Using Multinomial Logistic Regression model to study factors that affect chest pain

Assist. Prof. Dr. Samira Muhamad Salh
College of Administration and Economics
University of Sulaimani

Assist. Lecturer: Hozan Taha Abdalla
College of Administration and Economics
University of Sulaimani

Assist. Lecturer: Zhyan Mohammed Omer
College of Administration and Economics
University of Sulaimani

Abstract:

In this work Logistic Regression model was utilized which is one of the significant techniques for categorical data analysis, the purpose of this study to distinguish a use of Multinomial Logistic Regression method when model arrangements one (nominal/ordinal) response variable that has multiple classifications, regardless of whether nominal or ordinal variable. This method has been applied in medical area and for application the data for heart disease was taken, the data that contains nine variables such as (Chest Pain, Age, Sex, Cholesterol, Fasting Blood Sugar, Thalac "Maximum Heart Rate", Exercise, Oldpeak "ST Segment Depression induced by Exercise relative to rest" and Blood Pressure). Where the Chest Pain is the Response variable and the eight other variables are explanatory variables, after analyzing the data we conclude that there are many significant variables in each reference categories in the model, specifically (Thalac "Maximum Heart Rate" and Exercise) were significant in each different categories in the Multinomial Logistic Regression Model.

Keywords: Binary logistic Regression, Multinomial Logistic Regression (MLR), Generalized linear models (GLM), Categorical data, Information criterion.

استخدام نموذج الانحدار اللوجستي المتعدد الحدود لدراسة العوامل المؤثرة في آلام الصدر

أ.م.د. سميرة محمد صالح
كلية الإدارة والاقتصاد
جامعة السليمانية
Samira.muhamad@univsul.edu.iq

م.م. هوزان طه عبدالله
كلية الإدارة والاقتصاد
جامعة السليمانية
hozan.abdulla@univsul.edu.iq

م.م. زيان محمد عمر
كلية الإدارة والاقتصاد
جامعة السليمانية
zhyan.omer@univsul.edu.iq

المستخلص:

في هذه الدراسة، تم استخدام نموذج الانحدار اللوجستي الذي يعد إحدى الأساليب الهامة لتحليل البيانات (المصنفة) الفئوية، والغرض من هذه الدراسة هو التمييز بين استخدام نموذج الانحدار اللوجستي المتعدد الحدود عندما يكون متغير استجابة النموذج بترتيب (اسمي/ترتيبي) لها تصنيفات متعددة، بغض النظر عن اكان متغيرًا اسميا أو ترتيبيًا. تم تطبيق هذا النموذج في المجال الطبي للتطبيق تم أخذ البيانات الخاصة بأمراض القلب و التي تحتوي على تسعة متغيرات (ألم الصدر، العمر، الجنس، الكوليسترول، صيام سكر الدم، الحد الأقصى لمعدل ضربات القلب،

التمارين، Oldpeak انخفاض شريحة ST الناجم عن التمرين، وضغط الدم) حيث يكون ألم الصدر هو متغير الاستجابة و المتغيرات الثمانية الأخرى متغيرات توضيحية، بعد تحليل البيانات استنتجنا أن هناك العديد من المتغيرات المهمة في كل فئة مرجعية في النموذج، على وجه التحديد الحد الأقصى لمعدل ضربات القلب ويعد التمارين ضروريا في كل فئة مختلفة في نموذج الانحدار اللوجستي المتعدد الحدود.

الكلمات المفتاحية: الانحدار اللوجستي الثنائي، الانحدار اللوجستي متعدد الحدود، نموذج الخطي المعمم، البيانات المصنفة، معيار المعلومات.

1. Introduction:

In recent years, specialized statistical methods for analyzed categorical data have expanded, especially for application in biomedical and sociology. Regression analysis is one of these statistical tools that utilize the relationship between two or more variables. The regression models can be partitioned into two groups, the first related to linear relationship models, and the second related to non-linear relationship models. The linear models, considered so far, are palatable for most regression applications. Nonlinear model utilized when the linear model is not appropriate at any rate. Huge numbers of statisticians believe that the logistic regression model is one of the important models can be applied to analyze a categorical data; this model is an extraordinary instance of generalized linear models (GLM). The multinomial logistic regression (MLR) model utilized in commonly powerful where the response variable is composed of more than two levels or categories.

The essential idea was generalized from binary logistic regression. Continuous variables are not utilized as response variable in logistic regression, and only one response variable can be used. The MLR model can be used to predict a response variable based on the continuous and/or categorical explanatory variables to decide the percent of difference in the response variable clarified by the explanatory variables, to rank the general significance of independents, to survey connection impacts, and to comprehend the effect of covariate control variables. The MLR model permits the simultaneous comparison of more than one differentiation, that is, the log odds of three or more contrasts are estimated simultaneously (Garson, 2009: 461-463). The logistic regression model expects that the categorical dependent variable has just two values, all in all, 1 for progress and 0 for failure. The logistic regression model can be reached out to circumstances where the response variable has more than two values, and

there is no regular requesting of the categories. Common ordering can be treated as nominal scale, such information can be analyzed by marginally changed strategies utilized in dichotomous outcomes, and this method is called the multinomial logistic. The impact of indicator is ordinarily explained regarding of odds ratios. After transforming the dependent into a logit variable (the natural log of the odds of the dependent occurring or not) logistic regression applies maximum likelihood estimation Logistic models computes changes in the log odds of the response, not changes in the outcome itself as ordinary least square (OLS) regression does. Logistic regression has various analogies to OLS regression: logit coefficients represent to be coefficients in the logistic regression equation, the normalized logit coefficients correspond to beta weights, and a pseudo R square (R^2) statistic is accessible to sum up the quality of the relationship. In contrast to OLS regression, not withstanding, logistic regression does not expect linearity of relationship between the independent variables and the response, it is not necessary that the variables are normally distributed, does not presume homoscedasticity, and all in all has less severe prerequisites. It does, nonetheless, necessitate that observations be independent and that the independent variables be linearly related to the logit of the dependent (Abdalla, 2012: 272).

2. Literature Review:

- In 2017, Asampana, Nantomah and Tungosiamu, their focal point of this investigation is to utilize multinomial logistic regression model to research to explore the determinants of understudies' scholastic presentation in science. a simple arbitrary example of 393 understudies was chosen from a companion of first year understudies of Zamse Senior High/Technical inside the Bolgatanga Municipality. the researchers were conceded inside the 2015/2016 year term to seek after different software engineers inside the school. A poll was wont to accumulate information from the researchers. The outcomes demonstrate that the event of fantastic execution in science is essentially dependent on sex of researchers with male understudies indicating altogether great execution than female understudies. Another critical indicator of fine scholastic execution in science was the age of students; with more youthful understudies showing great scholarly execution than older students. Mother's business additionally contributes altogether to great execution in science with

understudies whose moms are utilized indicating acceptable scholastic execution than their partners whose moms don't seem to be waged.

- In 2016, Erkan intends to take a gander at the association between youngsters' work status and their segment attributes utilizing multinomial logistic regression. to the current end, data collected by TUIK's (Turkish Statistical Institute) "Child labour Survey, 2012," directed with the interest of 27,118 children, were used. At the primary stage of the analysis, eight independent variables on the demographic characteristics of the participants were examined using the chi-square test of independence, the variable that wasn't significant was removed, and therefore the subsequent analyses were directed using the remaining seven variables. The legitimacy of the model inside the investigation was examined utilize maximum likelihood estimation, and also the model was significant. Odds ratios of the variables within the model were calculated, and two category comparisons were made on the idea of the baseline category using odds ratio coefficients. In comparisons 1 and a pair of, odds ratio coefficients for the variables rustic/metropolitan, sex, age group, family unit size, proficiency, school participation, and level of instruction of the zenith of family unit were significant.
- In 2016 Erkan and Aydin focused on factors influencing the kinds of force against women decided by multinomial logistic regression model. during this specific situation, they utilized the info of "Research on force against Women in Turkey" that was applied by Turkish Statistical Institute in 2008. within the study, the variable of the kinds of force against women was used as variable that has four levels. moreover, twelve independent variables were used removing irrelevant variables from the info set via chi-square test of independence. After that, the most likelihood estimates and therefore the odds ratios of the variables of the model were obtained. Also, the legitimacy of the model was tested by likelihood ratio test. Finally, comparisons were made for 3 categories counting on the chances ratio in step with the chosen reference category. Regarding of odds ratios, the variables of "education level of woman" and "husband's work sector" were statistically significant in precisely comparison one; the variables of "agnation with husband, "education level of husband", "recurrence of seeing alcoholic husband", and "frequency of gambling of husband" were statistically significant in both comparison one and three; the variables of

"region", "deceived by husband", "common-law female for husband" were statistically significant all told comparisons.

- In 2015, Coughenour, Paz, de la Fuente-Mella and Singh, their purpose of this investigation was to grasp perceptions and likelihood of using various bicycle infrastructures for transportation by Las Vegas residents. An overview was created and administered (n = 457). Multinomial regression was accustomed create predictions to work out which foundations were seen as protected and perhaps to be utilized for transportation; frequencies were analyzed. so as to extend active transportation rates effectively, residents' perceptions of safety and infrastructure preferences should be considered. Results from this examination indicated that respondents had numerous security concerns with this bicycling framework in Las Vegas and gave thoughts to future foundation speculations and related strategies.
- In 2015 Murata, Fujii and Naitoh, , their point of this examination was to investigate the effectiveness of behavioral evaluation measures for predicting drivers' subjective drowsiness. Behavioral measures included neck vending angle (horizontal and vertical), back pressure, foot pressure, COP (Center of Pressure) movement on sitting surface, and tracking error in driving simulator task. Sluggish states were anticipated by methods for the multinomial logistic regression model where physiological and behavioral measures and subjective evaluation of drowsiness corresponded to independent variables and a variable, respectively. First, they thought about the adequacy of two techniques (correlation coefficient-based method and odds ratio-based method for deciding the request for entering behavioral measures into the expectation model it had been discovered that the expectation precision didn't contrast between the two techniques. Second, the prediction accuracy was compared among the numbers of behavioral measures. The forecast precision neglected to contrast among 4, 5, and 6 behavioral measures, and it had been reasoned that entering at least four behavioral measures into the expectation model is sufficient to acknowledge higher expectation exactness. Third, the forecast accuracy was compared between the strongly drowsy and therefore the weakly drowsy group. The expectation exactness varied between the two groups, and hence the proposed strategy was powerful (the prediction

accuracy was significantly higher) particularly under the condition where drowsiness was induced to a bigger extent.

- In 2014, Madhu, Ashok and Balasu bramanian, their study was: a. Is there a difference within the pattern of carcinoma cases in numerous socio-economic status with relevancy their zone of residence. b. Demonstrate the appliance of multinomial logistic multivariate analysis to look at the factors related to carcinoma in major league salary, middle and low pay families. Carcinoma cases reported to the Bharath Hospital and Institute of Oncology (BHIO) from 2007 to December 2011 were analyzed. Descriptive analysis like chi-square analysis and multinomial multivariate analysis is performed. MLR analysis demonstrated that Illiteracy, nulliparity, young ladies (< 40 years) having a place with family units had higher chances of carcinoma in middle and low income families contrasted with major league salary families.
- In 2007, Woo-Yong and Ditton studied about exploring the connections among variables regarding the eagerness to substitute one location for one more location. The targets of the investigation are: 1. to establish and predict the extent to which saltwater anglers were willing to substitute fishing at one location for fishing at another location; and 2. identify the link between independent variables like demographic characteristics, constraints, and anglers' specialization variables as predictors and anglers' willingness to substitute one fishing location for an additional. From the results of the multinomial logistic regression analysis, anglers' willingness to substitute was affected negatively by age, and affected positively by a restriction variable; and anglers' willingness to substitute was negatively related to specialization variables.

3. Multinomial logistic regression model:

3.1. The logit (logistic) regression model:

In fact, the multinomial logistic regression (MLR) model is a fairly straight forward generalization of the binary model, and both models depend mainly on (logistic analysis) or logistic regression. Logistic regression in different ways is the natural complement of ordinary linear regression whenever the response is in a categorical variable. When some discrete variables occur between explanatory variables, they are dealt with by the introduction of one or several (0, 1) dummy variable, but when the response variable belong to this kind of data, the multiple regression model

not suitable logistic regression provides a ready alternate.

For a response variable Y with two measurement levels (dichotomous) and explanatory variable X, let:

$$\pi(x) = p(Y = 1|X = x) = 1 - p(Y = 0|X = x) \quad \ldots(1)$$

the logistic regression model has linear form for logit of this probability:

$$\text{logit}[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x, \text{ Where the odds} = \frac{\pi(x)}{1-\pi(x)}, \ldots(2)$$

The odds=$\exp(\alpha + \beta x)$, and the logarithm of the odds is called logit, so

$$\text{logit}[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \log[\exp(\alpha + \beta x)] = \exp(\alpha + \beta x) \quad \ldots(3)$$

The logit has linear approximation relationship and logit = logarithm of the odds. The parameter ß is determined by the rate of increase or decrease of the S shaped curve of p (x).

The sign of ß indicates whether curve ascends (ß> 0) or descends (ß < 0), and the rate of [Chatterjee and Hadi,2006: 318-319]

3.2. Multiple logistic regressions:

The logistic regression can be extending to models with multiple explanatory variables. Let k denotes number of predictors for a binary response Y by $x_1, x_2, \ldots, x_k$, the model for log odds is:

$$logit[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad \ldots(4)$$

and the alternative formula, directly specifying $\pi(x)$, is

$$\pi(x) = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)} \ldots(5)$$

The parameter β_i refers to the x_i effect of on the log odds that $Y = 1$, controlling other x_j , for instance, $\exp(\beta_i)$ is the multiplicative effect on the odds of a one unit increase in x_i, at fixed levels of other x_j .

If we have n independent observations with p-explanatory variables, and the qualitative responses variable have k categories, to build the logit in the multinomial cases, one of the categories must be considered the basic level and all logits are built relative for it and all categories can be taken as the basic level, so we will take category k as the base level. Since there are not ordering, it is apparent that any category may be labeled k.

Let p_j denote the multinomial probability of an observation falling in the jth category, to discover between the relationship this probability and the p explanatory variables, X1, X2, ,....., Xp, the multiple logistic regression model then is

$$log\left[\frac{\pi_j(x_i)}{\pi_k(x_i)}\right] = \alpha_{0i} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi}\,, \qquad \ldots(6)$$

Where j= 1, 2, …, (k-1), i = 1, 2, …,n. Since all the π 's add to unity, this reduces to:

$$\log\left(\pi_j(x_i)\right) = \frac{\exp\left(\alpha_{0i}+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\cdots+\beta_{pj}x_{pi}\right)}{1+\sum_{j=1}^{k-1}\exp\left(\alpha_{0i}+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\cdots+\beta_{pj}x_{pi}\right)} \qquad \ldots(7)$$

For j = 1, 2, …, (k-1), the model parameters are estimated by the method of ML. Practically, we use statistical software to do this fitting (Chatterjee and Hadi, 2006: 329-330).

3. 3. Baseline-Category Logit Model:

In MLR model, the estimate for the parameter can be identified compared to a(reference) baseline category. We defined bold letter as matrix or vector, let $\pi_j(x) = p(Y = j|x)$ at a fixed setting **x** for explanatory variables, with$\sum_j \pi_j(x) = 1$, for observations at that setting, we search the counts at the J categories of Y as multinomial with probabilities, $\{\pi_1(x), \ldots., \pi_J(x)\}$, logit models pair each response category with a baseline category, often the most popular model is:

$$log\frac{\pi_j(x)}{\pi_J(x)} = \alpha_j + \acute{\beta_j}x, \ldots(8)$$

where j = 1,…, (*J -1*), simultaneously simultaneously characterize impact of x on these (J-1)logits, the different effective according to responses paired with baseline, these (J-1) equations determine parameters to sign in with another pairs of response categories.

Since

$$log\frac{\pi_a(x)}{\pi_b(x)} = log\frac{\pi_a(x)}{\pi_J(x)} - log\frac{\pi_b(x)}{\pi_J(x)} \qquad \ldots(9)$$

with categorical predictors, Pearson chi-square statistic X^2 and the likelihood ratio chi-square statistic G^2 goodness-of-fit statistics provide a model check when data are not sparse.

When an explanatory variable is continuous or the data are sparse, such statistics are still valid for comparing nested models differing by relatively few terms (Agresti, 2002: 267-268).

4. The information criterion:

4. 1. Akaike information criterion (AIC):

Akaike's information criterion (AIC) compares the quality of a set of statistical models to each other. For example, you might be interested in

what variables contribute to the status of low socioeconomic and how the variables contribute to that status. if you create several regression models for different reasons like education, family members, or disability status; The AIC will take each model , rank them from best to worst. The "best" rank will be the one that neither under-fits nor over-fits. Although the AIC will chooses the best model from a set, it won't say anything about absolute quality. In other meaning, if all of your models are poor, it will choose the best of a bad bunch. Thus if you selected the best model it had considered a running a hypothesis test to detect the relationship between the variables in next model to get perfective result. Akanke's Information Criterion is usually calculated with software.

Thebasic formula is defined as:

$$AIC = -2(\text{log-likelihood}) + 2K \text{ or } AIC = 2K - 2\ln(L) \quad \ldots(10)$$

Where:

- K is the number of model parameters (the number of variables in the model plus the intercept).
- Log-likelihood is a measure of model fit .This is usually obtained from statistical output.

For small sample sizes ($n/K < \approx 40$), use the second-order AIC:

$$AIC = -2(\text{log-likelihood}) + 2K + (2K(K+1)/(n-K-1)) \ldots..(11)$$

Where:

- n = sample size,
- K= number of model parameters,
- Log-likelihood is a measure of model fit.(Kenneth and David, 2002: 60-62)

4.2: Bayesian information criterion (BIC)

In statistics, the Bayesian information criterion (BIC) or Schwarz information criterion (also SIC, SBC, SBIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than

in AIC.The BIC is formally defined as

BIC=K ln(n)-2ln(L)…(12)

Where:

L= the maximized value of the likelihood function of the model M, i.e. $= p(x, \widehat{\theta}, M)$, where $\widehat{\theta}$, are the parameter values that maximize the likelihood function,

n= the number of observations, or equivalently, the sample size,

K= number of model parameters. (Claeskins and Hkort, 2008: 70-97), (Ward andAhlquist, 2018: 38).

4. 3. Maximum likelihood estimation (MLE):

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate. The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference (Chambers, Steel, Wang & Welsh, 2012: 17), (Hendry & Nielsen, 2007: 36-37), (Ward & Ahlquist, 2018: 9).

If the likelihood function is differentiable, the derivative test for determining maxima can be applied. In some cases, the first-order conditions of the likelihood function can be solved explicitly; for instance, the ordinary least squares estimator maximizes the likelihood of the linear regression model. (Press, Flannery, Teukolsky and Vetterling, 1992: 651-653) Under most circumstances, however, numerical methods will be necessary to find the maximum of the likelihood function.

4. 3. 1. The Likelihood Function:

Maximum likelihood estimation endeavors to find the most "likely" values of distribution parameters for a set of data by maximizing the value of what is called the "likelihood function" This likelihood function is largely based on the probability density function (pdf) for a given distribution. As an example, consider a generic pdf:

$f(x, \theta_1, \theta_2, \ldots, \theta_k)$ …(13)

where x represents the data (times to failure), and θ_1, θ_2,..., θ_k are the parameters to be estimated. For a two-parameter Weibull distribution, for example, these would be beta (β) and eta (η). For complete data, the likelihood function is a product of the pdf functions, with one element for each data point in the data set:

$$L = \prod_{i=1}^{n} f(xi, \theta_1, \theta_2, \dots, \theta_k) \quad \dots(14)$$

where R is the number of failure data points in the complete data set, and x_i is the i^{th} failure time. It is often mathematically easier to manipulate this function by first taking the logarithm of it. This log-likelihood function then has the form:

$$\Lambda = \ln L = \sum_{i=1}^{n} \ln f(xi, \theta_1, \theta_2, \dots, \theta_k) \quad \dots(15)$$

It then remains to find the values for the parameters that result in the highest value for this function. This is most commonly done by taking the partial derivative of the log-linear equation for each parameter and setting it equal to zero:

$$\frac{\partial \Lambda}{\partial \theta_j} = 0, j = 1,2,3, \dots, k \quad \dots(16)$$

This results in a number of equations with an equal number of unknowns, which can be solved simultaneously. This can be a relatively simple matter if there are closed-form solutions for the partial derivatives. In situations where this is not the case, numerical techniques need to be employed. (Ward and Ahlquist, 2018: 21-22)

4. 4. The Chi-square test:

The logic of hypothesis testing was first invented by Karl Pearson (1857-1936), Pearson's Chi-square distribution and the Chi-square test also known as test for goodness-of-fit and test of independence are his most important contribution to the modern theory of statistics. (Magnello, 2005-2006: 1) The importance of Pearson's Chi-square distribution was that, the statisticians could use the statistical methods that did not depend on the normal distribution to interpret the findings. He invented the Chi-square distribution to mainly cater the needs of biologists, economists, and psychologists. His paper in 1900 published in Philosophical magazine elaborates the invention of Chi-square distribution and goodness of fit test. (Rana and Singhal, 2015: 69) Chi-square test is a nonparametric test used

for two specific purpose: (a) To test the hypothesis of no association between two or more groups, population or criteria (i.e. to check independence between two variables); (b) and to test how likely the observed distribution of data fits with the distribution that is expected (i.e., to test the goodness-of-fit). It is used to analyze categorical data (e.g. male or female patients, smokers and non-smokers, etc.), it is not meant to analyze parametric or continuous data (e.g., height measured in centimeters or weight measured in kg, etc.).

The formula for calculating a Chi-square statistic is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad \ldots(17)$$

5. The Data Description

The data that we used in this research is a heart disease data that contains nine variables such as (Chest Pain, Age, Sex, Cholesterol, Fasting Blood Sugar, Thalac "Maximum Heart Rate"), Exercise, Oldpeak (ST Segment Depression induced by Exercise relative to rest and Blood Pressure) where the Chest Pain is the Response variable and the eight other variables are explanatory variables. We took this data from this site https://www.kaggle.com/ronitf/heart-disease-uci By making some changes in some variables such as Chest Pain and Blood Pressure we changed them from a scale data type to classified nominal data type and the data analyzed in SPSS 26 program. We used Multinomial Logistic Regression for data analysis by taking the chest pain as a Response variable and we have four level in the chest pain variables we took all of the reference categories so we can see the differences between all of the categories.

Table (1): Represents the Summary of the Data

Chest Pain Classes	N	Marginal Percentage
0-0.89	143	47.2%
0.90-1.79	50	16.5%
1.80-2.69	87	28.7%
2.70-3.59	23	7.6%
Total	303	100.0%

Table (2): Represents the Information about the model Fitting

The Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
The Intercept	736.645	747.787	730.645			
Final	652.757	753.027	598.757	131.889	24	0.000

Table (3): Represents the Model Fitting for the Data

Pearson			Deviance		
Chi-Square	df	Sig.	Chi-Square	df	Sig.
830.731	879	0.876	598.757	879	1.000

Table (4): Represents the Pseudo R-Square

Pseudo R-Square		
Cox and Snell	Nagelkerke	McFadden
0.353	0.388	0.181

Table (5): Represents the Likelihood Ratio Tests for all of the Predictors in the Model

	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	654.114	743.244	606.114	7.357	3	0.061
Age	647.476	736.606	599.476	0.720	3	0.869
Sex	651.766	740.896	603.766	5.010	3	0.171
Cholesterol	648.209	737.339	600.209	1.453	3	0.693
Fasting Blood Sugar	650.542	739.672	602.542	3.785	3	0.286
Thalac (Maximum Heart Rate)	658.193	747.323	610.193	11.436	3	0.010
Exercise	682.526	771.656	634.526	35.769	3	0.000
Oldpeak	665.496	754.625	617.496	18.739	3	0.000
Blood Pressure	651.493	740.622	603.493	4.736	3	0.192

Table (6): Represents the Parameter Estimation of First Reference Category (0-0.89)

ChestPain Classes		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval forExp(B)	
								Lower Bound	Upper Bound
0.90-1.79	Intercept	-3.703	2.578	2.063	1	0.151			
	Age	-0.003	0.025	0.012	1	0.912	0.997	0.950	1.047
	Sex	-0.170	0.410	0.173	1	0.678	0.843	0.377	1.884
	Cholesterol	0.000	0.004	0.004	1	0.952	1.000	0.992	1.008
	FastingBloodSugar	0.050	0.605	0.007	1	0.934	1.051	0.321	3.444
	Thalac(Maximum Heart Rate)	0.029	0.012	5.936	1	0.015	1.029	1.006	1.053
	Exercise	-1.963	0.581	11.424	1	0.001	0.140	0.045	0.438
	Oldpeak	-0.997	0.299	11.117	1	0.001	0.369	0.205	0.663
	Blood Pressure	-0.081	0.238	0.114	1	0.736	0.923	0.578	1.472
1.80-2.69	Intercept	-1.496	1.954	0.586	1	0.444			
	Age	-0.006	0.020	0.093	1	0.761	0.994	0.956	1.034
	Sex	-0.434	0.333	1.700	1	0.192	0.648	0.338	1.244
	Cholesterol	-0.003	0.003	0.701	1	0.403	0.997	0.991	1.003
	Fasting Blood Sugar	0.760	0.427	3.169	1	0.075	2.138	0.926	4.937
	Thalac (Maximum Heart Rate)	0.019	0.008	5.135	1	0.023	1.019	1.003	1.036
	Exercise	-1.828	0.388	22.249	1	0.000	0.161	0.075	0.344
	Oldpeak	-0.215	0.152	2.007	1	0.157	0.806	0.598	1.086
	Blood Pressure	0.060	0.189	0.100	1	0.751	1.062	0.734	1.536
2.70-3.59	Intercept	-8.384	3.538	5.616	1	0.018			
	Age	0.022	0.033	0.444	1	0.505	1.022	0.959	1.089
	Sex	0.815	0.625	1.702	1	0.192	2.260	0.664	7.695
	Cholesterol	-0.005	0.005	0.885	1	0.347	0.995	0.985	1.005
	Fasting Blood Sugar	0.468	0.624	0.562	1	0.454	1.596	0.470	5.424
	Thalac (Maximum Heart Rate)	0.032	0.014	5.051	1	0.025	1.032	1.004	1.062
	Exercise	-1.631	0.621	6.906	1	0.009	0.196	0.058	0.661
	Oldpeak	0.158	0.193	0.672	1	0.412	1.171	0.802	1.710
	Blood Pressure	0.576	0.288	3.993	1	0.046	1.780	1.011	3.133

Table (7): Represents the Parameter Estimation of Second Reference Category (0.90-1.79)

Chest Pain Classes		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
0-0.89	Intercept	3.703	2.578	2.063	1	0.151			
	Age	0.003	0.025	.012	1	0.912	1.003	0.955	1.053
	Sex	0.170	0.410	.173	1	0.678	1.186	0.531	2.650
	Cholesterol	0.000	0.004	.004	1	0.952	1.000	0.993	1.008
	Fasting Blood Sugar	-0.050	0.605	.007	1	0.934	0.951	0.290	3.115
	Thalac (Maximum Heart Rate)	-0.029	0.012	5.936	1	0.015	0.972	0.950	.994
	Exercise	1.963	0.581	11.424	1	0.001	7.120	2.281	22.221
	Oldpeak	0.997	0.299	11.117	1	0.001	2.711	1.508	4.873
	Blood Pressure	0.081	0.238	.114	1	0.736	1.084	0.679	1.729
1.80-2.69	Intercept	2.206	2.577	.733	1	0.392			
	Age	-0.003	0.024	.019	1	0.891	0.997	0.950	1.046
	Sex	-0.263	0.391	.453	1	0.501	0.769	0.357	1.654
	Cholesterol	-0.002	0.004	.362	1	0.547	0.998	0.990	1.005
	Fasting Blood Sugar	0.710	0.567	1.565	1	0.211	2.034	0.669	6.183
	Thalac (Maximum Heart Rate)	-0.010	0.012	.697	1	0.404	0.990	0.968	1.013
	Exercise	0.135	0.640	.044	1	0.833	1.144	0.326	4.013
	Oldpeak	0.782	0.301	6.737	1	0.009	2.186	1.211	3.945
	Blood Pressure	0.140	0.235	.356	1	0.551	1.151	0.726	1.824
2.70-3.59	Intercept	-4.682	3.968	1.392	1	0.238			
	Age	0.024	0.036	.457	1	0.499	1.025	0.955	1.100
	Sex	0.986	0.675	2.132	1	0.144	2.680	0.714	10.067
	Cholesterol	-0.005	0.006	.641	1	0.423	0.995	0.984	1.007
	Fasting Blood Sugar	0.418	0.747	.313	1	0.576	1.518	0.351	6.564
	Thalac (Maximum Heart Rate)	0.003	0.017	.039	1	0.844	1.003	0.971	1.037
	Exercise	0.332	0.811	.167	1	0.683	1.393	0.284	6.827
	Oldpeak	1.156	0.333	12.024	1	0.001	3.176	1.653	6.104
	Blood Pressure	0.657	0.327	4.040	1	0.044	1.929	1.016	3.661

Table (8): Represents the Parameter Estimation of Third Reference Category (1.80-2.69)

Chest Pain Classes		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
0-0.89	Intercept	1.496	1.954	0.586	1	0.444			
	Age	0.006	0.020	0.093	1	0.761	1.006	0.967	1.047
	Sex	0.434	0.333	1.700	1	0.192	1.543	0.804	2.962
	Cholesterol	0.003	0.003	0.701	1	0.403	1.003	0.997	1.009
	Fasting Blood Sugar	-0.760	0.427	3.169	1	0.075	0.468	0.203	1.080
	Thalac (Maximum Heart Rate)	-0.019	0.008	5.135	1	0.023	0.981	0.966	0.997
	Exercise	1.828	0.388	22.249	1	0.000	6.221	2.911	13.296
	Oldpeak	0.215	0.152	2.007	1	0.157	1.240	0.921	1.671
	Blood Pressure	-0.060	0.189	0.100	1	0.751	0.942	0.651	1.363
0.90-1.79	Intercept	-2.206	2.577	0.733	1	0.392			
	Age	0.003	0.024	0.019	1	0.891	1.003	0.956	1.053
	Sex	0.263	0.391	0.453	1	0.501	1.301	0.604	2.801
	Cholesterol	0.002	0.004	0.362	1	0.547	1.002	0.995	1.010
	Fasting Blood Sugar	-0.710	0.567	1.565	1	0.211	0.492	0.162	1.495
	Thalac (Maximum Heart Rate)	0.010	0.012	0.697	1	0.404	1.010	0.987	1.033
	Exercise	-0.135	0.640	0.044	1	0.833	0.874	0.249	3.064
	Oldpeak	-0.782	0.301	6.737	1	0.009	0.458	0.253	0.826
	Blood Pressure	-0.140	0.235	0.356	1	0.551	0.869	0.548	1.378
2.70-3.59	Intercept	-6.888	3.604	3.653	1	0.056			
	Age	0.028	0.033	0.705	1	0.401	1.028	0.964	1.097
	Sex	1.249	0.629	3.940	1	0.047	3.488	1.016	11.973
	Cholesterol	-0.002	0.005	0.198	1	0.656	0.998	0.987	1.008
	Fasting Blood Sugar	-0.292	0.615	0.226	1	0.634	0.747	0.224	2.490
	Thalac (Maximum Heart Rate)	0.013	0.015	0.807	1	0.369	1.013	0.985	1.043
	Exercise	0.197	0.679	0.084	1	0.772	1.217	0.322	4.608
	Oldpeak	0.374	0.208	3.240	1	0.072	1.453	0.967	2.183
	Blood Pressure	0.517	0.296	3.053	1	0.081	1.677	0.939	2.993

Table (9): Represents the Parameter Estimation of Last Reference Category (2.70-3.59)

Chest Pain Classes		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
0-0.89	Intercept	8.384	3.538	5.616	1	0.018			
	Age	-0.022	0.033	0.444	1	0.505	0.979	0.918	1.043
	Sex	-0.815	0.625	1.702	1	0.192	0.442	0.130	1.506
	Cholesterol	0.005	0.005	0.885	1	0.347	1.005	0.995	1.015
	Fasting Blood Sugar	-0.468	0.624	0.562	1	0.454	0.626	0.184	2.128
	Thalac (Maximum Heart Rate)	-0.032	0.014	5.051	1	0.025	0.969	0.942	0.996
	Exercise	1.631	0.621	6.906	1	0.009	5.110	1.514	17.251
	Oldpeak	-0.158	0.193	0.672	1	0.412	0.854	0.585	1.246
	Blood Pressure	-0.576	0.288	3.993	1	0.046	0.562	0.319	0.989
0.90-1.79	Intercept	4.682	3.968	1.392	1	0.238			
	Age	-0.024	0.036	0.457	1	0.499	0.976	0.909	1.048
	Sex	-0.986	0.675	2.132	1	0.144	0.373	0.099	1.401
	Cholesterol	0.005	0.006	0.641	1	0.423	1.005	0.993	1.017
	Fasting Blood Sugar	-0.418	0.747	0.313	1	0.576	0.659	0.152	2.847
	Thalac (Maximum Heart Rate)	-0.003	0.017	0.039	1	0.844	0.997	0.965	1.030
	Exercise	-0.332	0.811	0.167	1	0.683	0.718	0.146	3.518
	Oldpeak	-1.156	0.333	12.024	1	0.001	0.315	0.164	0.605
	Blood Pressure	-0.657	0.327	4.040	1	0.044	0.518	0.273	0.984
1.80-2.69	Intercept	6.888	3.604	3.653	1	0.056			
	Age	-0.028	0.033	0.705	1	0.401	0.973	0.911	1.038
	Sex	-1.249	0.629	3.940	1	0.047	0.287	0.084	0.984
	Cholesterol	0.002	0.005	0.198	1	0.656	1.002	0.992	1.013
	Fasting Blood Sugar	0.292	0.615	0.226	1	0.634	1.339	0.402	4.468
	Thalac (Maximum Heart Rate)	-0.013	0.015	0.807	1	0.369	0.987	0.959	1.016
	Exercise	-0.197	0.679	0.084	1	0.772	0.821	0.217	3.109
	Oldpeak	-0.374	0.208	3.240	1	0.072	0.688	0.458	1.034
	Blood Pressure	-0.517	0.296	3.053	1	0.081	0.596	0.334	1.065

Table (10): Represents the Classification Statistics of the Predictors by the Model

Observed	Predicted				
	0-0.89	0.90-1.79	1.80-2.69	2.70-3.59	Percent Correct
0-0.89	106	13	22	2	74.1%
0.90-1.79	10	15	24	1	30%
1.80-2.69	33	13	41	0	47.1%
2.70-3.59	10	2	11	0	0%
Overall Percentage	52.5%	14.2%	32.3%	1%	53.5%

6. The Results of the Data:

- The Model Fitting Information table contains a Likelihood Ratio Chi-Square test, comparing the Full Model (i.e., containing all the predictors) with the Null Model (i.e., containing intercept only or no predictors). Statistical significance indicates that the Full Model represents a significant improvement fit over the Null Model. We can see that this model is a significant improvement in fit over a null model since the $x^2_{(24)} = 131.889, p < 0.05$.
- The Goodness of Fit table contains the Pearson and Deviance Chi-Square tests, which are useful for determining whether a model exhibits good fit to the data. Non-significant test results are indicators that the model fits the data well. Both of the Two different tests indicates that the model fits the data well since Pearson's Chi-Square test is $x^2_{(879)} = 830.731, p = 0.876$ and Deviance's Chi-Square test is $x^2_{(879)} = 598.757, p = 1.000$
- These are Pseudo R-Square values that are treated as rough analogues to the R-Square value in OLS Regression. In general, there is no strong guidance in the literature on how these should be used or interpreted.
- These results contain Likelihood Ratio Tests of the overall contribution of each independent variable to the model. Using the conventional $\alpha = 0.05$ threshold, we see that each of the (Thalac (Maximum Heart Rate), Exercise and Old peak) predictors in the model are Significance since their values are greater than the $\alpha = 0.05$.
- The results in table (6) provide information comparing each chest pain level against the reference category. The first set of coefficients represents

comparison between the first category (0-0.89) with the other three categories where each of (Thalac "Maximum Heart Rate", Exercise and Oldpeak) were significant in the model, since all of the values respectively are equal to (0.015, 0.001 and 0.001) and in the second set of coefficients both of (Thalac (Maximum Heart Rate), Exercise) were significant since their values are equal to (0.023 and 0.000) in the third set of coefficients (Thalac "Maximum Heart Rate", Exerciseand Blood Pressure) were significant since their values are equal to (0.025, 0.009 and 0.046) by comparing all of them with $\alpha = 0.05$. We also have a nearly significant value which is Fasting Blood Sugar that is equal to 0.075.

- In table (7) which provide information comparing each chest pain level against the reference category. The first set of coefficients represents comparison between the Second category (0.90-1.79) with the other three categories where each of (Thalac "Maximum Heart Rate", Exercise and Oldpeak) were significant in the model, since all of the values respectively are equal to (0.015, 0.001 and 0.001) and in the second set of coefficients only (Oldpeak) was significant since the values is equal to (0.009) in the third set of coefficients (Oldpeak and Blood Pressure) were significant since their values are equal to (0.001 and 0.044) by comparing all of them with $\alpha = 0.05$.
- The results in table (8) provide information comparing each chest pain level against the reference category. The first set of coefficients represents comparison between the Third category (1.80-2.69) with the other three categories where each of (Thalac "Maximum Heart Rate" and Exercise) were significant in the model, since all of the values respectively are equal to (0.023 and 0.000) and in the second set of coefficients only (Oldpeak) was significant since the values is equal to (0.009)by comparing all of them with $\alpha = 0.05$. There are also nearly significant values such as Fasting Blood Sugar that is equal to 0.075, Oldpeak (0.072) and Blood Pressure (0.081).
- The results in table (9) provide information comparing each chest pain level against the reference category. The first set of coefficients represents comparison between the Lastcategory (2.70-3.59) with the other three categories where each of (Thalac "Maximum Heart Rate", Exercise and Blood Pressure) were significant in the model, since all of the values

respectively are equal to (0.025, 0.009 and 0.046) and in the second set of coefficients (Oldpeak and Blood Pressure) were significant since their values are equal to (0.001 and 0.044), in the third set of coefficients only sex was significant since its value is equal to (0.047) by comparing all of them with $\alpha = 0.05$. There are also nearly significant values such as (Oldpeak and Blood Pressure) since their values are equal to (0.072 and 0.081).

- Table (10) represents the classification statistics used to determine which group membership were best predicted by the model. Where the first reference category that correctly predicted by the model is 74.1% of the time [as 106 of 143 cases in the first level of chest pain were predicted to do so by the model;106/(106 + 13 + 22 + 2) = 0.741. and the second reference category that correctly predicted by the model is 30.0%, the third reference category that correctly predicted by the model is 47.1%. But the last reference category that correctly predicted by the model is 0%.

7. Conclusion: In this study we conclude that:

- By comparing the null model and the full model we conclude that the full model was significant also both of the Pearson and Deviance Chi-Square tests indicates that the model fits the data well.
- In the Likelihood Ratio Tests each of the (Thalac "Maximum Heart Rate", Exercise and Oldpeak) predictors in the model are Significance.
- There are many significant variables in all four reference categories in the model, specifically (Thalac "Maximum Heart Rate" and Exercise) were significant in each different categories in the Multinomial Logistic Regression Model.

8. Recommendations:

- We recommend our hospitals in the Kurdistan Region, especially Sulaymaniyah city, to create a database on individuals in the community about different types of diseases so that we can use our country's internal data.
- patients with high blood pressure, tachycardia (Maximum Heart Rate) and IHD or MI (Ischemic Heart Disease or Myocardial Infarction) ST segment depression (Oldpeak) must use antihypertensivedrugs, antiarrhythmic drugs, MI or IHD medication to control and prevent this situations.
- Choose an aerobic activity such as walking, swimming, light jogging, or biking. Do this at least 3 to 4 times a week. Always do 5 minutes of

stretching or moving around to warm up your muscles and heart before exercising. A well-balanced diet with whole grains, fruits, vegetables, and lean protein is always smart.

- Lifestyle modifications are recommended for lipid management in all patients, including daily physical activity and weight management. The recommended dietary therapy for all patients should include reducing intake of saturated fats to less than 7% of total calories, reducing intake of trans-fatty acids to less than 1% of total calories, and reducing daily cholesterol intake to less than 200 mg. In addition, moderate- to high-dose statin therapy should be prescribed in the absence of contraindications or documented adverse effects.

Reference:

1. A. Erkan, (2016), Using Multinomial Logistic Regression to Examine the Relationship between Children's Work Status and Demographic Characteristics, Siyaset Ekonomive Y önetim Araştırmaları Dergisi 4.
2. A. Erkan and N. Aydin, (2016), Examination by Multinomial Logistic Regression Model of the Factors Affecting the Types of Domestic Violence against Women: A Case of Turkey, International Journal of Scientific & Technology Research Vol. 5, Issue 11. pp. 67-74.
3. A. Murata, Y. Fujii and K. Naitoh, (2015), Multinomial logistic regression model for predicting driver's drowsiness using behavioral measures, Journal of Traffic and Transportation Engineering. Vol. 3, No. 2. pp. 2426-2433.
4. Abdalla M. El-Habil, (2012), an Application on Multinomial Logistic Regression Model, Pakistan Journal of Statistics and Operation Research 8(2).
5. Agresti, A., (2002), Categorical Data Analysis, John Wiley & Sons, Inc.
6. B. Madhu, N. C. Ashok and S. Balasubramanian, (2014), A Multinomial Logistic Regression Analysis to Study the Influence of Residence and Socio-Economic Status on Breast Cancer Incidences in Southern Karnataka, International Journal of Mathematics and Statistics Invention (IJMSI). Vol. 2, Issue 5. pp. 1-8.
7. C. Coughenour, A. Paz, H. de la Fuente-Mella and A. Singh, (2015), Multinomial logistic regression to estimate and predict perceptions of bicycle and transportation infrastructure in a sprawling metropolitan area, Journal of Public Health. Vol. 38, No. 4, pp. e401-e408.
8. Chambers, Raymond L.; Steel, David G.; Wang, Suojin; Welsh, Alan, (2012), Maximum Likelihood Estimation for Sample Surveys, Boca Raton: CRC Press.
9. Chatterjee, S., and Hadi, A., (2006), Regression Analysis by Example, John Wiley & Sons.

10. Claeskins, G. &Hkort, N., (2008), Model Selection and Model Averaging (Cambridge Series in Statistical and Probabilistic Mathematics), 1st Edition. Cambridge University Press.
11. G. Asampana, K. K. Nantomah and E. A. Tungosiamu, (2017), Multinomial Logistic Regression Analysis of the Determinants of Students' Academic Performance in Mathematics at Basic Education Certificate Examination, Higher Education Research.Vol. 2, No. 1, 2017, pp. 22-26.
12. Garson, D., (2009), Logistic Regression with SPSS, North Carolina State University, Public administration Program.
13. H. Woo-Yong and R. B. Ditton, (2007), Using multinomial logistic regression analysis to understand angler's willingness to substitute other fishing locations, Proceedings of the 2006 Northeastern Recreation Research Symposium. pp. 248-255.
14. Hendry, David F.; Nielsen, Bent, (2007), Econometric Modeling: A Likelihood Approach, Princeton: Princeton University Press.
15. Kenneth P. Burnham and David R. Anderson, (2002), Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Springer Science & Business Media.
16. Magnello ME Karl Pearson, (2005-2006), the origin of modern statistics: An elastician becomes a statistician, Rutherford J, Vol. 1, 2005-2006.
17. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T., (1992), Least Squares as a Maximum Likelihood Estimator. Numerical Recipesin FORTRAN: The Art of Scientific Computing (2nd ed.). Cambridge: Cambridge University Press. pp. 651–655.
18. R. Rana, R. Singhal, (2015), Chi‑square Test and its Application in Hypothesis Testing, Statistical Section, Central Council for Research in Ayurvedic Sciences, Ministry of AYUSH, GOI, New Delhi, India, Vol. 1, Issue. 1, pp. 69-71.
19. Ward, Michael Don; Ahlquist, John S., (2018), Maximum Likelihood for Social Science: Strategies for Analysis, New York: Cambridge University Press.