significance 1%, and achieved a high odds ratio of (0.7767). The variables (age - occupation - smoking - blood pressure) are not significant and the variables (weight - water source - place of residence - other diseases - wealth index) are not significant.

References

[1] Rasbash, Fiona Steele, William J. Browne & Harvey Goldstein, (2012), "A User's Guide to MLwiN" , Version 2.26, ISBN: 978-0-903024-97-6 , Printed in the United Kingdom First Printing November 2004. Updated for University of Bristol, October 2005, February 2009 and September.

[2] Sommet, N. and Morselli, D. (2017). “Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS”. International Review of Social Psychology, 30(1), 203–218, DOI: https://doi.org/10.5334/irsp.90

[3] Kindu Kebede Million Wesenu, (2021), "Multilevel Modeling to identifying associated factors with Anemia Status of Women under Reproductive Age in Ethiopia.

[4] Jan de Leeuw and Erik Meijer, (2007), Handbook of Multilevel Analysis, Springer. Berlin Heidelberg New York Hong Kong London Milan Paris Tokyo

[5] Peter C. Austin; Juan Merlo, (2016), "Intermediate and advanced topics in multilevel logistic regression analysis", Wiley Online Library, (wileyonlinelibrary.com) DOI: 10.1002/sim.7336

[6] Md. Hasinur Rahaman Khan and J. Ewart H. Shaw, (2011),"Multilevel Logistic Regression Analysis Applied to Binary Contraceptive Prevalence Data", Journal of Data Science 9(2011), 93-110.

[7] Marc Callens & Christophe Croux (2005), "Performance of likelihood-based estimation methods for multilevel binary regression models", Journal of Statistical Computation and Simulation, Vol. (75), No.12, 1003-1017, DOI: 10.1080/00949650412331321070.

[8] Adeniyi Francis Fagbamigbe; Babatunde Bowale Bakre, (2018), "Evaluating Likelihood Estimation Methods in Multilevel Analysis of Clustered Survey Data", African Journal of Applied Statistics Vol. 5 (1), 2018, pages 351–376. DOI: http://dx.doi.org/10.16929/ajas/351.220

[9] Xihong Lin, , (), "Estimation using penalized quasi likelihood and quasi-pseudo-likelihood in Poisson mixed models", Lifetime Data Anal (2007) 13:533–544 DOI 10.1007/s10985-007-9071-zSpringer Science+Business Media, LLC

JRUCS

Journal of AL-Rafidain University College for Sciences

PISSN: (1681-6870); EISSN: (2790-2293)

مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

AL- Rafidain University College

استخدام احتمالية Penalized Quasi (PQL) لتقدير النموذج اللوجستي الثنائي متعدد المستويات لتحديد عوامل فقر الدم

زينب محمد رضا	نبأ عباس محسن
zainab.reda@s.uokerbala.edu.iq	nabaa.abbas@s.uokerbala.edu.iq
قسم الاحصاء - كلية الادارة والاقتصاد - جامعة كربلاء، كربلاء، العراق	
مهدي وهاب نصر الله	
mehdi.wahab@uokerbala.edu.iq	
قسم الاحصاء - كلية الادارة والاقتصاد - جامعة كربلاء، كربلاء، العراق	

المستخلص

يستخدم تحليل الانحدار اللوجستي الثنائي متعدد المستويات (Multilevel Binary Logistic Regression Analysis) في الكثير من المجالات منها الصحية، الطبية ، الجغرافيا، الاجتماعية والتربوية ، حيث اهتم الباحثين ضمن هذا التحليل بتحديد ودراسة طبيعة العلاقة بين سلوك المفردات او وحدات الدراسة، والمتغيرات الاجتماعية والبيئية والاقتصادية في البيئات المختلفة التي يعيشون فيه وينتمون اليها، وفي مثل هذه البيانات الهرمية، في هذا البحث تم استعمال بيانات متعددة المستويات اخذت عن مرض الانيميا لدى الاطفال اذ مثل المتغير المعتمد يمثل الاصابات بالانيميا (مصاب- غير مصاب) من اربعة مستشفيات تابعة لدائرة صحة بابل وهي (مستشفى الحلة التعليمي العام – مستشفى الامام الصادق (ع) – مستشفى مرجان التعليمي العام – مستشفى بابل للنسائية والاطفال) والذي يمثل المستوى الثالث في التحليل وحسب نوع الردهة (عام – خاص) والذي يمثل المستوى الثاني وتم اخذ من مستشفى الحلة التعليمي العام (50) حالة مرضية بالانيميا من الجناح العام و (20) حالة مرضية من الجناح الخاص ، ومن مستشفى الامام الصادق (ع) السلام اخذت (75) حالة مرضية من الجناح العام و (20) حالة مرضية من الجناح الخاص. ومن مستشفى مرجان التعليمي العام اخذت (30) حالة مرضية من الجناح الخاص و (20) حالة مرضية من الجناح العام، ومن مستشفى الهاشمية اخذت (40) حالة مرضية من الجناح العام و (25) حالة مرضية من الجناح العام وهذه الحالات المرضية تمثل المستوى الثالث في التحليل متعدد المستويات ليكون مجموع الحالات الكلي للحالات المرضية (290) حالة والمتغيرات المستقلة التي ممكن ان تؤثر على الانيميا اخذت (الجنس- العمر- الوزن- المهنة- الحالة الزوجية - التدخين- التحصيل الدراسي- مكان السكن- الاصابة بأمراض اخرى – ضغط الدم) اي ان المستوى الاول فيه (10) متغيرات مستقلة. تم اجراء التحليل اللوجستي لثنائي المتعدد المستويات باستعمال برنامج NCSS 2022 باستعمال طريقة الامكان شبه الأرجحية الجزائية .

معلومات البحث

تواريخ البحث:

تاريخ تقديم البحث: 2022/12/20

تاريخ قبول البحث: 2023/3/3

تاريخ رفع البحث على الموقع: 2023/12/31

الكلمات المفتاحية:

تحليل الانحدار اللوجستي الثنائي، تحليل الانحدار اللوجستي الثنائي متعدد المستويات، شبه الارجحية الاحتمالية، الانحدار، البيانات الهرمية.

للمراسلة:

زينب محمد رضا

zainab.reda@s.uokerbala.edu.iq

https://doi.org/10.55562/jrucs.v54i1.617