IRAQI Academic Scientific Journals

University of Kirkuk Journal For Administrative and Economic Science (2025) 15 (2): 394-404

UKJAES

University of Kirkuk Journal For Administrative and Economic Science

ISSN:2222-2995 E-ISSN:3079-3521
University of Kirkuk Journal For Administrative and Economic Science

Hamad Akhterkhan Saber & Hussein Mohammad Mahmood Faqe. Comparative Analysis of Beta Regression Models (BRM) with Different Link Function Using Real and Simulation Data: Identifying the Optimal Model. ***University of Kirkuk Journal For Administrative and Economic Science*** **(2025) 15 (2):394-404.**

Comparative Analysis of Beta Regression Models (BRM) with Different Link Function Using Real and Simulation Data: Identifying the Optimal Model

Akhterkhan Saber Hamad [1], Mohammad Mahmood Faqe Hussein[2]

[1,2] *University of Sulaimani, College of Administration and Economics, Sulaymaniyah, Iraq*

Akhterkhan.hamad@university.edu.iq [1]
Mohammad.faqe@university .edu.iq [2]

Abstract: The statistical framework of beta regression is a powerful tool for examining continuous response variables that are bounded within the open interval (0,1). Because of its adaptability it is widely used in fields like financial modeling healthcare analytics and social sciences. This study compares various versions of multiple beta regression in-depth using simulated and empirical datasets in order to determine which predictive model is the most accurate and dependable. Among the evaluation metrics used are the Bayesian Information Criterion Akaike Information Criterion and Mean Square Error. The analysis shows that the β-regression model with a log-log link function performs better as evidenced by the lowest AIC and BIC values and the highest log-likelihood estimates. The findings in Table 5 show that the two most important factors affecting the severity of anemia are hemoglobin (HGB) and red cell distribution width percentage (RDW percent represented by X_2). This model strikes a balance between parsimony and predictive accuracy consistently producing optimal performance across all sample sizes.

Keywords: log, probit, log-log link functions, complementary log-log transformation, beta regression modelling.

التحليل المقارن لنماذج الانحدار بيتا (BRM) مع دوال الارتباط المختلفة باستخدام البيانات الحقيقية والمحاكاة: تحديد النموذج الأمثل

أ.م.د. اخترخان صابر حمد[١]، أ.م.د. محمد محمود فقى حسين[٢]

[١،٢] *جامعة السليمانية-كلية الإدارة والاقتصاد، السليمانية، العراق*

المستخلص: يُعد الإطار الإحصائي للانحدار بيتا أداةً فعّالة لفحص متغيرات الاستجابة المستمرة المحدودة ضمن الفاصل المفتوح (٠،١). ونظرًا لقدرته على التكيف، يُستخدم على نطاق واسع في مجالات مثل النمذجة المالية، وتحليلات الرعاية الصحية، والعلوم الاجتماعية. تُقارن هذه الدراسة إصداراتٍ مختلفة من الانحدار بيتا المتعدد بشكلٍ مُعمّق باستخدام مجموعات بيانات مُحاكاة وتجريبية لتحديد النموذج التنبئي الأكثر دقةً وموثوقية. من بين مقاييس التقييم المُستخدمة معيار المعلومات البايزي (معيار المعلومات أكايكي) ومتوسط مربعات الخطأ. يُظهر التحليل أن نموذج الانحدار بيتا مع دالة ارتباط لوغاريتمي-لوغاريتمي يُحقق أداءً أفضل، كما يتضح من أدنى قيم لـ

AIC وBIC وأعلى تقديرات احتمالية لوغاريتمية. تُظهر النتائج الواردة في الجدول ٥ أن أهم عاملين يؤثران على شدة فقر الدم هما الهيموغلوبين (HGB) ونسبة عرض توزيع خلايا الدم الحمراء (مُمثلة بنسبة RDW بـ X_2). يُحقق هذا النموذج توازنًا بين الاقتصاد والدقة التنبؤية، مُنتجًا أداءً مثاليًا باستمرار في جميع أحجام العينات.

الكلمات المفتاحية: اللوغاريتم، بروبت، وظائف ربط اللوغاريتم-اللوغاريتم، التحويل اللوغاريتمي-اللوغاريتمي التكميلي، نمذجة الانحدار بيتا.

Corresponding Author: E-mail: Akhterkhan.hamad@university.edu.iq

Introduction

When modeling data that is restricted to the (0,1) intervals like rates and proportions which are commonly found in fields like environmental science and public health beta regression is especially well-suited[12]. Applying conventional regression techniques to such bounded data frequently results in inaccurate estimations and decreased modeling efficiency. By consuming a selection of link functions containing logit, probit, complementary log-log (clog-log) and log-log .Beta regression offers flexibility in capturing a wide range of data patterns thereby overcoming these difficulties. Because the appropriateness of the link function depends on the particular distributional characteristics of the data choosing the right one is essential to maximizing model accuracy and interpretability. This study compares different link functions in order to determine the best modeling strategy for simulating actual disease prevalence data.

1st: Research Problems

Illustration reliable assumptions from Analysis of health data requires precisely modeling the association between danger features and disease incidence such as anemia. Beta regression is a suitable and efficient modeling technique because anemia prevalence is usually expressed as a proportion limited within a bounded interval. However because different link functions may produce differing degrees of model performance and interpretability the choice of link function within the beta regression framework can have a substantial impact on the outcomes. The current study aims to determine the best β-regression specification for modeling illness prevalence data by examining the effects of various link functions on model performance due to this methodological challenge.

2nd: Goal of the Research:

The purpose of this study is to compare multiple link functions in order to calculate the best beta regression for analyzing bounded outcome variables. Both simulated and real-world datasets will be used to assess the performance of β-regression models which will include link functions like. A major goal is to examine the impact of different predictors especially those related to health and demographics on the prevalence of anemia in addition to comparing models. The study aims to offer insightful information that can guide the creation of focused public health initiatives by examining these determinants. Beyond improving knowledge of anemia a widespread worldwide health concern this study advances beta regression techniques paving the way for more precise and trustworthy predictive modeling.

3rd: Methodology

1- β −Regression Procedure:

A specific statistical method called beta regression is used to examine continuous response variables that are limited to the (01) interval. Because fractional data is inherently bounded like rates and proportions it works especially well for modeling such data. This includes variables that reflect limited quantities like the percentage of people in a population who display a particular trait the allocation of funds among groups or any other result that needs to fall within predetermined bounds. [1][3][4].

2- Main Characteristics of Beta Regression

A. Response Variable:

Retaining the response variable within the (0,1) interval is essential in β-regression. Researchers usually apply a transformation like adding a small constant or investigate different modeling techniques like fractional logit models or zero-inflated models to account for boundary cases that are precisely 0 or 1 in the dataset.[4][10]

B. Distribution:

The response variable is thought to have a beta distribution within the context of β-regression which is defined by two positive shape parameters γ and δ. These parameters enable the distributions shape to be changed in a variety of ways which makes it extremely adaptable and able to model a broad range of data patterns.[5][7]

$$f(x;\gamma,\delta)=\frac{x^{\gamma-1}(1-x)^{\delta-1}}{\frac{\Gamma(\gamma)\Gamma(\delta)}{\Gamma(\gamma+\delta)}} \quad \text{for} \quad 0<x<1, \qquad (1)$$

3- The Mean and Variance of the Beta Distribution:

- **The expected value**: The expected value of the β −distribution is calculated as:
$$\mu=\frac{\gamma}{\gamma+\delta} \qquad (2)$$
- The variance is expressed as:
$$\sigma^2=\frac{\gamma\delta}{(\gamma+\delta)^2(\gamma+\delta+1)} \qquad (3)$$
- The values of the shape parameters γ and δ have an impact on the β-distributions form.
- **Uniform Distribution:** Occurs when both **γ = 1** and **δ = 1**.
- **U-Shaped**: When both γ <1 and δ <1.
- **Bell-Shaped**: When both γ >1 and δ >1.
- **Skewed:** The distribution becomes skewed when the values of **γ** and **δ** are unequal.

4- β-Regressions advantages

Ideal for Proportional Data: β −Regression is more effective than linear regression when working with data restricted to the (0, 1) range.

▪**Versatility**: The model supports various link functions, offering researchers the flexibility to accurately model the relationship between forecasters and the response variable.
▪**Managing Non-Normal Data**: With the ability to model non-normal response variables, the beta distribution facilitates more accurate and reliable predictions.

5- The Link Function:

In β −Regression, a link function is utilized to define the association between the explanatory variables and the predictable value of the response variable. Some of the commonly applied link functions are:[7]

A. β −Regression Model using the Logit

Many regression models, particularly logistic regression, which is frequently used for binary outcomes, depend on the logit link function. It facilitates the modeling of the link between the outcome and its covariates by converting the event probability into a linear mixture of explanatory variables. [6][7]

$$g(p)=\log\left(\frac{p}{1-p}\right)=X\beta \qquad (4)$$

In this case, p denotes the probability or proportion, and the logit function allows the model to handle bounded outcomes while accurately capturing the relationships with the predictors.

The response variable X~β(γ, δ), which is parameterized by α and β:

$$X \sim \beta(\gamma, \delta)$$

The expected value of the β-distributed response variable is associated to the forecasters complete the logit link:

$$\text{Mean} = \frac{1}{1 + e^{X\beta}}$$

And the variance:

$$\text{Variance} = \frac{\text{Mean}(1 - \text{Mean})}{1 + \frac{1}{\psi}}$$

The parameter, ψ, controls the variability of the β-distribution. When ψ is larger, the data shows less variability, while a smaller ψ leads to greater variability. The parameters $\boldsymbol{\delta}$ are commonly assessed through MLE. By maximizing the likelihood function, which is based on the β-distributed, the coefficients are estimated[10][11].

$$\text{Likelihood}(\delta) = \prod_{i=1}^{n} \frac{X_i^{\gamma_i - 1}(1-X)^{\delta_i - 1}}{\beta(\gamma_i, \delta_i)} \qquad (5)$$

B. β-Regression Model using the Probit:

Regression models require the probit link function, particularly when using probit regression, which is frequently used to binary outcomes. It uses a new mathematical method, akin to the logit link function, to translate the likelihood that a certain event will occur given a linear combination of predictor factors[8][9]

$$\text{Expected value of } Y_i \text{ given } X_i = \phi^{-1}(X_i'\beta) = \varphi(X_i'\beta) \qquad (6)$$

The following is an expression for the β-distributed response variable's mean and variance:

$$\text{Eexpected value of}(Y) = \text{Mean} = \varphi(X\beta)$$

$$\sigma^2 = \text{Mean}(1 - \text{Mean}).\frac{1}{1+1/\varphi} \qquad (7)$$

Where: The mean, determined by φ(Xβ), φ is the dispersion parameter that he higher the value of ϕ, the lower the variance between the values, and the term $\text{Mean}(1 - \text{Mean})$ represents the variance of the underlying **β**-distributed before accounting for the dispersion.

The following is a possible construction for the beta regression model's likelihood function L given n observations:[2][3].The probit link is used to calculate the expected value for each observation X_i:

$$\text{Expected value of } Y_i \text{ given } X_i = \varphi(X_i'\beta) \qquad (8)$$

For a single observation X_i, the L is provided by:

$$L(x_i, \text{expected value}, \emptyset) = \frac{x_i^{\gamma - 1}(1-x_i)^{\delta - 1}}{\beta(\gamma, \delta)} \qquad (9)$$

In this case, α and β are functions of the dispersal parameter ϕ and μ:

$$\alpha = \text{Expected value} \times \phi \quad , \beta = (1 - \text{Expected value}) \times \phi$$

The dispersal parameter ϕ controls the variance of the β-distributed. MLE is typically used to estimate β. The beta distribution is used to determine the L function for β-regression.

$$\log L(\beta, \emptyset) = \sum_{i=1}^{n} [(\text{Mean} \times \phi - 1)\log(X_i) + ((1 - \text{Mean}_i) \times \phi - 1)\log(1 - X_i) - \log(\beta(\text{Mean}_i \times \emptyset, (1 - \text{Mean}_i \times \emptyset))]$$

C. A model of β-Regression using the Clog-log

The linear regression model and the mean of the **β**- distributed are linked by the clog-log. The expected value μ_i of X_i for a given observation i is written as: [6][7]

$$\text{Mean}_i = \emptyset^{-1}(X_i^T\beta) = 1 - \text{Exp}(-\text{Exp}(X_i'\beta)) \qquad (10)$$

We use the clog-log link function to maximize the probability function of the β -distribution in order to estimate the parameters β and ϕ .[9]

1. Parameter Estimation via Maximum Likelihood

To estimate the parameters β and ϕ, we maximize the likelihood function of the **β**-distribution under the clog-log link function.[9]

1. The mean μ_i is calculated for each observation Y_i using the clog-log link function:

$$\text{Mean}_i = 1 - e^{(-e^{(X_i'\beta)})} \quad (11)$$

2. The parameters γ and δ of the **β**-distribution are subsequently connected to the mean μ_i and the dispersion parameter ϕ in the following manner:

$$\alpha = \text{Mean}_i \times \phi \quad , \beta = (1 - \text{Mean}_i) \times \phi \quad (12)$$

3. If X_i is a single observation, the likelihood function is

$$L(X_i, \text{Mean}_i, \emptyset) = \frac{X_i^{\gamma-1}(1-X_i)^{\delta-1}}{\beta(\gamma,\delta)} \quad (13)$$

When considering n observations, the log-likelihood function is :

$$\log L(\beta, \emptyset) = \sum_{i=1}^{n} [(\gamma - 1)\log(X_i) + (\beta - 1) \times (\phi - 1)\log(1 - X_i) - \log \beta(\gamma, \delta))]$$

Where

$$\gamma = \mu_i \times \phi \quad , \delta = (1 - \text{Mean}_i) \times \phi \quad (14)$$

D. β-Regression Model with Log-Log link function

We connect the linear predictors to the Y mean using a log-log link function in the manner described below:[3][7]

$$\mu_i = \emptyset^{-1}(X_i'\beta) = e^{(-e^{(X_i'\beta)})} \quad (15)$$

The dispersion parameter, ϕ, and the beta distribution's likelihood function under the log-log link function are estimated using this method.[3][7]

(1) Using the log-log link, the expected mean μ_i is computed for every observation Y_i:

$$\mu_i = e^{(-X_i'\beta)} \quad (16)$$

(2) The dispersion parameter ϕ and the mean μ_i are used to express the α - shape parameter and β-shape parameter of the β-distribution, respectively:

$$\alpha = \mu_i \times \phi \quad , \beta = (1 - \mu_i) \times \phi \quad (17)$$

(3) Y_i: is the probability for a single observation, is calculated using

$$L(Y_i, \mu_i, \emptyset) = \frac{Y_i^{\alpha-1}(1-Y_i)^{\beta-1}}{\beta(\alpha,\beta)} \quad (18)$$

The log-likelihoods of each individual observation are summed together to find the log-likelihood for a sample with n sample size.

$$\text{Log}L(\beta, \emptyset) = \sum_{i=1}^{n}[((\text{Mean}_i \times \phi) - 1)\log(Y_i) + (((1 - \text{Mean}_i) \times \phi) - 1)\log(1 - Y_i) - \log \beta((\text{Mean}_i \times \phi), ((1 - \text{Mean}_i) \times \phi))] \quad (19)$$

The log-likelihood function is maximized in order to determine the parameter estimates for β and ϕ. Since there are no closed-form solutions, numerical optimization techniques are usually used to achieve this.[9][11]

E- The Algorithm and Development of Monte Carlo Techniques

Random collection of samples is used in Monte Carlo procedures to produce numerical results. They are particularly helpful in addressing complex problems that may be challenging to mathematically examine. An structured method for using Monte Carlo techniques can be seen below.

(1) Identify the Problem:

Clearly state the issue and the function or distribution that requires analysis. This could entail tasks like optimization, integration, or stochastic process simulation.

(2) Set Parameters:

Choose how many samples (n) to produce. Although this raises the computational needs, a bigger N usually enhances the accuracy of the output.

(3) Create Random Samples:

To generate n random samples according to the specified distribution (such as uniform or normal), use a random number generator. The sample technique will change based on the particular problem being solved.[10][11]

(4) Evaluate the Function:

Calculate the function value x_i for every random sample $f(x_i)$. This stage is particular to the problem and could involve:

- Calculating a stochastic process's results
- Simulating rewards or financial return
- Calculating integration function values

(5) Assess the Function:

For every x_i random sample: Determine the value of the function $f(x_i)$ that is pertinent to the analysis. This could involve analyzing a financial payout, computing an integral, or simulating a process.

(6) Aggregate Results: Integrate the function evaluation findings based on the analysis:

- **For integration**: Calculate a Mean of the function values:

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \qquad (20)$$

- **For optimization, use the assessments to get the optimal (minimum or maximum) function value.**

F. Estimate Statistical Measures:

To assess the precision and fluctuation of the estimations, calculate the mean and variance. For instance, use the following method to estimate the variance:

$$S^2 = \frac{\sum_{i=1}^{n}(f(x_i)-\text{Mean})^2}{n-1} \qquad (21)$$

G. Repeat as necessary:

Repeat steps 3-6 to improve the estimates, which may need doing several iterations or expanding the sample size, responsing on the level of precision needed.

H. Current Findings:

Give the final estimates, standard errors, and confidence ranges, and explain the findings in light of the initial problem.

(1) Use Case: Calculating: **π**

The estimation of π is a well-known use of Monte Carlo methods

- Make use of a unit circle to estimate the value of π.
- Choose a big n, like 10,000.
- Create N random points (x,y) with uniform x and y distributions ranging from 0 to
- Determine the number of locations that are inside the unit circle that is specified by

$$x^{\wedge}2 + y^{\wedge}2 < 1 \qquad (22)$$

➢ The collection results, or the ratio of the points inside the circle to the total number of points, can be used to determine the area:

$$\text{Calculate } \pi = 4 \times \frac{\text{Number of point inside the circle}}{\text{Sample Size}} \qquad (23)$$

➢ Results as of right now: Show the suspected value of π.

H. Constructing Statistical Model

Certain metrics are typically employed to evaluate the effectiveness of statistical models while evaluating their fit. These include log-likelihood, pseudo R^2, and the Bayesian and Akaike information criteria (BIC and AIC). Below is a summary of each them.[1][9][10]

(1) Log-Likelihood (LL)

The model's ability to explain the observed data is measured by log-likelihood. Given the model parameters, it shows the likely of witnessing the data as the logarithm of the likelihood function.

$$LL = \sum_{i=1}^{n} \log(P(y_i|\theta)) \qquad (24)$$

(2) Akaike Information Criterion (AIC)

It is an indicator employed for model select that equilibrates model fit and density.

$$AIC = -2 \times LL + 2k \qquad \textbf{(25)}$$

where k: is the of responding variables .

(3) Bayesian Information Criterion (BIC)

Though it imposes a heavier penalty on models with additional parameters, BIC is comparable to AIC. It is frequently used in Bayesian settings and helps choose models according to fit and difficulty.

$$BIC = -2 \times LL + k\log(n) \qquad \textbf{(26)}$$

where n represents the sample size.

I. Outcome and Discussion

The (n=150) samples in this dataset were taken from patients at the Children's Hospital in the Sulaymaniyah Governorate. Studying the possible variables influencing children's hemoglobin (HGB) levels is the aim of this dataset. An explanation of each variable in the dataset is provided below:

Explanatory Variables (X_1 to X_8):

X1 = WBC: Shows the proportion of white blood cells in a given amount of blood.
X2= Hemoglobin (HGB, g/dL): Red blood cells contain the protein hemoglobin, which binds and
X3=RDW Red blood cell size or volume variation is expressed by RDW% .
X_4 = PLT (Platelet Count): Indicates the quantity of platelet count, which are clotting cells.
X_5 = Serum Iron : Indicates how much iron is in the blood.
X_6= Serum Ferritin : Shows the amount of ferritin, a protein that the body uses to store iron.
X_7= Folic Acid: Indicates the amount of folic acid, often known as vitamin B9, which is essential for the creation of DNA and red blood cells.
X_8: Vitamin (V.B12): Indicates the amount of vitamin B12, which is necessary for the development of red blood cells and neurological function.

First: Test the anemia severity response variable (Y).

The response variable (Y) is examined for fit to the β −distribution. We use the χ^2, A-D, and K-S tests to determine whether the data fits the beta distribution, as indicated in the table below:

The hypothesis test:

H_0: A beta distribution fits the data.
H_1: The data provided does not fit the ~ Beta distribution.

Table (1): A beta distribution test should be performed on the response variable.

	K-S		AD		χ^2	
□	**0.01**	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**
The basic value	0.15462	0.1289	3.9074	2.5018	16.812	12.592
Value of Statistics	0.09366		0.52936		8.7378	

- The null hypothesis cannot be rejected since the Kolmogorov-Smirnov (K-S) test statistics value (0.09366) is less than the critical value (0.15462 or 0.1289). This supports the previous conclusion that there isn't any solid proof that the data don't follow the beta distribution.
- You are unable to rule out the null hypothesis since the Anderson-Darling (AD) statistics value (0.52936) is smaller than the crucial value (3.9074 or 2.5108). This supports the former conclusion that there isn't any solid proof that the data don't follow the beta distribution.
- The null hypothesis cannot be rejected since the Chi-Squared statistic value (8.7378) is below the crucial threshold (16.812 or 12.592). This supports the earlier conclusion that the data does not significantly deviate from the beta distribution.

Second: Identify the Multicollinearity problem

We looked into the multicollinearity issue with the model using the variance inflation factor (VIF). according to Table 2.

The test for hypotheses:

H_0: There is no multicollinearity.
H_1: There is multicollinearty.

Table (2): Detect multicollinearty problem use VIF test

Variables	Value of Variance Inflation Factor(VIF)	Decisions
X_1	2.26	< 5
X_2	2.18	< 5
X_3	1.53	< 5
X_4	1.23	< 5
X_5	1.14	< 5
X_6	1.13	< 5
X_7	1.13	< 5
X_8	1.12	< 5

There is no problem with multicollinearity in the data because each variable's VIF value is less than 5. The absence of multicollinearity indicates that most of the variance in $\mathbf{X_j}$ is unrelated to the other predictors when the 1/VIF ratio is getting close to 1.

Third: test the problem of heteroskedasticity.

The following is how we applied the Breusch-pagan-Cook-Weisberg (BPCW) test to investigate the heteroskedasticity problems:

Test of Hypothesis:

H_0: The data follows homoskedastic
H_a: The data follows heteroskedastic

Table (3): Detect for the heteroskedastic Problem

Test	χ^2-10	P-Value
The White Test	53.19	0.1614

According to Table 3, White's test has a p-value greater than 0.01.This demonstrated that heteroskedasticity does not exist.

Fourth: Models of Beta Regression Calculated Using Various Link Functions:

Table 4 evaluates the Beta Regression model using several kinds of link functions to determine which model is best suited for this data.

Table (4): Estimation the Parameters of Different Types of Beta Regression model

Variables	Different Link function of $\beta -$ Regression			
	Logit	Probit	Clog-log	Log-log
X_1	0.00213	0.000137	0.00346	-0.001
X_2	0.5623	0.31156	0.50525	0.25991
X_3	-0.0517	-0.0219	-0.0553	-0.0119
X_4	-0.0005	-0.0002	-0.0005	-1E-04
X_5	0.00126	0.00051	0.00141	0.00023
X_6	0.00051	0.00012	0.00079	-7E-05
X_7	0.00386	0.00197	0.00358	0.00134
X_8	3.8E-05	-2E-05	0.0001	-4E-05
Constant	-8.1318	-4.612	-7.5309	-3.6485

As shown by Table 4's p-value of less than 0.01 for the LR $\chi 2$ test, four beta regression models are significant. Furthermore, the coefficients in the four $\beta -$Regression models for the constant and the variables (x_2, x_3) are significant. This may be the root cause of anemia issues. Table 5 explains how we employed goodness criteria to choose the most successful model in order to determine that the link function is the best of the models.

Table (5): Measures for choosing the beta models that use Link Functions

Link Functions	LL	AIC	BIC
Logit	298.76	-577.52	-550.42
Probit	317.31	-614.62	-587.53
Clog-log	289.651	-559.3	-532.21
Log-Log	247.196	-674.39	-647.3

The β-regression model is used to investigate the four link functions in Table 5. We determined that the model with link function (log-log) was the best fit for our data after observing that it had the lowest BIC and AIC values and the maximum log-likelihood.

Fifth: Link function-based β-Regression Model Log-log::

The **β**-Regression model with a log-log is definite as follows:
γ = -3.6485+0.2599 X_2-0.0119 X_3

Determine Mean

Take the negative exponent to calculate μ:

Mean=$e^{(-e^{(-3.6485+0.2599\ X_2-0.0119\ X_3)})}$

Table (6): Measuring for assessing β-Regression with log-log Functions

Variables	Coefficients	Standard Error	Z-Test	Sig.	Confidence Interval 95% CI	
X_1	-0.001	0.002	-0.5	0.616	-0.00492	0.002916
X_2	0.2599	0.0051	51.06	0.0000*	0.249929	0.269884
X_3	-0.0119	0.0053	-2.23	0.0260*	-0.0224	-0.00146
X_4	-0.0001	0.0001	-1.25	0.212	-0.00025	0.000055
X_5	0.00023	0.0002	1.22	0.222	-0.00014	0.000608
X_6	-0.0001	0.0005	-0.12	0.903	-0.00114	0.00101
X_7	0.0013	0.0019	0.72	0.471	-0.0023	0.004972
X_8	-4E-05	0.0001	-0.64	0.524	-0.00017	0.000088
Constants	-3.6485	0.0991	-36.83	0.0000*	-3.84268	-3.45437

Coefficient Analysis

▪When X2 and X3 are both zero, α0=−3.6485 is the baseline value of the converted mean η.
▪As demonstrated by the positive and significant coefficient α2=0.2599, an increase in X_2 is linked to an ↑ in η, which in turn results in a ↓ in μ due to the log-log connection. In particular, η rises by 0.2599 units for each unit increase in X_2, indicating that μ (the anticipated proportion or rate) is probably going to fall.
▪A rise in X3 reduces η, which raises μ, in accordance with the significant and negative coefficient α_3=-0.0119. η drops by 0.0119 units for every unit increase in X_3, which causes μ to slightly increase.

Sixth: Simulate data with the Monte Carlo approach:

Use the following R code to create data using the Monte Carlo methods using the given distributions. According to my specifications, a Monte Carlo simulation framework will be employed in this application, yielding n=50,250,350 observations for each variable.

Table (7): Simulation results with varying sample sizes

Sample Size	Functions	LL	AIC	BIC
Sample Size n=50	Logit	25.6893	-31.379	-12.258
	Probit	25.6919	-31.384	-12.264
	Clog-log	25.5863	-31.173	-12.052
	Log-log	25.8016	-31.603	-12.483
Sample Size n=250	Logit	58.3475	-96.695	-66.589
	Probit	58.3481	-96.696	-66.59
	Clog-log	58.3414	-96.683	-66.577
	Log-log	58.3593	-96.719	-66.612
Sample Size n=350	Logit	122.941	-225.88	-187.3
	Probit	122.941	-225.88	-187.3
	Clog-log	122.927	-225.85	-187.27
	Log-log	122.954	225.909	187.329

Findings and Discussion:

The table compares four models (Logit, Probit, Clog-log, and Log-log) with different sample sizes (n=50, n=250, and n=350) using the performance metrics of Log-Likelihood, AIC, and BIC. With the lowest AIC and BIC values, the highest Log-Likelihood, and the best trade-off between model fit and complexity, the Log-log model consistently performs best across all sample sizes.

4th: Conclusions and Suggestions

1- Conclusions

This study demonstrates how useful beta regression models are for precisely evaluating sparse health data particularly when estimating the incidence of anemia.

a. Since it minimizes the AIC and BIC indicating a more efficient model with fewer parameters and maximizes the log-likelihood indicating a better fit to the data the beta regression model with a log-log link function is the best.

b. The variables X_2 = hemoglobin (HGB g/dL) and X_3 = red cell distribution width (RDW percent) have a significant impact on the likelihood or severity of anemia according to a β-Regression model with a log-log link function and the data from table (5).

c. The log-log link function in the β-regression model provides the best fit to the data consistently exhibiting the highest log-likelihood values across all sample sizes. Its simplicity and efficacy are demonstrated by its lowest AIC and BIC values. This comparison shows that the log-log link functions can handle extreme values and skewed data.

2- Suggestions

In interpretation of the results of the analysis the following recommendations are offered.

a. In situations where the illness prevalence involves extreme values or a skewed distribution, use the log-log link function.

b. When evaluating anemia-related health data, rely on results like AIC and BIC to choose the optimal model.

c. To find risk variables linked to the prevalence of anemia, encourage health officials to use beta models.

References

1- Cook, R. D., McCullough, B. D., & Kieschnick, R. (2008). Beta regression and related models. Journal of Statistical Planning and Inference, 138(4).1067-1076
2- Cribari-Neto, F., and Ferrari, S. L. P. (2004). To model proportions and rates, use beta regression. Journal of Applied Statistics, 31(7), 799–815.
3- Greenberg, E., and Ferris, J. A. (1966). "Beta distribution in economic modeling." 61(314), 225-238, Journal of the American Statistical Association.
4- Hastie, T. J., and Tibshirani, R. J. (2017). "Generalized Additive Models." *Monographs on Statistics and Applied Probability*,43.
5- Mohammad, S. H., and Hussein, M. M. F. (2024). Diagnosis Anemia Disease using The Partial Least Square (PLS) Models and Support Vector Machines (SVM), Journal for Kurdistani for Strategic Studies, 2(9).
6- Pereira, R., and F. Cribari-Neto (2013). "*Tested and comparison for beta regression models"*. Mathematical Methods in the Applied Sciences, 36(12), 1460–1474.
7- Salh, S. M., Faqe, M. M., & Salih, D. T. M. (2022). Estimation Parametric Regression Model dependent on (Time-To-Event) Survival Time Distributions with Application. Journal of Administrative Sciences in Iraq, 18 (74).
8- Simas, A. B., Rocha, G., & Barreto-Souza, W. (2010). *Modeling proportional data with beta regression models: A Comparative study*, Statistical Modeling. 10(3), 211–236.
9- Smith, A. B., Brown, E. F., & Jones, C. D. (2005). "Beta Regression: A Methodological Review." Statistical Modelling, 5(2), 125-144.
10- Wafa, D. A. & Faqe, M. M., (2022). Constructing a Multilevel Modeling to High-Resolution CT (HRCT) Lung in Patients with COVID-19 Infection. IRAQ JOURNAL OF STATISTICAL SCIENCES,19(2) .
11- Website : https://www.tandfonline.com/doi/abs/10.1080/0266476042000214501
12- Zou, H., and Li, R. (2008). "Regularized Beta Regression." *Journal of Statistical Planning and Inference*, 138(4), 1212-1220.