Application of Limited Dependent Variable Models to Study the Most Prognostic Factors for Thalassemia Patients in Erbil City

ID No. 493

(PP 365 - 376)

https://doi.org/10.21271/zjhs.27.2.22

Kurdistan Ibrahim Mawlood

College of Administration and Economics, Statistics Department, Salahaddin University-Erbil
kurdistan.mawlood@su.edu.krd

Received: 14/08/2022
Accepted: 27/09/2022
Published:17/04/2023

Abstract

The main purpose of this study is to compare three models known as limited dependent variable models which are the binary logit, Tobit model and the binary probit regression models. In the most fields surveys are done with limited options due to their nature, in these cases the data not provide assumptions of linear regression models. The data of this study was obtained from Erbil thalassemia center, which is the only health center specific for thalassemic patients in Erbil city, the values of two criteria measures (Bayesian information criterion BIC and Akaike Information Criterion AIC) were obtained from the estimated models for selection the best model fit for these three models. Furthermore, the results indicated that the results of the Logit and Probit models are similar, but the parameter estimations of the two models are not directly comparable. Stata V. 16 and SPSS V.25 software's were used for fitting the models.

Keywords: limited dependent variable models, binary logit, Tobit model, probit regression models, thalassemia, AIC.

1. Introduction

Models with a two-category dependent variable, such as male-female, yes-no, successful-unsuccessful, are categorical models with dependent variables represented as "0" and "1" models and called Two-ended or dummy dependent variable models. The most commonly models used to estimating the functional relationship among both dependent and independent variables in practical application are the logit and probit models. The standard least squares method (OLS) cannot be used to estimate this model if the predictor variable is unobserved or the dependent variable is binary. Alternatively, the maximum probability estimate is utilized, which necessitates assumptions regarding error distribution. Frequently, the selection is between probit model normal errors and logit model logistic errors. The two types of limited dependent variables regression models are censored and truncated regression models. Least squares estimates are skewed when the dependent variable is censored. The censored regression model or Tobit model allows us to develop consistent and asymptotically efficient predictors when censored is applied to the predictor variable.

2. Limited Dependent Variable Models

In this study three limited dependent variable models namely the binary logit, Tobit model and the and binary probit regression models are studied.

2.1 Binary Logit Mode

The binary Logit model used to estimate the probability function of a class or events with two choses. Or the dependent variable yi takes "0" and "1", although the model is called by binary logit. Logistic regression analysis is one of the most important statistical techniques that may be used in many fields of life.; When a response variable is binary or a categorical nature relationship, takes the logistic distribution function model's formula, Moreover, the predictor variables may be binary, ordinal, mixed, qualitative, or quantitative. (Mawlood, K. I. 2000). The reasons for the importance of logistical regression are: ease of use, requires a few assumptions as a model using, the logistic formulas result from a wide variety of basic assumptions about explanatory variables, and its ability to estimate regardless of the method of sampling, whether the Prospect or Retrospective (David, W & Hosmer, JR. 2013).

Consider that the response variable Y is a binary variable, and $P(Y=1)$ is dependent on a vector of predictor variables $\vec{x}$. The object is to model

$$p(\vec{x}) \equiv P(Y=1 \mid \vec{x})$$

However the response variable Y is binary, modeling $p(\vec{x})$ means modeling $E(Y \mid \vec{x})$, if model $p(\vec{x})$ modeled as a linear function of explanatory variables, $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

In that case, the model may produce estimated probability values that exceed a certain range [0,1]. Therefore, to overcome that, the logistic function can be utilized, it is assuming that:

$$p(\vec{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}$$

Where $x_1, \ldots, x_p$ represent explanatory variables, $p(\vec{x})$ lies between [0,1] or satisfy the probability condition, and by making transformation for $p(\vec{x})$ we obtain another function: (David, W & Hosmer, JR. 2013)

$$\log\left(\frac{p(\vec{x})}{1 - p(\vec{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \qquad \ldots(1)$$

This model is called the logistic regression model.

Estimation of coefficients or parameters, ($\beta_0, \beta_1, \ldots, \beta_p$) in logistic regression can be made by maximum likelihood method (Mawlood K. I., 2019).

$$\prod_{i=1}^{n} \left\{ p(\vec{x})^{\;y_i} \left[1 - p(\vec{x}_i)\right]^{1-y_i} \right\}$$

$$\log_e \mathrm{L}(\beta) = \sum_{i=1}^{n} \mathrm{Y_i}\,(X_i'\beta) - \sum_{i=1}^{n} \log_e [1 + \exp(X_i'\beta)]$$

The maximum likelihood estimators are calculated by solving numerically, by maximizing the log likelihood. We get the equations are nonlinear in the parameters by taking the first derivative of the log maximum likelihood equation and then equating it to zero. The solution can be estimated numerically, so we use the Newton Rafson iterative method, which after a few cycles of progression formed appropriate estimates of the parameters (Hosmer, D. & Lemeshow, S., 2000).

2.2 Tobit Model

Tobin originally used the Tobit model, which is defined as a model in which the dependent variable has a lower or upper limit, to examine household expenditures by focusing on durable consumer items, despite the fact that expenditures cannot be negative. Tobin's approach was to adjust the probability function to represent the probability of unequal sampling for each observation based on whether the latent dependent variable rises or falls below a certain threshold. The standard tobit model (Tobin, J., 1958) was created to account for censoring in the dependent variable and to prevent the bias that occurs with assumption a linear functional form in the presence of censoring. Later on, this standard Tobit model is developed. When the y variable is bound (or censored) from above or below, the standard Tobit Model forms as (Chay, K. Y. & Powell, J. L., 2001):

$y_i^* = x_i'\beta + u_i$ $\qquad u_i \sim \mathrm{N}(0,\sigma^2)$ $\quad$ i =1,………..,n

$y_i = y_i^*$ if $y_i^*>0$

$y_i = 0$ otherwise

where:

x_i : is a vector of predictors for the ith household,

y_i : are observed costs by the ith household,

y_i^* : is an unobserved continuous dependant variable supposed to determine the value y_i.

Maximum likelihood estimation is used as a standard estimator for various types of models (MLE). Under proper assumptions for instance homoscedasticity and normality of an error terms, MLE yields consistent estimates of the parameters of the tobit model.

The Tobit model utilizes a censored predictor variable that is censored to a certain value. $y > \tau$ is an indicator variable equal to 1, the observation is uncensored. If $y = \tau$, that is, the observation is censored, then it is equal to 0. The Tobin model's likelihood function is as follows (Powell, J. L., 1986):

$$L(\beta,\sigma^2)=\prod_{i=1}^{n}\left[\frac{1}{\sigma}\phi\left(\frac{(y_i-x_i\beta)}{\sigma}\right)\right]^{di}-\left[1-\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]^{1-di}$$

The log likelihood function is:

$$\ln L(\beta,\sigma^2)=\sum_{i=1}^{n}d_i\left[-\ln\sigma+\ln\phi\left(\frac{(y_i-x_i\beta)}{\sigma}\right)\right]-\sum_{i=1}^{n}(1-d_i)\ln\left[1-\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \qquad \ldots(2)$$

There are two parts to the log-likelihood in this case. The first part relates to the standard regression for uncensored observations, whereas the part two relates to the censoring variables.

when $\tau = 0$, the censored model has three expected values (Cameron, A., (2011):

The expected value of the latent variable y_i^* is $E(y_i^*)=\beta x_i$. Estimated value of $(y|y>0)$ is $(y|y>0)=\beta x_i+\ \sigma\lambda(\alpha)$, where $\alpha=(\tau-x_i\beta)/\sigma_u$ and then we can find that the expected of y is $E(y)=\beta x_i+\ \Phi((\beta x_i)/\sigma)\,[\beta x_i+\ \sigma\lambda(\alpha)]$

2.3. Probit Model

The most evident problem in the linear probability model, that is one of the qualitative preference models with qualitative variables that can have two values, is that the predicted probability values are not in the range of "0" and "1". The probit model is one of the models that used to solve this problem. In terms of coefficients, this model is nonlinear, permitting the probabilities to remain between "0" and "1", and assumes that y_i are statistically independent, there is no exact or multicollinearity among all the predictor variables. When the response variable yi is binary, Pi is represented as follows (Amemiya, T., 1981).

$$P_i=E\left(y={}^{1}/_{x_i}\right)=\Phi((x_i\beta)$$

Where Φ is the standard normal distribution's cumulative distribution function and β its maximum likelihood coefficients. The basic dependent variable is assumed to be normally distributed in the probit model, but the y dependent variable is assumed to be based on the logistic curve in the y dependent variable. As a result, the tail regions of these two models' logit cumulative distribution functions are bigger than in the probit model. The probit model, often known as the "normit model" in the literature, employs the cumulative normal distribution function. Because the probit model is based on [McFadden 1973] utility theory, it is reliant on the unobservable utility index (I_i).

Whenever fitted as a model with a latent variable the probit probability model based on the normal cumulative distribution function represents as in the following equation:

$$y_i^* = I_i = \alpha + \beta x_i$$

Where x_i can be observed but y_i^* is not, if $y_i = 1$ then $y_i^* > 0$, and if $y_i^* < 0$ then $y_i = 0$.

The result of the variable y, depends on the value of τ used as the threshold value which is usually taken as "0" however another integer value can be used instead of zero. It indicates that if y_i exceeds the value y_i^* , the event will occur; if it cannot, the event will not occur, this means that:

$$y_i = \begin{cases} 1, & y_i^* > 0 \\ 0, & otherwise \end{cases} \qquad \ldots(3)$$

Then under assumption of normality, the case when $y_i^*$is less than or equals to y_i is determined using standard cumulative distribution functions. Furthermore, for the standard normal variable Z with mean "0" and variance "1", the definition of the cumulative distribution function is as follows (Demaris, A., 2004):

$$\phi(Z) = P(Z \leq z)$$

and:

$$P(y_i = 1) = 1 - \phi(\frac{-\alpha-\beta x_i}{\sigma})$$

$$P(y_i = 0) = \phi(\frac{-\alpha-\beta x_i}{\sigma})$$

As a result, the model can be represented as follows:

$$F^{-1}(P_i) = F^{-1}(I_i) = \alpha + \beta x_i \qquad \ldots(4)$$

Weighted least square method and maximum likelihood method can be used to estimate binary probit model coefficients.

2.4 Evaluating the models performance

T, F tests, and residuals are used in linear regression analysis to test the coefficients and the model. Ordinary least square is used to fit the model. In limited dependent variable models, the situation is different, The likelihood ratio test, approximation chi-square and z tests are used.

to test the hypothesis:

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_p = 0$$

H_1: at Least two of them are not equal zero.

Utilizing a chi-square test that is built on the difference between both estimated log likelihoods related to the two models, the test statistics represented as:

$$LR(p) = -2[LnL(\alpha) - LnL(\alpha, \beta)] \qquad \ldots(5)$$

Where:

$LnL(\alpha)$ indicates the reduced model's logarithm of likelihood function, that only includes the intercept parameter. $LnL(\alpha,\beta)$ denotes the final model's logarithm of likelihood. The LR is distributed as chi-square P, where P represents the number of explanatory variables in the model. (Archer, J. & Lemeshow, S. 2006)

To test the hypothesis:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

The test statistics is given as:

$$Z = \frac{\hat{\beta}_j}{\sqrt{\hat{V}ar(\hat{\beta}_j)}} \qquad \ldots(6)$$

Where $Z \sim^{H_0}_{approx} N(0,1)$

The deviance shows how much the saturated model's likelihood overrides the likelihood of the suggested model. The deviance will be low if the suggested model fits the data well. The deviance will be high if the suggested model does not fit the data well. A model's deviance is represented as (Hosmer, D. & Lemeshow, S. & May, R., 2007):

$$D(y,\hat{\mu}) = 2\left(\log\left(p\left(y \mid \hat{\theta}_s\right)\right) - \log\left(p\left(y \mid \hat{\theta}_0\right)\right)\right) \quad \ldots(7)$$

Where:

y : is the outcome.

$\hat{\mu}$: is the estimate of the model.

$\hat{\theta}_s$ and $\hat{\theta}_0$: are correspond to the respective parameters for the fitted saturated and suggested models. A saturated model contains the same number of parameters as training points, that is, $p = n$. As a result, it fits perfectly. Any other model may be the proposed model

$p(y \mid \theta)$: Is the likelihood of the data provided the model.

2.5 Criterion of the Model Selection

The two criteria utilized in this study to choose the best model were estimated for each model, and the model with the lowest value was chosen as the best model (Lee, T. & Wang, W. 2003).

2.5.1 Akaike's Information Criterion

The Akaike's Information Criterion (AIC) compares a number of statistical models to each other. AIC is calculated as follows:

AIC= -2log-likelihood + 2K …(8)

K is the numbers of model parameters (the number of variables in the model plus the intercept). A measure of model fit is the log-likelihood, which is normally determined from statistical output (Menard, S. 2002).

2.5.2 The Bayesian Information Criterion

One of the most well-known and often used tools for choosing statistical models is the Bayesian information criterion (BIC). Calculating each model's BIC is all that is necessary to compare them using the Bayesian information criteria; the model with the lowest BIC is determined to be the best model. (Lee, T. & Wang, W. 2003).

BIC=-2lnL +2*lnN*k …(9)

Where L is the value of the likelihood, N is the number of recorded measurements, and k is the number of estimated parameters.

3. Results and Discussions:

Is in this section three limited dependent variable models (binary logit, Tobit model and the and binary probit regression model) were used for analysis data of thalassemia patients in Erbil city. Also; A comparison of the two models was conducted after all corresponding results were presented. Two statistical measures (AIC and BIC) were used to evaluate the best model in our data. The following programs were used to analyzing the data:

1. Stata V.16.
2. SPSS V.25.

3.1 Data Collection

The data set for this study about thalassemic patients in Erbil city was collected from Erbil thalassemia center**.** The data consisted of 100 cases have been collected on thalassemic patients admitted to Erbil thalassemia center during **4** months' period; beginning from **1st** January **2021** through **31st** April**.**

Furthermore, for comparing three models, 13 variables were taken into account. The dependent variable was determined to be the type of thalassemia. Variables that are interdependent given in table (1);

Table 1. The Interdependent Variables

Variable names	Categorizations	N
Age grouped	1=1 -9 2=10-14 3=15-19 4=20-24 5=25-29 6=30-35	25 28 25 15 4 3
Gender	1=male 2=female	59 41
Education	1= Illiterate 2= Primary school 3= Secondary school 4= Diploma 5= College degree 6= Post under graduate	17 65 12 3 2 1
Marital Status	1=Single 2=Married	94 6
Weight(Binned)	1=5 -13 2=14-22 3=23-31 4=32-40 5=41-49 6=50-58 7=59 -67 8=68-74	7 14 22 18 10 22 4 3
Height (Binned)	1=10 -44 2=45-59 3=60-74 4=75-88 5=90 -104 6=105-119 7=120-134 8=135-149 9=150-164 10=165-180	2 2 7 6 16 16 14 10 19 8
Blood Group	1= B- 2= AB- 3= O- 4= A- 5= B+ 6= AB+ 7= O+ 8= A+	12 13 10 15 13 11 11 15
Duration of Receiving	1=Every month	40

Blood	2=once in two month 3=once in three month 4=once in six month	30 20 10
Hemoglobin	1=1 - 7 Mg/dl 2=7.1 - 8 Mg/dl 3=8.1 - 9 Mg/dl 4=9.1 - 10.5 Mg/dl	23 45 28 4
Iron	1=Normal 2=Abnormal	22 78
BMT	1=Yes 2=No	6 94
Place of Residence	1= Inside the City 2= Outside the Citr	61 39
Blood Quantity	1= One blood bag 2= two blood bag 3= three blood bag	56 37 7

From table 1. The results showed that most of the patients were at the age group of 10 to 14, A total of 59 patients were male and 41 patients were female. Of the 100 patients with **Thalassemia**, most of the patients (40) are need receiving blood every month, and 10 patients are need receiving blood once in six months. Only 4 patients were found to have (9.1 - 10.5 Mg/dl) hemoglobin levels. In addition, our result showed that normal level of Iron observed in 22 patients, while 78 had abnormal level of Iron. Regarding Blood Quantity, the results show most of the cases were 56 patients need only One blood bag every time need receiving blood and 7 0f them need to receiving three blood bag.

Descriptive statistics corresponding to explanatory variables are given in Table 2.

Table 2. Descriptive statistics of explanatory Variables

Covariate	N	Minimum	Maximum	Mean	Std. Deviation
Age (Binned)	100	1	6	2.54	1.290
Gender	100	1	2	1.41	.494
Level of Education	100	1	7	2.15	1.925
Marital Status	100	1	2	1.06	.239
Weight (Binned)	100	1	8	4.07	1.805
Height (Binned)	100	1	10	7.54	2.316
Blood Group	100	1	8	4.56	2.324
Receiving blood	100	1	4	2.00	1.005
Hemoglobin	100	1	4	2.13	.812
Iron	100	1	2	1.78	.936
BMT	100	1	2	1.94	.239
Place of residence	100	1	3	1.41	.514
Blood quantity	100	1	3	1.51	.628

The response or dependent variable Y is binary; that is, it can have only two possible outcomes that we denote as 0 and 1, in this study 0 represents type one and 1 represents type two Thalassemia.

3.2 model Fitting

Binary logit, tobit and probit models are used to analyze a data set from thalassemic patients in Erbil city, corresponding to **types of Thalassemia** as a response variable *Yi*.

3.2.1 Results for Logit Model

Binary logit model results are examined in Table 3, includes the chi squear statistic and the corresponding significance level test of each of the independent variables in the model, ratio of the coefficient B and its standard error. If the chi squear statistic is significant (i.e., its p-value less than 0.05) then the parameter is significant in the model. Of the independent variables **it is seen that**, (**Marital Status, Height, Iron and Blood Quantity**) are significant **and the other variables are not significant their p-values (p>0.05)**.

Table 3. Binary Logit Regression Model

Variables	Coef.	Std. Err.	z	P>t	95% Confidence Limits	
					Lower	Upper
Age	.092	.9126	.010	.920	-1.697	1.880
Gender	1.810	1.3517	1.793	.181	-.839	4.459
Level of education	1.701	1.1441	2.210	.137	-.541	3.944
Marital Status	-10.282	4.8722	4.454	.035	-19.832	-.733
Weight	-.624	.6154	1.028	.311	-1.830	.582
Height	1.479	.7253	4.157	.041	.057	2.900
Blood Group	-.308	.2488	1.528	.216	-.795	.180
Duration of Receiving Blood	-.920	.7632	1.454	.228	-2.416	.575
Hemoglobin	1.272	.6654	3.654	.056	-.032	2.576
Iron	-3.882	1.2774	9.235	.002	-6.386	-1.378
BMT	3.361	2.4647	1.860	.173	-1.469	8.192
Place of Residence	1.321	1.1087	1.419	.234	-.852	3.494
Blood Quantity	8.963	3.1882	7.903	.005	2.714	15.212
constant	-13.407	10.4712	1.639	.200	-33.930	7.117

Correlation between type of thalassemia and other variables according to binary logit model was found to be as follows:

Z = -13.407+ 0. 092 ∗ Age + 1.810∗ Gender + 1.701∗ Level of Education -10.282 ∗ Marital Status - 0.624 ∗ Weight + 1.479 ∗ Height - 0.308 ∗ Blood Group - 0.92 ∗ Duration of Receiving Blood + 1.272 ∗ Hemoglobin – 3.882 ∗ Iron + 3.361 ∗ BMT + 1.321 ∗ Place of Residence + 8.9 ∗ Blood Quantity.

3.2.2 Results for Tobit Model

Table (4) shows model fitting and parameter estimation of Tobit Regression model. The results show that five independent variables (**Marital Status, Height, Iron,** Hemoglobin **and Blood Quantity**) are significant but **and the other variables are not**.

Table 4. Tobit Regression Model

Variables	Coefficient	Std. Err.	z	P>t	95% Confidence Limits	
					Lower	Upper
Age grouped	-.013	.5467	.001	.981	-1.084	1.059
Gender	1.401	.9122	2.358	.125	-.387	3.189
Level of education	1.144	.6759	2.863	.091	-.181	2.468
Marital Status	-6.758	3.4483	3.841	.049	-13.516	.001

Weight	-.467	.3537	1.746	.186	-1.161	.226
Height	1.005	.4817	4.357	.037	.061	1.949
Blood Group	-.183	.1733	1.119	.290	-.523	.156
Duration of Receiving Blood	-.446	.5509	.655	.418	-1.526	.634
Hemoglobin	.902	.4501	4.020	.045	.020	1.785
Iron	-2.466	.9070	7.392	.007	-4.244	-.688
BMT	2.276	1.8522	1.510	.219	-1.354	5.906
Place of Residence	.838	.7595	1.217	.270	-.651	2.326
Blood Quantity	6.068	2.4305	6.233	.013	1.304	10.832
constant	-11.005	7.5057	2.150	.143	-25.716	3.706

Correlation between type of thalassemia and other variables according to Tobit model was found to be as follows:

y_i = -11.005 - 0. 013 ∗ Age + 1.401∗ Gender + 1.144∗ Level of Education -6.758 ∗
Marital Status - 0.467 ∗ Weight + 1.005 ∗ Height - 0.183 ∗ Blood Group - 0.446 ∗
Duration of Receiving Blood + 0.902 ∗ Hemoglobin – 2.466 ∗ Iron + 2.276 ∗ BMT +
0.838 ∗ Place of Residence + 6.068 ∗ Blood Quantity

3.2.3 Results for Probit Model

The estimated probit coefficients are the marginal effects of a change in independent variable on response variable type of thalassemia given in Table (5), the result indicates that only four variables (**Marital Status, Height, Iron and Blood Quantity**) are significant **and the other variables are not significant in the model.**

Table 5. Binary Probit Regression Model

Variables	coefficient	Std. Err.	z	P>t	95% Confidence Limits	
					Lower	Upper
Age grouped	0.065	.5064	0.017	0.898	-0.927	1.058
Gender	1.120	0.7652	2.143	0.143	-0.380	2.620
Level of education	0.990	0.6107	2.625	0.105	-0.207	2.186
Marital Status	-6.005	2.7174	4.883	0.027	-11.331	-0.679
Weight	-0.382	0.3483	1.203	0.273	-1.065	0.301
Height	0.874	0.4102	4.535	0.033	0.070	1.678
Blood Group	-0.182	0.1449	1.586	0.208	-0.466	0.102
Duration of Receiving Blood	-0.527	0.4423	1.420	0.233	-1.394	0.340
Hemoglobin	0.749	0.3845	3.795	0.051	-0.005	1.503
Iron	-2.265	0.7278	9.684	0.002	-3.691	-0.838
BMT	2.002	1.4611	1.877	0.171	-0.862	4.865

Place of Residence	0.792	o.6349	1.557	0.212	-0.452	2.037
Blood Quantity	5.306	1.8782	7.980	0.005	1.624	8.987
constant	-8.601	6.0942	1.992	0.158	-20.546	3.343

The relation between type of thalassemia and other variables according to probit model was found to be as follows;

Z = -8.601+ 0. 065 ∗ Age + 1.12∗ Gender + 0.99∗ Level of Education -6.005 ∗ Marital Status - 0.382 ∗ Weight + 0.874 ∗ Height - 0.182 ∗ Blood Group - 0.527 ∗ Duration of Receiving Blood + 0.749 ∗ Hemoglobin – 2.265 ∗ Iron + 2.002 ∗ BMT + 0.792 ∗ Place of Residence + 5.306 ∗ Blood Quantity

Table 6. Goodness of Fit

	Logit Model			Tobit Model			Probit Model		
	Value	df	P Value	Value	df	P Value	Value	df	P Value
Deviance	28.076			27.825			27.421		
Pearson Chi-Square	27.500	85	0.000	24.202	85	.000	25.563	85	.000
Log Likelihoodb	-14.038			-13.912			-13.710		
Akaike's Information Criterion (AIC)	56.076			55.825			55.421		
Bayesian Information Criterion (BIC)	92.549			92.297			91.893		

Table (6), determines which of the three models is more suitable for fitting our data, In the comparison of the models, the probit model had the lowest Deviance, AIC and BIC values. As a result, it can be said that probit model is better than tobit and logit models in estimating regression models, however; The Logit model and the probit model have the same significant coefficients but the Measures of the Model Selection of logite model are slightly higher than probit model.

4. Conclusions

The most important conclusions made by the current study, according to the results of model fitting a data set of thalassemic patients in Erbil city and parameter estimation of three models are:

1. The results showed that 59 patients were male and 41 patients were female, with the majority of patients falling within the age range of 10 to 14.
2. Out of the 100 patients with thalassemia, 40 require blood transfusions monthly, and 10 require blood transfusions every six months. Hemoglobin values between 9.1 and 10.5 mg/dl were only reported in 4 cases.
3. Our results indicated that while 78 patients had abnormal levels of iron, 22 patients had normal levels. Regarding blood quantity, the results show that in the majority of cases, 56 patients only need one blood bag each time they needed to receive blood, while only seven required three bags.
4. The Logit model and the probit model identified the same prognostic factors that influence in response variable type of thalassemia.
5. The Significance variables estimated in tobit model are: (Marital Status, Height, Iron, Hemoglobin and Blood Quantity).
6. After comparing the outcomes of the three regression models using the AIC and BIC criteria, it was found that the probit model had the lowest value of the AIC

and BIC criteria or was the model that was most appropriate for the data set applied in this research.

7. According to the results of probite model of this study the most significant factors that effecting on thalassemic patients for our data set are: (Marital Status, Height, Iron and Blood Quantity).

References

1. Archer, J. & Lemeshow, S. (2006) "***Goodness-of-fit test for a logistic regression model fitted using survey sample data***". The Stata Journal, Vol. 6, N0. 1, PP. 97–105.
2. Amemiya, T., (1981) "***Qualitative Response Models: A Survey***", Journal of Economic Literature, Vol. 19 No.4, 481- 536.
3. Cameron, A., (2011)"***Limited Dependent Variable Models (Brief) Binary, Multinomial, Censored, Treatment Effects***", Second Edition, Wiley, New York ,USA .
4. David, W & Hosmer, JR. (2013) "***Applied Logistic Regression***". Third Edition, John Wiley & Sons, Inc., Hoboken, New Jersey, Canada.
5. Demaris, A., (2004) "***Regression with Social Data: Modeling Continuous and Limited Response Variables***", John Wiley & Sons, Inc. Hoboken, New Jersey.
6. Hosmer, D. & Lemeshow, S. (2000). "***Applied Logistic Regression***". Second Edition, New York: Johnson Wiley & Sons, Inc.
7. Hosmer, D. & Lemeshow, S. & May, R. (2007). "***Applied Survival Analsis: Regression Modeling of Time to Event Data***" . Second Edition, Wiley, New York ,USA .
8. Lee, T. & Wang, W. (2003). " ***Statistical Methods for Survival Data Analysis***". Second Edition. Wiley, New York.
9. Mawlood, K. I. (2000):" ***The Use of Discriminant Analysis for Diagnosing the Most Effective factors in Clinical Classification for Heart Patients*** " M.Sc., Thesis in statistics/ Salahddin University-Erbil, College of Administration &Economics.
10. Mawlood, K. I. (2019):" ***Using Logistic Regression and Cox Regression Models to Studying the Most Prognostic Factors for Leukemia patients***" QALAAI ZANIST SCIENTIFIC JOURNAL, Vol. 4 No. 2, PP. 705-724.
11. Menard, S. (2002)" ***Applied Logistic Regression Analysis***" Second Edition, Sage Publication, Inc.
12. Tobin, J., (1958). "***Estimation of Relationships for Limited Dependent Variables***", Econometrica, Vol. 46 No.1, PP. 24-36.
13. Chay, K. Y. and Powell, J. L., (2001) "***Semiparametric Censored Regression Models***", Journal of Economic Perspectives, Vol. 15 No. 4, PP. 29-42.
14. Powell, J. L., (1986)."**Symmetrically Trimmed Least Squares Estimation for Tobit Models", Econometrica**", Vol. 54 No. 6, PP. 35-60.

بەکارهێنانی مۆدێلی گۆڕاوی وابەستەی سنووردار بۆ لێکۆڵینەوە لە زۆرترین هۆکاری پێشبینی بۆ نەخۆشانی تالاسیمیا لە شاری هەولێر

کوردستان ابراهیم مولود

بەشی ئامار وزانیاریەکان، کۆلیژی بەرێوەبردن وئابووری، زانکۆی سەلاحەددین-هەولێر

kurdistan.mawlood@su.edu.krd

پوختە

ئامانجی سەرەکی ئەم توێژینەوەیە بەراوردکردنی سێ مۆدێلە کە بە مۆدێلی گۆڕاوە وابەستە سنوورداره‌کان ناسراون کە بریتین لە لۆجیتی دووانەیی، مۆدێلی تۆبیت و مۆدێلی لاری دووانەیی پرۆبیت. لە زۆربەی بوارەکاندا ڕاپرسییەکان بە بژاردەی سنووردار ئەنجام دەدرێن بەهۆی سروشتی خۆیانەوە، لەم حاڵەتانەدا داتاکان گریمانەکانی مۆدێلی لاری هێڵی جێبەجێ ناکەن. داتاکانی ئەم توێژینەوەیە لە سەنتەری تالاسیمیای هەولێر وەرگیراوە، کە تاکە بنکەی تەندروستی تایبەتە بە نەخۆشانی تالاسیمی لە شاری هەولێر، بەهاکانی دوو پێوەری پێوەر (پێوەری زانیاری بەیزی BIC و پێوەری زانیاری ئەکایکی AIC) لە مۆدێلە خەمڵێنراوەکان بۆ هەڵبژاردنی باشترین مۆدێل کە گونجاوە لە نێوان ئەم سێ مۆدێلە. جگە لەوەش، ئەنجامەکان ئاماژەیان بەوە کرد کە ئەنجامی مۆدێلی لۆجیت و پرۆبیت هاوشێوەن، بەڵام خەمڵاندنی پارامێتەرەکانی دوو مۆدێلەکە ڕاستەوخۆ بەراورد ناکرێت. بۆ خەمڵاندنی مۆدێلەکان پرۆگرامە ئماریەکانی Stata V. 16 و SPSS V.25 بەکارهێنران.

کلیلی ووشەکان: مۆدێلی گۆڕاوە وابەستە سنوورداره‌کان، لۆجیتی دووانەیی، مۆدێلی تۆبیت، مۆدێلی پاشەکشەی پرۆبیت، تالاسیمیا، پێوەرەکانی زانیاری ئەکایکی .

تطبيق نماذج المتغير المعتمد المحدودة لدراسة العوامل التنبؤية لمرضى الثلاسيميا في مدينة أربيل

كوردستان ابراهيم مولود

قسم الاحصاء والمعلوماتية، كلية الادارة و الاقتصاد، جامعة صلاح الدين-اربيل

kurdistan.mawlood@su.edu.krd

ملخص

الغرض الرئيسي من هذه الدراسة هو مقارنة ثلاثة نماذج تُعرف بالنماذج المتغير المعتمد المحدودة وهي نموذج اللوغاريتم الثنائي ونموذج توبيت ونماذج الانحدار الاحتمالي الثنائي. في معظم المجالات ، يتم إجراء الاستطلاعات بخيارات محدودة بسبب طبيعتها ، وفي هذه الحالات لا تطبق البيانات افتراضات لنماذج الانحدار الخطي. تم الحصول على بيانات هذه الدراسة من مركز الثلاسيميا في أربيل ، وهو المركز الصحي الوحيد المخصص لمرضى الثلاسيميا في مدينة أربيل ، تم الحصول على قيم معيارين هما معيار معلومات بايزي BIC ومعيار معلومات اكايكي AIC من النماذج المقدرة لمرضى الثلاسيميا. لاختيار أفضل نموذج يناسب هذه النماذج الثلاثة. أشارت النتائج إلى أن نتائج نموذجي نموذج اللوغاريتم الثنائي ونموذج توبيت متشابهة ، لكن تقديرات المعلمات للنموذجين غير قابلة للمقارنة بشكل مباشر. تم استخدام برامج Stata V. 16 و SPSS V.25 لتركيب النماذج.

الكلمات المفتاحية: النماذج المتغيرة المعتمدة المحدودة ، اللوغاريتم الثنائي ، نموذج توبيت ، نماذج الانحدار الاحتمالي ، الثلاسيميا ، معيار معلومات اكايكي.