International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 2
ISSN: 2088-5334

# Constructing a Model with Binary Response to Some of the Factors Affecting the Incidence of Chronic Kidney Failure

Nabaa Naeem Mahdi[a,*]

[a] *College of Management and Economics, University of Mustansiriyah, Iraq*
*E-mail: *nabaanaeemmahdi@uomustansiriyah.edu.iq*

***Abstract*— Chronic kidney failure has become a disease of widespread diseases and increasingly in our society and must take precautions and prevention of causes and identify the most critical factors that lead to the disease. Since the binary logistic regression response (0,1) handles such as the model, the researcher took a sample consisting of 100 people between infected and non-infected chronic kidney. The researcher considered three factors that affect the disease: sex, age, and blood pressure. The model of logistic regression was constructed, and the parameter was estimated using the Maximum likelihood. The significance of the parameter was tested through the Wald test, and the full estimated model was tested with quality testing. The researchers recommend using logistics models for flexible use and application in the medical, economic, and social fields. Study of a larger number of variables affecting the disease and linear Multi-collinearly through the method of partial least square regression. Spreading awareness and culture among the community in identifying the causes of the disease and how to reduce treatment spread. We reached that the pressure is a factor, which is in line with the medical terms of interpretation that the blood pressure affecting the likelihood of developing high blood pressure leads significantly to chronic kidney failure, and the low blood pressure leads to the disease at a low rate. Besides, the sex variable also has a role in this model and the probability of injury for males more than females injured, but age did not appear significant to these parameters.**

***Keywords*— Logistic regression; odd ratio; Wald test; logit transformation; maximum likelihood.**

*Manuscript received 6 Jul. 2020; revised 24 Dec. 2020; accepted 19 Jan. 2021. Date of publication 30 Apr. 2021.*

## I. INTRODUCTION

Chronic kidney failure is a deficiency in kidney function gradually any kidney's inability to purify waste and excess fluid from the blood, which leads to increased toxins in the body because of the accumulation of dangerous levels of fluid and waste product. There are several reasons for the occurrence of this disease, including Diabetes disease, high blood pressure, excessive use of analgesic medications for pain and some antibiotics, inflammation of the recurrent urinary tract, kidney stones, and risk factors, aging, obesity, heart failure and liver disease [1]. Symptoms of the disease are not taken to feeling developed symptoms only if the patient arrived in the advanced stages, and these symptoms are fatigue, length of time and shortness of breath after a simple effort and a feeling of weakness, dizziness, and swelling of the hands as a result of fluid retention of blood and stomach upset, nausea and loss of appetite [2], [3].

The logistic regression analysis is used in epidemiological and medical studies and, through explanatory variables, is determined whether descriptive or quantitative that affect the probability of occurrence of the dependent variable [4]. This research's problem lies in the fact that many people suffer from Chronic kidney diseases and in the late time the number of patients has increased significantly, which drew the researcher's attention in finding some factors that significantly affect the occurrence of this disease [5]. Since the dependent variable takes the binary formula, the logistic regression model was the most appropriate in analyzing the data and interpreting the results to reach the study's objectives [6], [7]. The study aims to use a logistic binary regression model (0, 1) and the effect of several explanatory variables to reach the best model to determine the likelihood of developing chronic kidney disease.

## II. MATERIALS AND METHOD

In order to achieve the desired objectives of this research, the researcher adopted the descriptive and analytical characterization of the logistic regression. The researcher took up the most important characteristics with a focus on how to estimate its parameter and significant test and a test for the analysis of indicators and data derived from the research sample on the subject of the research.

### *A. Regression Model*

The regression model is an analysis and interpretation of the relationships between the dependent variable and explanatory variables. Through the mathematical model construct, this may be a linear or non-linear model. The estimation of model parameters and the model's quality test were at the beginning of determining the relationship's form. It was continued by the prediction stage on the nature of the study. However, the dependent variable may be nominal variables rather than quantitative ones. When this variable takes two values, such as (0,1), it is called binary logistic regression, and it may take more than two values. In this case, it is called multiple logistic regression, and there is another type called rank logistic regression in the form of ordered variables [8].

Using logistic regression can arrange the effects of independent variables, which helps researchers understand anything more important than other variables. It is also less sensitive to the direction of the deviations from the normal distribution, such as linear regression analysis, and discriminant analysis can also exceed the assumptions of ordinary fewer squares (OLS) [9][10]. It can be said that logistic regression is a model used to predict the probability of events or a statistical method to check the relationship between variable qualitative variables and variables or multiple explanatory variables logistic regression equation [11][12].

### *B. The Concept of Logical Regression*

Logistic regression is used to describe the relationship between response variables(Y) and influence variables according to the following formula [13]:

$$P(x) = \frac{1}{1+e^{-\alpha-\beta\, x_i}} \tag{1}$$

Whereas:

$P(x)$: Represents the probability of response

$(\alpha, \beta)$ : Represents the model parameters, and $(\beta \rangle 0)$.

$(x_i)$: Represent variables affecting as the value ranging from minus infinity to positive infinity $(-\infty \ \langle\ x_i \langle \infty)$.

It is noted that the above function is a non-linear exponential function and can be converted into linear through known and transfers, including logarithmic transformation. In 1944, Berkson found a logarithmic relationship to transform the relationship between influencing variables and the probability of a response to a linear relationship by drawing logit function. The logarithm of the odds is based on the formula below:

$$\mathrm{log\,i\,t}\ (\mathrm{P_i}) = \ln(\mathrm{e}^{\underline{\mathrm{x_i}}\underline{\beta_\mathrm{i}}}) = \underline{\mathrm{x_i}}\underline{\beta_\mathrm{i}} \tag{2}$$

### *C. Estimation parameters logistic regression*

It is worth mentioning that the estimate model parameters logit is (maximum likelihood), which is the most famous estimation method. It measures the observation probability of the number (n) of explanatory variables, which are located in the sample and represents the product of the greatest possible probability function $M.L = prob(x_1, x_2, \ldots\ldots, x_n)$ [14][15].

The researcher relies appreciation on Newton Raphson method, which is an iterative method that depends on the frequency of calculations several times to be reached to the best estimate of the parameter. The data interpretation of the phenomenon studied where to stop differences between the previous session and the subsequent very small [16][17].

To calculate the standard error of the parameters to be estimated based on the following formula:

$$S.E(\hat{\beta}_i) = h_{ii} \tag{3}$$

Whereas

$h_{ii}$ : Represent the diagonal elements of the matrix of covariant estimated by the following formula:

$$Cov\ (\hat{\beta}) = \{X\ Diag(n_i \hat{p}_i (1-\hat{p}_i)X)\}^{-1} \tag{4}$$

To test the significant parameters, we use the following tests and formulas:

$$Wald = \left[\frac{\hat{\beta}_i}{S.E(\hat{\beta}_i)}\right]^2 \tag{5}$$

The previous formula is distributed in one degree of freedom. To fully test the quality of the model, we use statistical$(R^2, F)$ linear regression in the case of the logistics model, the log-likelihood ratio, which follows the distribution of Chi-Square, is used as follows [18]:

$$\chi^2 = 2[log_e L_0 - log_e L_1] \tag{6}$$

Whereas

$L_1$ : It represents the value of the possible function that contains (i) variable.

$L_0$ : It represents the value of the possible function that contains (i-1) variable.

## III. Results and Discussion

This section describes the research sample and analysis. The researcher analyzed the data through the application of logistic regression model. This section also presents the interpretation of the data based on the desired objectives of the study.

### *A. Description of the Research Sample*

The study was conducted on a sample selected randomly from a hospital Medent al-Teb. The sample size of 100 people included people living with disease, chronic kidney failure, and non-infected. They represent variable response indexers infected (1) and non-infected (0) and has also been taking a number of influential variables. It is expected that it can affect the probability of occurrence of the disease and excluded a group of influential variables as causing the problem of multi collinearly. Variables in the study, $(x_1)$ represents sex and the second variable is$(x_2)$ represents the age where is divided into four age groups and the third variable is$(x_3)$ covers the blood pressure and was divided into high, low, and normal [19][20].

### *B. Analysis of the results*

Using the statistical program SPSS (20) were obtained the following results:

TABLE I
EXPLAIN CASE PROCESSING SUMMARY

| Unweighted Cases | | N | Percent |
|---|---|---|---|
| Selected Cases | Included in Analysis | 100 | 100 |
| | Missing Cases | 0 | .0 |
| | Total | 100 | 100 |
| Unselected Cases | | 0 | .0 |
| Total | | 100 | 100.0 |

TABLE II
EXPLAIN CATEGORICAL VARIABLES CODINGS

| | | Frequency | Parameter coding | | |
|---|---|---|---|---|---|
| | | | (1) | (2) | (3) |
| AgeX$_2$ | 1 | 15 | 1.000 | .000 | .000 |
| | 2 | 35 | .000 | 1.000 | .000 |
| | 3 | 38 | .000 | .000 | 1.000 |
| | 4 | 12 | .000 | .000 | .000 |
| Blood pressureX$_3$ | 1 | 62 | 1.000 | .000 | |
| | 2 | 12 | .000 | 1.000 | |
| | 3 | 26 | .000 | .000 | |
| Sex X$_1$ | 1 | 54 | 1.000 | | |
| | 2 | 46 | .000 | | |

Table 1 summarizes the data entered and the size of the sample studied, and the missing data. Furthermore, Table 2 refers to the study sample occurrences where the number of females is 46, and the number of males is 54. The second variable represents the age where the number of the first category is 15 people, while the second category is 35 people, the third category is 38 people, and the fourth category included 12 people.

The third variable is blood pressure; the number of people suffering from high blood pressure reached 62, and those who suffer from low blood pressure were 26 people. The last category was those who have naturally blood pressure of about 12 people.

TABLE III
EXPLAIN ITERATION HISTORY A, B, C, D

| Iteration | | -2 Log likelihood | Coefficients | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Constant | SexX1(1) | Age X2(1) | Age X2(2) | Age X2(3) | Blood pressure X3(1) | Blood pressure X3(2) |
| Step 1 | 1 | 61.066 | -1.722- | .142 | -.401- | -.083- | .260 | 3.361 | 2.599 |
| | 2 | 53.800 | -2.444- | .338 | -.880- | -.185- | .585 | 4.667 | 3.276 |
| | 3 | 52.735 | -2.831- | .511 | -1.236- | -.283- | .849 | 5.309 | 3.578 |
| | 4 | 52.682 | -2.935- | .561 | -1.339- | -.327- | .929 | 5.487 | 3.675 |
| | 5 | 52.681 | -2.941- | .563 | -1.346- | -.331- | .934 | 5.499 | 3.683 |
| | 6 | 52.681 | -2.941- | .564 | -1.346- | -.331- | .934 | 5.499 | 3.683 |

Table 3 contains some repetitive iteration of the maximum likelihood's derivative function to get a lower value to a negative function double of the maximum likelihood. It aims to obtain optimal estimation parameters for the model's negative derivative double of the maximum likelihood function. In the sixth session, the researcher gained less value where the value was 52.681. The difference was less than 0.001. Table 3 shows that the fourth, fifth, and sixth sessions were very slow changes. The estimated parameters were very similar with minor differences, and the stop was at the sixth session and considered the best parameters as minus double the logarithm of a likelihood function in the minimum value.

TABLE IV
VARIABLES IN THE EQUATION

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Lower | Upper |
| Step 1$^a$ | Sex X1(1) | -2.941- | 1.467 | 4.022 | 1 | .045 | .053 | .365 | 8.455 |
| | Age X2 | | | 3.889 | 3 | .274 | | | |
| | Age X2(1) | -1.346- | 1.524 | .781 | 1 | .377 | .260 | .013 | 5.156 |
| | Age X2(2) | -.331- | 1.426 | .054 | 1 | .816 | .718 | .044 | 11.742 |
| | Age X2(3) | .934 | 1.438 | .422 | 1 | .516 | 2.545 | .152 | 42.608 |
| | Blood pressure X3 | | | 28.677 | 2 | .000 | | | |
| | Blood pressure X3(1) | 5.499 | 1.027 | 28.672 | 1 | .000 | 244.562 | 32.671 | 1830.681 |
| | Blood pressure X3(2) | 3.683 | 1.130 | 10.625 | 1 | .001 | 39.769 | 4.343 | 364.202 |

Table 4 indicates all the parameters of the model estimated and the standard error for each parameter. The statistical of Wald for each parameter, the degrees of freedom, and parameters significance are also presented. Table 4 above also indicates the efficiency and quality of the model with Goodness of fit. The percentage of maximum likelihood use by equation (6) is also indicated. Table 4 above also presents the confidence limits of the estimated parameters and it can be written based on the values of (β) and are units (Log-odd) according to the following formula.

$$\log\left(\frac{\widehat{\mathrm{P}}}{1-\widehat{\mathrm{P}}}\right) = -2.941\ \mathrm{x}_{11} + 5.499\ \mathrm{x}_{31} + 3.683\mathrm{x}_{32}$$

Where (P̂) is the possibility of infection of the kidney failure. The estimate shows the relationship between the independent variables and the variable (logit units). The standard deviation of the transaction's column was calculated according to the equation (3) to obtain test Wald be based on the equation (5). To interpret the odds ratio Exp (B) was for the sex variable is equal to 0.053 that is the possibility of male cases of the disease greater by 0.053 from injury for females. The odds ratio for high pressure is estimated at 244.562 that is the risk of disease chronic renal failure at high blood pressure will be increased by 244.562. The low blood pressure when the researcher found that the odds ratio is equal to 39.769. Overall, the researcher found that the effect of high blood pressure was in the greater incidence of the disease.

TABLE V
OMNIBUS TESTS OF MODEL COEFFICIENTS

| | | Chi-square | Df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 71.139 | 6 | .000 |
| | Block | 71.139 | 6 | .000 |
| | Model | 71.139 | 6 | .000 |

TABLE VI
CONTINGENCY TABLE FOR HOSMER AND LEMESHOW TEST

| | | Y = 0 | | Y = 1 | | |
|---|---|---|---|---|---|---|
| | | Observed | Expected | Observed | Expected | Total |
| Step 1 | 1 | 10 | 10.686 | 1 | .314 | 11 |
| | 2 | 11 | 10.078 | 0 | .922 | 11 |
| | 3 | 5 | 5.790 | 7 | 6.210 | 12 |
| | 4 | 3 | 1.715 | 7 | 8.285 | 10 |
| | 5 | 2 | 1.068 | 9 | 9.932 | 11 |
| | 6 | 0 | .780 | 13 | 12.220 | 13 |
| | 7 | 0 | .211 | 5 | 4.789 | 5 |
| | 8 | 0 | .502 | 17 | 16.498 | 17 |
| | 9 | 0 | .170 | 10 | 9.830 | 10 |

Table 6 above shows the test for non-parametric of the he Goodness of fit model, which is based on the calculation statstica$(\chi^2)$ of the difference between observed values and expected values. Using the$(\chi^2)$ detection of deviations from the logistic model, the researcher found that the expected values are very close to the observed values. Thus, this confirms the significant test results as shown in Table 7 below.

TABLE VII
HOSMER AND LEMESHOW TEST

| Step | Chi-square | df | Sig. |
|---|---|---|---|
| 1 | 6.560 | 7 | .476 |

Table 7 above shows that the value of Hosmer and Lemeshow Test on the Goodness of fit model. The results show the significant compatibility. This is related to Table 6, where the observation values and the expected values are remarkably close.

TABLE VIII
CLASSIFICATION TABLE

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | Y | | Percentage |
| | | | 0 | 1 | Correct |
| Step 1 | Y | 0 | 24 | 7 | 77.4 |
| | | 1 | 3 | 66 | 95.7 |
| | Overall Percentage | | | | 90.0 |

Table 8 shows the percentage of correct classification, which amounted to 90 percent. It is divided into two groups of classification, which belong to ((24+66)/100 = 90%). Thus, there were only 10 percent of the samples who were classified in incorrect responses. It can be said that it is a good percentage that indicates the sample represents the data in a good manner.

## IV. CONCLUSION

The researcher shed light on the binary logistic regression analysis response through a theoretical study of the concept. The analysis and the possibility of estimating features depending on the way the maximum likelihood. By the study, the researcher reached the following model:

$$log\left(\frac{\hat{P}}{1-\hat{P}}\right) = -2.941\ x_{11} + 5.499\ x_{31} + 3.683x_{32}$$

which indicates the possibility of chronic kidney failure due to high blood pressure where the regression factor was equal to 5.499. This study showed high morale on the dependent variable and a degree of freedom. Besides, the low pressure also showed significant results, and the value of the regression factor was equal to 3.683.

Sex has had a stake in this model, but less where the regression coefficient was -2.941 the males more probability to the disease than females. Age did not show any significant effect on the disease probability. Perhaps this disease can affect all age groups, which can be seen in the survey data. Where model proves its efficiency through the test of Goodness of fit model and it is valuable$(\chi^2 = 71.139)$.

This study recommends that logistics models can be used for flexible use and application in the medical, economic, and social fields. This study shows the more significant number of variables affecting the disease and linear Multi-collinearly through Partial Least Square (PLS) regression. Awareness and culture should be spread among the community in identifying the causes of the disease and how to reduce treatment spread. Besides, the database in hospitals should be provided. The hospital should also provide technical staff capable of building this database to represent statistics for documenting data in the statistical sense and use them in medical research. The database has suffered most researchers when going to institutions and departments to take data and information. The basis of any research is the raw material of correct and accurate data.

## REFERENCES

[1] Stanifer, J. W., Maro, V., Egger, J., Karia, F., Thielman, N., Turner, E. L., & Patel, U. D. (2015). The epidemiology of chronic kidney disease in Northern Tanzania: a population-based survey. PloS one, 10(4), e0124506.

[2] Ahmed, R. M., & Alshebly, O. Q. (2019). Prediction and factors affecting of chronic kidney disease diagnosis using artificial neural

networks model and logistic regression model. Iraqi Journal of Statistical Sciences, 16(28), 140-159.

[3] Hasegawa, T., Sakamaki, K., Koiwa, F., Akizawa, T., Hishida, A., & CKD-JAC Study Investigators. (2019). Clinical prediction models for progression of chronic kidney disease to end-stage kidney failure under pre-dialysis nephrology care: results from the Chronic Kidney Disease Japan Cohort Study. Clinical and experimental nephrology, 23(2), 189-198.

[4] Pencina, M. J., Parikh, C. R., Kimmel, P. L., Cook, N. R., Coresh, J., Feldman, H. I., ... & Star, R. A. (2019). Statistical methods for building better biomarkers of chronic kidney disease. Statistics in medicine, 38(11), 1903-1917.

[5] Aitken, G. R., Roderick, P. J., Fraser, S., Mindell, J. S., O'Donoghue, D., Day, J., & Moon, G. (2014). Change in prevalence of chronic kidney disease in England over time: comparison of nationally representative cross-sectional surveys from 2003 to 2010. BMJ open, 4(9).

[6] Sharaf, H. K., Ishak, M. R., Sapuan, S. M., Yidris, N., & Fattahi, A. (2020). Experimental and numerical investigation of the mechanical behavior of full-scale wooden cross arm in the transmission towers in terms of load-deflection test. Journal of Materials Research and Technology, 9(4), 7937-7946.

[7] Chen, Z., Zhang, X., & Zhang, Z. (2016). Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. International urology and nephrology, 48(12), 2069-2075.

[8] Sharaf, H. K., Ishak, M. R., Sapuan, S. M., & Yidris, N. (2020). Conceptual design of the cross-arm for the application in the transmission towers by using TRIZ–morphological chart–ANP methods. Journal of Materials Research and Technology, 9(4), 9182-9188.

[9] Molnar, M. Z., Kalantar-Zadeh, K., Lott, E. H., Lu, J. L., Malakauskas, S. M., Ma, J. Z., ... & Kovesdy, C. P. (2014). Angiotensin-converting enzyme inhibitor, angiotensin receptor blocker use, and mortality in patients with chronic kidney disease. Journal of the American College of Cardiology, 63(7), 650-658.

[10] Sharaf, H. K., Salman, S., Dindarloo, M. H., Kondrashchenko, V. I., Davidyants, A. A., & Kuznetsov, S. V. (2021). The effects of the viscosity and density on the natural frequency of the cylindrical nanoshells conveying viscous fluid. The European Physical Journal Plus, 136(1), 1-19.

[11] Vejakama, P., Ingsathit, A., McKay, G. J., Maxwell, A. P., McEvoy, M., Attia, J., & Thakkinstian, A. (2017). Treatment effects of renin-angiotensin aldosterone system blockade on kidney failure and mortality in chronic kidney disease patients. BMC nephrology, 18(1), 1-9.

[12] Sharaf, H. K., Salman, S., Abdulateef, M. H., Magizov, R. R., Troitskii, V. I., Mahmoud, Z. H., ... & Mohanty, H. (2021). Role of initial stored energy on hydrogen microalloying of ZrCoAl (Nb) bulk metallic glasses. Applied Physics A, 127(1), 1-7.

[13] Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., ... & Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. Journal of translational medicine, 17(1), 1-13.

[14] Salman, S., Hilo, A., Nfawa, S. R., Sultan, M. T. H., & Saadon, S. (2019). Numerical Study on the Turbulent Mixed Convective Heat Transfer over 2D Microscale Backward-Facing Step. CFD Letters, 11(10), 31-45.

[15] Ashham, M. (2017). Simulation of Heat Transfer in a Heat Exchanger Tube with Inclined Vortex Rings Inserts. International Journal of Applied Engineering, 12(20), 9605-9613.

[16] Flayyih, H. H., Mohammed, Y. N., Talab, H. R., & Radhi, N. R. (2020). Integration of the system of Activity Based Costing and liability accounting. Integration, 1(2), 1-9.

[17] Alzabari, S. A. H., Talab, H. R., & Flayyih, H. H. (2019). The Effect of Internal Training and Auditing of Auditors on Supply Chain Management: An Empirical Study in Listed Companies of Iraqi Stock Exchange for the Period 2012-2015. Int. J Sup. Chain. Mgt Vol, 8(5), 1070.

[18] Ashham, M., Aliywy, A. M., Raheemah, S. H., Salman, K., & Abbas, M. (2020). Computational Fluid Dynamic Study on Oil-Water Two Phase Flow in A Vertical Pipe for Australian Crude Oil. Journal of Advanced Research in Fluid Mechanics and Thermal Sciences, 71(2), 134-142.

[19] Talab, H. R., Flayyih, H. H., & Ali, S. I. (2017). Role of Beneish M-score model in detecting of earnings management practices: Empirical study in listed banks of Iraqi Stock Exchange. International Journal of Applied Business and Economic Research, 15(23), 287-302.

[20] Hasan, R. F., & Mahdi, N. N. (2020). Robust non-parametric regression models for some petroleum products. Periodicals of Engineering and Natural Sciences, 8(1), 263-271.