*Sys Rev Pharm 2020;11(12):1956-1963*
A multifaceted review journal in the field of pharmacy

# Re-Weighted Least Squares: The Best Negative Binomial Regression Methods In Determining The Congenital Anomalies' Risk Factors

Ahmed Hamza Abood *(1), Suhair Khatan Ismail (2), Adnan Fadhil Toma Sabah (3)

(1) Department of Anatomy, Histology and Embryology/Faculty of Medicine/Karbala University/Iraq
(2) Middle Technical University / Institute of Administration Al Russafa /Iraq
(3) Statistics Department, Administration and Economics College, Kerbala University, Iraq
Corresponding author: ahmed.eabod@uokerbala.edu.iq

## ABSTRACT

Nowadays, the prevalence of congenital anomalies (CAs) is increasing. Our study aimed to assess the best method among various negative binomial regression methods that identifies the possible risk factors of CAs among infants attending the Pediatric Hospital in Karbala government, Iraq using mean square error MSE and Determination Coefficient R2 as a Comparative criterions. We did a cross-sectional retrospective study in which a review of the record checklists of a 257 neonates admitted in the hospital over a three year period (January 2016–December 2019). Entering and analyzing of the data were performed using statistical program SPSS version 21. The results showed that firstly, the best distribution for CAs risk factors is negative binomial. Secondly, IRLS has higher R2 and lower MSE values than PLS. Thirdly, the variables (Mother health, Degree of parents kinship, type and present births) were significant, with a value of less than 0.05 in all methods. We concluded that IRLS is the best negative binomial regression method that determines the congenital anomalies' risk factors.

**Keywords:** Congenital Anomalies, Risk Factors, Negative Binomial Regression, Re-weighted Least Squares, Partial Least Square

**Correspondence**:
Ahmed Hamza Abood
Department of Anatomy, Histology and Embryology/Faculty of Medicine/Karbala University/Iraq

*Corresponding author: Ahmed Hamza Abood email-address: ahmed.eabod@uokerbala.edu.iq

## INTRODUCTION

According to the definition of the World Health Organization (WHO), the congenital anomalies (CAs) are includes all the structural, behavioral, functional and metabolic disorders that are present at birth and could cause physical or mental disability, or sometimes result in fetal death. According to data published by the WHO, the CAs accounts at least 2% of all newborn babies over the world and it has a major impact on infant mortality and childhood morbidity. Moreover, the number of babies born with birth defects annually were eight millions, of them, 3.3 million die before 5 years old; 3.2 million of survivors with either mental or physical disabilities. [1]

The prevalence of CAs in Iraq is still under debate because of the low diagnostic capabilities and unreliability of medical registration. Therefore, the statistical knowledge of this problem is necessary for a real image about CAs, the image that will serve as a basis to approach and manage the medical and social implications of the CAs[2].

Our aim in this study is to focus on three points. First, to study the negative binomial model and its characteristics. Second, the estimate of parameters of the negative binomial model by using by using methods of estimation (iteratively re-weighted least squares and partial least square. Third, to compare between these methods to know the best one. To do so, we must study the response variable (dependent variable) which is a type of CAs and independent variables that represent the risk factors.

Recognizing the risk factors of CAs is a key target for prevention and genetic counseling but their determination is difficult because a CA may have different causes. Risk factors increase the risk of malformation pathology. There are many predisposing factors related to the environment, heredity, stress, etc. When dealing with risk factors for CAs, they are considered either pre-conception risk factors such as family history of both mother and father, origin, social and cultural environment or risk factors after conception. According to international statistics [3]. The chief risk factors in the determination of the CAs are

**Maternal factors** [4] [5]

- Maternal age: As the age of mother over 35 years at conception increases, the risk of the chromosomal syndromes increases
- Maternal parity: it is cited that the frequency of the CAs increases in the multiparous women specially who had more than 3briths
- Type of birth: it is mentioned in many literatures that major CAs are more frequent in twins specially monozygotic twins
- Maternal heath: it is well known that maternal health and presence of fever are important factors for the embryogenesis and favor the occurrence of CAs. More commonly, maternal infections like rubella, herpes simplex virus and toxoplasmosis increases the risk of developing CAs. Women exposed to radiation, consuming potentially teratogenic drugs and carriers of metabolic disease (diabetes) have more chance for deliver babies with chromosomal abnormities and CAs
- Obstetrics history: it is reported that CAs is adversely associated with bad obstetric history of the mother, there is study mentions that females with oligohydramnios and polyhydramnios have higher chance for deliver babies with multiple malformation syndromes, moreover, female with breech presentation of fetus frequently associated with congenital anomalies compared to female with cranial presentation.
- Degree of kinship of parents: as the kinship degree of parents' increases, the chance for developing CAs and chromosomal abnormalities is increased

**Paternal factors** [6] [7]

- Paternal age: It is estimated that advanced paternal is associated with an increased risk of CAs especially congenital heart defects, hypospadias and velopalatal defect.
- Paternal health: presence of the congenital genetic diseases of father such as Duchenne muscular dystrophy, hemophilia, diseases linked to chromosome X

## MATERIALS AND METHODS

**Negative binomial distribution** [8][9]

probability mass function (PMF) of negative binomial distribution is a probability of failure (y) observation before( r-th) of successes of Bernoulli trails, (r) is positive integer , this is the deference of states successes and failures among binomial , geometric and negative binomial , binomial distribution describes success number in (n) Bernoulli trails , while geometric distribution describes number of failure before first success , and negative binomial distribution describes number of failure before (r-th) success.

Let $y \sim NB(\alpha,\theta)$

$$p(Y=y)=\frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y+1)}\left(\frac{1}{1+\theta}\right)^{\alpha}\left(\frac{\theta}{1+\theta}\right)^{y} \quad ;\alpha,\theta\in R^{+};y=0,1,2,... \quad (1)$$

**Negative binomial regression** [10][12]

Response variable (y) has count observations distributed negative binomial, (y) values non negative integers ( 0 , 1 ,2 ,3 ,.....) , in poisson model, mean and variance are equal but negative binomial model, the mean less than variance.

Negative binomial model one of the general linear models with response variable (y) takes countable integers for any phenomena or event, parameters of negative binomial model; $\theta=\alpha\mu$

$$p(y)=p(Y=y)=\frac{\Gamma(y+1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \quad ...(2)$$

$\mu>0$ mean of (y) , $\alpha>0$ heterogeneity parameter , we can to derive from (poisson-gamma mixture)

$$ln\mu=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p \quad .....(3)$$

$:x_1,.....,x_p$ are independent variable ;

$\beta_1,.....,\beta_p$ are regression coeficients

In a random sample (n) for any subject or event takes response variable (y) and independent variables ($x_i$) and have values $(x_{1i},x_{2i},...,x_{pi})$ and $\beta=(\beta_0\beta_1.....\beta_p)^T$ as a vector of parameters.

Independent variables matrix is:

$$X=\begin{bmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{bmatrix}$$

We can re-arrangement $i^{th}$f for (x row) to get $x_i$ :

$$p(y_i)=\frac{\Gamma(y_i+1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha e^{x_i\beta}}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha e^{x_i\beta}}{1+\alpha e^{x_i\beta}}\right)^{y_i},\ i=1,2,...,n \quad ...(4)$$

To estimate the parameters (α, β) using maximum likelihood estimation, becomes maximum likelihood function:

$$L(\alpha,\beta)=\prod_{i=1}^{n}P(y_i)=\prod_{i=1}^{n}\frac{\Gamma(y_i+1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha e^{x_i\beta}}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha e^{x_i\beta}}{1+\alpha e^{x_i\beta}}\right)^{y_i} \quad ...(5)$$

Log. Maximum function:

$$lnL(\alpha,B)=\sum_{i=1}^{n}\left(y_i ln\alpha+y_i(x_i.\beta)-\left(y_i+\frac{1}{\alpha}\right)ln(1+\alpha e^{x_i\beta})+ln\Gamma\left(y_i+\frac{1}{\alpha}\right)-ln\Gamma(y_i+1)-ln\Gamma\left(\frac{1}{\alpha}\right)\right) \quad ...(6)$$

$(\alpha,\beta)$ values which maximize $lnL(\alpha,B)$ will form maximum likelihood estimates .

Variance-covariance matrix for estimates is( $\Sigma=-H^{-1}$) :H is the second derivative of log- likelihood function .

**Derivation of negative binomial distribution** [13]

$$f(y;\lambda,u)=\frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} \quad ... \quad (7)$$

The gamma mixture correlated Poisson counts, the mean of (y) conditioned on (u),poisson by conditioned mean and variance given by (u):

$$f(y;x,u)=\int_0^{\infty}\frac{e^{-(\lambda_i u_i)}(\lambda_i u_i)^{y_i}}{y_i!}g(u_i)\partial u_i \quad ... \quad (8)$$

The distribution of (y) specified by how we define g(u) for gamma model is given

$u=\exp(\varepsilon)$ ,where $ln(\mu)=xb+\varepsilon$ ,with mean equal one to gamma distribution :

$$f(y;x,u)=\int_0^{\infty}\frac{e^{-(\lambda_i u_i)}(\lambda_i u_i)^{y_i}}{y_i!}\frac{v^v}{\Gamma(v)}u_i^{v-1}e^{-vu_i}du_i \quad ... \quad (9)$$

$$=\frac{\lambda_i^{y_i}}{\Gamma(y_i+1)}\frac{v^v}{\Gamma(v)}\int_0^{\infty}e^{-(\lambda_i+v)u_i}u_i^{(y_i+v)-1}du_i \quad ... \quad (10)$$

The derivation carries by moving: $\frac{\lambda_i^{y_i}}{\Gamma(y_i+1)}\frac{v^v}{\Gamma(v)}\frac{\Gamma(y_i+v)}{(\lambda_i+v)^{y_i+v}}$ to left integral, we remain terms equal one under integral.

$$=\frac{\lambda_i^{y_i}}{\Gamma(y_i+1)}\frac{v^v}{\Gamma(v)}\Gamma(y_i+v)\left(\frac{v}{\lambda_i+v}\right)^{v}\frac{1}{v^v}\left(\frac{\lambda_i}{\lambda_i+v}\right)^{y_i}\frac{1}{\lambda_i^{y_i}} \quad ...(11)$$

$$=\frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)}\left(\frac{v}{\lambda_i+v}\right)^{v}\left(\frac{\lambda_i}{\lambda_i+v}\right)^{y_i}$$

$$=\frac{\Gamma(y_i+v)}{\Gamma(y_i+1)\Gamma(v)}\left(\frac{1}{1+\frac{\lambda_i}{v}}\right)^{v}\left(1-\frac{1}{1+\frac{\lambda_i}{v}}\right)^{y_i} \quad ..... \quad (12)$$

The gamma scale parameter is inverted (v), overdispertion parameter or heterogeneity of negative binomial.

$$f(y;\mu,\alpha) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(1-\frac{1}{1+\alpha\mu_i}\right)^{y_i} \quad ....(13)$$

Is observed from of the negative binomial given (y)

$$\Gamma(y+1) = y!\ ,\Gamma(y+1/\alpha) = (y+1/\alpha-1)!\ ,\ (1/\alpha) = (1/\alpha-1)!$$

$$\frac{\Gamma(y_i+1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)} = \frac{(y_i+1/\alpha)!}{y_i!(1/\alpha-1)!} = \binom{y_i+1/\alpha-1}{1/\alpha-1} \quad ...(14)$$

The left term of gamma functions may be re-structured to a combination and right term may be formed to a single fraction, and result in another expression of negative binomial probability.

$$f(y;\mu,\alpha) = \binom{y_i+\frac{1}{\alpha}-1}{\frac{1}{\alpha}-1}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \quad ...(15)$$

## METHODS OF ESTIMATION

**Iteratively re – weighted least squares (IRLS)**[14]

This method depends on fisher scoring and it is a part of maximum likelihood estimation that is used to estimate linear model with first derivative that is called Hessian matrix, we know the probability density function of exponential family is:

$$f(y;\theta,\Phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\alpha_i(\Phi)} + c(y_i;\Phi)\right] \quad ..... (16)$$

Where: (θi) is canonical parameter or link function b (θi) is cumulant α (φ) is scale parameter, C (yi; φ) is the normalization term.

The unique form of exponential family in the first, second derivatives of cumulant with respect to (θ) to produce mean and variance functions.

$b'(\theta_i) = \text{mean}$

$b''(\theta_i) = \text{variance}$

Probability function of general linear model is:

$f(y;\theta,\Phi)$ .....(17)

Y: response variable Location parameter: θ Φ: scale parameter

To employ the natural log of likelihood function to facilitate estimation.

Log-likelihood function can be written:

$L(\theta,\Phi;y)$

IRLS can be derived as algorithum is based on Taylor expansion of log – likelihood function:

$$0 = f(y_0) + (y_1-y_0)f'(y_0) + \frac{(y_1-y_0)^2}{2!}f''(y_0) + \frac{(y_1-y_0)^3}{3!}f'''(y_0) + \ldots$$

Can be reduced the first two terms:

$$0 = f(y_0) + (y_1-y_0)f'(y_0)$$

$$y_1 = y_0 - \frac{f(y_0)}{f'(y_0)}$$

The gradient or the fisher score is the first derivative of log – likelihood function, calculate parameter estimations of maximum likelihood function by setting to zero and solve with respect to $\beta$.

Log – likelihood function has second derivation called Hessian matrix and the Log – likelihood function is peaked more than flat, minus inverse Hessian gives variance – covariance matrix, standard errors of parameters depend on diagonal elements of matrix, which is called the information matrix.

$$U = \partial L \quad \text{and} \quad H = \partial^2 L \quad .....(18)$$

Where:

U: first derivative of log likelihood function.

H: second derivative of log likelihood function.

By employ Newton –Raphson to estimate parameters:

$$\beta_r = \beta_{r-1} - H^{-1}U \quad .....(19)$$

Where

$$H = H_{r-1} \quad \text{and} \quad U = U_{r-1}$$

By iteratively finding solution of H and U , with Newton – Raphson algorithum to estimate the model's parameters .we have Observed information matrix also Expected information matrix used in (IRLS), to find gradient(U) :

In form of exponential family, log-likelihood function is:

$$L(\theta;y,\Phi) = \sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{\alpha_i(\Phi)} + c(y_i;\Phi) \quad .....(20)$$

By chain rule we solve L, with respect to $\beta$

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n}\left(\frac{\partial l_i}{\partial \theta_i}\right)\left(\frac{\partial \theta_i}{\partial \mu_i}\right)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\left(\frac{\partial \eta_i}{\partial \beta_i}\right) \quad .....(21)$$

Solve each term:

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i \theta_i - b'(\theta_i)}{\alpha_i(\Phi)} + \sum_{i=1}^{n} \frac{y_i - \mu_i}{\alpha_i(\Phi)} \quad \text{.....(22)}$$

We get formula above by solving each term of chain $b'(\theta_i) = \mu_i$

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = V(\mu_i)\,,\ \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{V(\mu_i)} \quad \text{.....(23)}$$

Also

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial (x_i \beta_j)}{\partial \beta_j} = x_{ij} \ , \text{since} \quad \eta_i = x_i \beta_j \quad \text{.....(24)}$$

Also

$$\frac{\partial \mu_i}{\partial \eta_i} = \left[g^{-1}(\eta_i)\right]' = \frac{1}{\partial \eta_i / \partial \mu_i} = \frac{1}{g'(\mu_i)} \quad \text{.....(25)}$$

This link function derivative with respect to $\mu$, η is inverse of link function.

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{\alpha_i(\Phi)V(\mu_i)g'(\mu_i)} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{\alpha_i(\Phi)V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0 \quad \text{.....(26)}$$

Y: response variable,

$\mu$: Fitted variable,

To find the second derivative we substitute H by I

$$I = -E\left[\frac{\partial^2 L}{\partial \beta_j \partial \beta_k}\right] = E\left[\frac{\partial L}{\partial \beta_j}\frac{\partial L}{\partial \beta_k}\right] \quad \text{.....(27)}$$

$$I = \frac{\partial}{\partial \beta_j}\left[\frac{(y_i - \mu_i)x_j}{\alpha_i(\Phi)V(\mu_i)}\left(\frac{\partial}{\partial \eta}\right)_i\right] * \frac{\partial}{\partial \beta_k}\left[\frac{(y_i - \mu_i)x_k}{\alpha_i(\Phi)V(\mu_i)}\left(\frac{\partial}{\partial \eta}\right)_i\right] \quad \text{...(28)}$$

$$I = \frac{(y_i - \mu_i)^2 x_j x_k}{\{\alpha_i(\Phi)V(\mu_i)\}^2}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2 \quad \text{.....(29)}$$

Where:

$$(y_i - \mu_i)^2 = \alpha_i(\Phi)V(\mu_i)$$

Let

$$V(y_i) = \alpha_i(\Phi)V(\mu_i) = (y_i - \mu_i)^2$$

$$I = \frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2 = \frac{x_j x_k}{V(y_i)g'^2}$$

Put equations together

$$\beta_r = \beta_{r-1} - \left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\right]^{-1}\left[\frac{(y_i - \mu_i)x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i\right] \quad \text{... ...(30)}$$

Multiply two sides by I

$$\left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\right]\beta_r = \left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\right]\beta_{r-1} + \left[\frac{(y_i - \mu_i)x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i\right] \quad \text{...(31)}$$

Let W is equal

$$W = \frac{1}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2$$

Linear predictor

$$\eta_i = x_{ik}\beta_{r-1}$$

$$\left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\right]\beta_r = [X'WX]\beta_r \quad \text{.....(32)}$$

Defined W,v(y)

$$\frac{(y_i - \mu_i)x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i = \frac{(y_i - \mu_i)x_k}{\frac{1}{w}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2}\left(\frac{\partial \mu}{\partial \eta}\right)_i \quad \text{.....(33)}$$

$$\left[\frac{x_j x_k}{V(y_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\right]\beta_{r-1} = x'w\eta_i \quad \text{.....(34)}$$

Combine the terms

$$[x'wx]\beta_r = x'w\eta_i + \left[x_k w(y_i - \mu_i)\left(\frac{\partial \eta}{\partial \mu}\right)_i\right] \quad \text{.....(35)}$$

$$[x'wx]\beta_r = x'w\eta_i + \left[\frac{(y_i-\mu_i)x_k}{\frac{1}{w}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2}\left(\frac{\partial\mu}{\partial\eta}\right)_i\right] \quad \text{.....(36)}$$

Let Z

$$z_i = \eta_i + (y_i-\mu_i)\left(\frac{\partial\eta}{\partial\mu}\right)_i$$

We have:

$$[x'wx]\beta_r = x'wz$$

$$\beta_r = [x'wx]^{-1}x'wz \quad \text{.....(37)}$$

**Partial least square** [15]

The partial least squares method is a linear combination of the least square of the correlation matrix and the covariance between the independent variables and the dependent variables depends on the Cross Block.

This method presents Factor Scores as linear groups between the original independent variables used in the predictive regression model.an analysis of the (x) and (y) matrix can be done:

$$x = Tp' + K \quad (38)$$

$$y = BC' + R$$

T: matrix of factors extracted from matrix (x)

P: is the vector of the (x)

B: matrix of factors extracted from matrix (y)

C: is the vector of the (y) matrix when finding a linear combination of columns of matrix (x).

We can extract the factors of matrix (x) to find linear combinations (t) , we have to find an initial vector that is multiplied by the matrix (x) :

$$t = XA \quad (39)$$

As (A) represents a vector of random values or is the first distinct vector corresponding to the first characteristic value of the matrix $(x'yy'x)$ ,$x'y$

Represented the co-variance matrix between (x and y) as well as the linear structure of the (y) matrix represented by the vector (B), in the same way:

$$B = YC$$

C: is the first characteristic vector compared to the first characteristic root of the matrix $x'yy'x$ . This method deals about a set of components called idempotent vectors that are explained co-variance between (x,y) , independent variables can be analyzed by :

$$X = TP'$$

T: is a linear set of predictive variables, but in the form of orthogonal factors, such column contains all the independent variables present in (x) but in the form of a linear set of weights.

As for P, it is a loaded vector and it is intended as a linear set between orthogonal factors (t) and original matrix of independent variables:

$$P = X't$$

Since (t): is a column of matrix T,$T'T = I$

After finding the first Eigen vector it is subtracted from both Y and X, and this procedure is repeated until X becomes a zero matrix.

**COMPARATIVE CRITERIONS** [16][17]

**Mean square error (MSE)**

$$\text{MSE} = \frac{\text{SSE}}{n-p} \quad \text{.....(40)}$$

Where: (n) sample size, (p) number of parameters of model.

$$\text{sse} = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{.....(41)}$$

**Coefficient of determination**

Using this criterion to determine model's ability to interpret the changes of (y)

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{\theta}_i)^2 + \sum_{i=1}^{n}(\hat{\theta}_i-\bar{y})^2 + 2\sum_{i=1}^{n}(y_i-\hat{\theta}_i)(\hat{\theta}_i-\bar{y}) \quad \text{...(42)}$$

The value **coefficient of determination** $0 \geq R^2 \leq 1$

## RESULTS AND DISCUSSION

257 inpatients records in the Karbala Pediatric Hospital were reviewed and their CAs were classified according to the ICD10 classification as in table (1). The data were analyzed statistically using the statistical program SPSS to study the effect of (14) independent factors (X1-X14) as shown in the table (2) on dependent variables of (y) that are called count response variables(types of CAs) , they take symbol's numbers (0-27) .

**Table (1) shows the congenital anomalies classification depending on (ICD10)**

| number | List of congenital anomalies | Symbol (ICD 10) | |
|---|---|---|---|
| 0 | Anencephaly | Q00 | 1 |
| 1 | Mirocephalus | Q02 | 2 |
| 2 | Congenital Hydrocephalus | Q03 | 3 |
| 3 | Congenital Anomalies Of Heart And Circulatory System | Q28 | 4 |
| 4 | Mongolism | Q90 | 5 |
| 5 | Other Chromosomal Anomalies | Q91-99 | 6 |
| 6 | Cleft lip | Q36 | 7 |

| | | | |
|---|---|---|---|
| 7 | Cleft Palate | Q35 | 8 |
| 8 | Cleft Lip And Palate | Q37 | 9 |
| 9 | Spina bifid | Q05 | 10 |
| 10 | Other Anomalies Of Brain And Spinal Cord | Q06 | 11 |
| 11 | Ambiguous Genitalia | Q56 | 12 |
| 12 | Hydrocele Congenital | Q83 | 13 |
| 13 | Undescended Testis | Q53 | 14 |
| 14 | Hypospadias And Epispadias | Q64 | 15 |
| 15 | Other Anomalies Of Genito - Urinary Organs | Q52,Q54,Q55 | 16 |
| 16 | Congenital Anomalies Of The Skin | Q82 | 17 |
| 17 | Anal Stenosis | Q42 | 18 |
| 18 | Other Congenital Malformation Of The Digestive System | Q38- Q41,Q43-Q45 | 19 |
| 19 | Congenital Anomalies Of The Eye | Q15 | 20 |
| 20 | Accessory Auricle | Q17 | 21 |
| 21 | Congenital Anomalies Of Upper Limb | Q71 | 22 |
| 22 | Congenital Anomalies Of Lower Limb | Q72 | 23 |
| 23 | Other congenital malformations of face and neck | Q18 | 24 |
| 24 | Other Congenital malformations of respiratory system | Q34 | 25 |
| 25 | Congenital malformations of musculoskeletal system , not elsewhere classified | Q79 | 26 |
| 26 | Other specified congenital malformation syndromes affecting multiple systems | Q87 | 27 |
| 27 | Other congenital malformations, not elsewhere classified | Q89 | 28 |

**Table (2) independent variables and their details**

| | Independent factor | Details |
|---|---|---|
| $X_1$ | Mother age | |
| $X_2$ | Mother job | 1_house wife 2_governmental worker 3_specially worker 4_free work |
| $X_3$ | Father age | |
| $X_4$ | Degree of parents kinship | 1_ found 2_not found |
| $X_5$ | Father job | 1_doesn't work 2_governmental worker 3_specailly worker 4_ free work |
| $X_6$ | Type birth | 1_single 2_twin |
| $X_7$ | Previous births | |
| $X_8$ | Present births | 1_ live 2_dead |
| $X_9$ | Mother health | 1_fever 2_radiation 3_take medicines 4_nothing is mentioned |
| $X_{10}$ | Previous abortions number | |
| $X_{11}$ | Habitat type | 1_city 2_village |
| $X_{12}$ | Infant sex | 1_male 2_female 3_hermaphrodite |
| $X_{13}$ | Infant weight | |
| Y | Type of anomaly & symbol | Write according to tenth classification demonstration |

By using the statistical program of easy fit, we are applied goodness of fit, the results showed that data followed negative binomial distribution with parameters **(n=2 , p=0.15498)** and **(p-value =0.000)** , and the data have over dispersion (variance 74.28 and 11.47 the mean of data).

To estimate and test of hypothesis coefficients of negative binomial regression of IRLS method. We used stata program as shown in table (3). We got a result of the determination factor (R2) equal to %19 and mean square error (MSE) equal to 54.73

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_{13} = 0$$
$$H_1 : \beta_0 \neq \beta_1 \neq \ldots \neq \beta_{13} \neq 0$$

**Table (3) coefficients of regression of (IRLS)**

| (x) | Standard error | coefficients | P-value |
|---|---|---|---|
| $X_1$ | 0.1041408 | -0.0829436 | 0.427 |
| $X_2$ | 4.070355 | -2.869986 | 0.481 |
| $X_3$ | 0.0870725 | 0.0552933 | 0.526 |
| $X_4$ | 0.4748609 | -0.9558305 | 0.045 |
| $X_5$ | 1.236127 | -0.6986716 | 0.527 |
| $X_6$ | 1.279343 | 3.061872 | 0.017 |
| $X_7$ | 2.809135 | -5.0951 | 0.071 |
| $X_8$ | 1.361783 | -3.5937 | 0.009 |
| $X_9$ | 0.4265287 | -1.433926 | 0.001 |
| $X_{10}$ | 1.074158 | -0.4209224 | 0.969 |
| $X_{11}$ | 1.142004 | 0.3859155 | 0.736 |
| $X_{12}$ | 1.014451 | 0.7500499 | 0.460 |
| $X_{13}$ | 0.0007725 | 0.0011191 | 0.149 |
| Constant | 7.469959 | 23.76104 | 0.002 |

From table (3) we observed that four independent variables have significant effect (x4, x6, x8, and x9) on response variable (y).

| | | |
|---|---|---|
| $X_4$ | Degree of parents kinship | 1_doesn't work 2_governmental worker 3_specailly worker 4_ free work |
| $X_6$ | type births | |
| $X_8$ | Present births | 1_ live 2_dead |
| $X_9$ | Mother health | 1_fever 2_radiation 3_take medicines 4_nothing is mentioned |

To estimate and test of hypothesis coefficients of negative binomial regression of PLS method. We used Minitab program as shown in table (4). We got a result of the determination factor ($R^2$) equal to %15.66 and mean square error (MSE) equal to 66

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_{13} = 0$$
$$H_1 : \beta_0 \neq \beta_1 \neq \ldots \neq \beta_{13} \neq 0$$

**Table (4) coefficients of regression of (PLS)**

| (x) | Standard error | coefficients | P-value |
|---|---|---|---|
| $X_1$ | 0.0959 | -0.0783 | **0.415** |
| $X_2$ | 3.75 | -2.96 | **0.431** |
| $X_3$ | 0.0802 | 0.0443 | **0.582** |
| $X_4$ | 0.421 | -0.942 | **0.026** |
| $X_5$ | 1.14 | -0.54 | **0.637** |
| $X_6$ | 1.18 | 2.86 | **0.016** |
| $X_7$ | 2.59 | -4.32 | **0.097** |
| $X_8$ | 1.24 | -3.48 | **0.006** |
| $X_9$ | 0.394 | -1.366 | **0.001** |
| $X_{10}$ | 0.992 | -0.444 | **0.655** |
| $X_{11}$ | 1.05 | 0.40 | **0.705** |
| $X_{12}$ | 0.939 | 0.864 | **0.358** |
| $X_{13}$ | 0.000712 | 0.001059 | **0.138** |
| Constant | 6.85 | 22.90 | **0.001** |

From table (4) we observed that four independent variables have significant effect ($x_4$ ,$x_6$ ,$x_8$ ,$x_9$ ) on response variable (y).

**Table (5) coefficient determination and mean square error**

| PLS | IRLS | المعيار |
|---|---|---|
| 66 | 54.73 | **MSE** |
| %15.66 | %19 | $R^2$ |

From table (5) , (IRLS) is the best method ,it has higher ($R^2$) and less (MSE) compared with (PLS).

## CONCLUSION

The current study concluded that the best distribution for data of the risk factor of the CAs is negative binomial and IRLS is more beneficial in determining the congenital anomalies' risk factors

## REFERENCES


1. **Mossey, P. and Castilla, E.** (2001): WHO Global Registry and Database On Congenital Anomalies. pp. 15–19
2. **Thaer, M.F., and Basim, A. A**. (2020): Craniofacial Anomaly Association with the Internal Malformations in the Pediatric Age Group in Al-Fallujah City-Iraq. BioMed Research International.
3. **Kalter, H. et al.** (1983): Congenital malformations: etiologic factors and their role in prevention (first of two parts). The New England Journal of medicine. 308, 424-31
4. **Rizk, F. et al.(2014)**:Congenital Anomalies: Prevalence and Risk Factors. Universal Journal of Public Health. 2(2), 58-63. http://www.hrpub.org
5. **Ahmed, A.M. et al**.(2011): Assessment of Risk Factors for Fetal Congenital Anomalies among Pregnant Women at Cairo University Hospitals. Journal of American Science. 7(12) http://www.americanscience.org
6. **Zhu, J.L.et al**.(2005): Paternal age and congenital malformations. Hum Reprod. 20(11), 3173–7
7. **Materna-Kiryluk, A. et al.**(2009): Parental age as a risk factor for isolated congenital malformations in a Polish population. Pediatr. Perinat Epidemiol. 23(1), 29–40
8. **Fredrik, L.** (2014): "Forecasting the Business Cycle using Partial Least Square", Master of science thesis Stockholm, Sweden.
9. **Gariballa, A.A. and Hussein, E**. (2016): "A Comparison Study of Ridge Regression and Principle Component Regression with Application" Vol.3,No.8.
10. **Hikmet, et al.** (2016): "prediction of agg weight from egg quality character stics via ridge regression and regression tree methods.
11. **Hilbe, J. M**.(2011): "Negative Binomial Regression , second edition , Jet Propulsion Laboratory , California Institute of Technology and Arizona State University , Cambridge University Press.
12. **Hilbe, J .M.** (1993): log negative binomial regression as a generalized linear model, technical report COS 93/94-5-26, department of sociology, Arizona state university.
13. **Nelder, J. A ., and Lee, Y.**(1992): likelihood , quasi-likelihood and pseudo-likelihood : some comparisons , Journal of the royal statistical society , B 54 :pp(273-284) .
14. **Winkelmann, R.:** "Count Data Models, Econometric theory and an application to labor mobility, Lecture Notes in Economics and Mathematical System No.(410)
15. **Ali, G. and Azin, Z**. (2014): "comparing multiple Regression, principal comonents Analysis, partial least square Regression and Ridg Regression in predicting Environment and Natural Resurces J.12(1),P:1-21.

16. **Alin, A. (**2010): Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics,. **2**(3): p. 370-374.

17. **Brian, S.E. and Graham, D**. (2001):"applied multivariate data analysis", Oxford university- 48-69, (2001).