Estimating Hazard Function and Survival Analysis of Tuberculosis Patients in Erbil city

Prof. Dr. Kurdistan Ibrahim Mawlood
College of Administration and Economics
Salahaddin University/Erbil

Researcher: Hogr Muhammed Qader
College of Administration and Economics
Salahaddin University/Erbil

Abstract:

The study aimed to estimate the effects of prognostic factors on tuberculosis (TB) survival. Two models have been studied (Logistic model and Cox regression models) in survival analysis.

Kaplan Meier has been applied to estimate the hazard function. The Kaplan Meier curve has been used to show the risks of dyeing of the factors in this study of tuberculosis data set. The data was obtained from Kurdistan Regional Government, Iraq /Ministry of health/General Directorate of Health, Hawler/Chest and Respiratory disease Center, in period 11th January 2015 through 23th November 2019 of all tuberculosis patients followed up by the hospital until 14th April 2020. Kaplan Meier estimator results indicates that in the factor X-ray result TB has the highest value of estimated mean time until death, the Kaplan Meier curves are clearly indicated that the risk of dying increased with the time especially after 15 months. The logistic regression model identifies that (Gender, Chest symptoms, Type of patient, Site of TB, Transpupillary thermotherapy (TTT-outcome)) are the prognostic factors that influence in tuberculosis survival. Moreover, the Cox regression model identifies that (Age group, Gender, Site of TB, TTT-outcome) are the most common factors that have an impact on tuberculosis.

Logistic regression model was selected to be the best model for our study data of tuberculosis by using the criterions; Akaike Information Criterion (AIC) and Bayesian information criterion (BIC) to comparing two models. It's worth mentioning; the results obtained by utilizing the statistical packages in Mat-lab and SPSS V.25, which was used to analyze our study data.

Keywords: Logistic regression, survival analysis, censoring, Kaplan Meier estimator, Cox regression**.**

تقدير دالة المخاطرة وتحليل البقاء لمرضى السل في مدينة أربيل

أ.د. كوردستان إبراهيم مولود
كلية الإدارة والاقتصاد
جامعة صلاح الدين/أربيل
hogrstat@gmail.com

الباحث: هؤكر محمد قادر
كلية الإدارة والاقتصاد
جامعة صلاح الدين/أربيل
kurdistan.mawlood@su.edu.krd

المستخلص:

هدفت الدراسة إلى تقدير العوامل المؤثرة على احتمالية بقاء مرضى السل استخدمت الدراسة نموذجين (نموذج الانحدار اللوجستي ونموذج الانحدار كوكس) في تحليل البقاء. وقد تم تطبيق كابلان ماير لتقدير دالة المخاطرة. تم الحصول على البيانات من المديرية العامة للصحة،

أربيل/مركز أمراض الصدر والجهاز التنفسي في الفترة من 11 كانون الثاني 2015 الى 23 تشرين الثاني 2019 لجميع مرضى المصابين بالسل الذين راجعوا المستشفى مع فترة متابعة الى 14 نيسان 2020. وأظهرت النتائج لكابلان ماير فرق كبير في البقاء لمرضى المصابين بالسل لجميع العوامل المؤثرة كما أظهرت منحنيات كابلان ماير بوضوح إلى أن خطر الموت يزداد مع مرور الوقت خاصة بعد 15 شهرًا. تم اختيار نموذج الانحدار اللوجستي ليكون أفضل نموذج لتمثيل بيانات الدراسة لمرضى السل؛ وتم استخدام معيارين للمفاضلة لاختيار أفضل نموذج للبيانات وهي معيار معلومات أكايكي ومعيار معلومات بيز.

الكلمات المفتاحية: الانحدار اللوجستي، تحليل البقاء، الرقابة، مقدر كابلان ماير، انحدار كوكس.

1. Introduction

Tuberculosis (TB) is a contagious disease caused by infection with Mycobacterium tuberculosis (MTB) bacteria. The bacteria that cause tuberculosis are spread from one person to another through tiny droplets released into the air via coughs, sneezes, speaks or sings, and people nearby breathe in these bacteria and become infected.

Scientific analysis, which depends on efficient statistical methods with quantitative and scientific measurements and parameters, must be used to study those specifications and characteristics. Using logistic regression and cox regression models, this study aimed to identify the important variables that influence tuberculosis disease.

2. Theoretical Background

2-1. Logistics Regression Model: Analysis of logistic regression is one of the important statistical methods that can be used in many areas of life, when there is a binary response variable or classified nature the relationship takes the formula of the logistic distribution function model and the explanatory variables can be quantitative, qualitative, mixed, ordinal and binary. What distinguishes the logistic regression model from the linear regression model is that the outcome in logistic regression is binary or dichotomous. (Hosmer & Lemeshow, 2000: 1)

Let the response variable Y be the binary variable, assuming that P(Y =1) is dependent on a vector of predictor variables $\bar{x}$. The goal is to model:

$$P(\bar{x}) = P(y = 1|\bar{x}) \qquad \dots (1)$$

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

Then the model may be result in estimated probability value which are over a restricted value [0,1]. Then to avoid this the logistic function can be used, it assumes that:

$$p(\bar{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)} \quad \ldots(2)$$

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

$$\log\left(\frac{p(\bar{x})}{1-p(\bar{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \qquad \ldots(3)$$

This model is called the logistic regression model.

2.1.1. Maximum Likelihood Estimation Method for Parameters: The regression coefficients are usually estimated using maximum likelihood estimation. Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximize the likelihood function, so that an iterative process must be used instead; for example, Newton's method.

Estimation of parameters in logistic regression (the coefficients, $\beta_1, \beta_2, \ldots, \beta_k$) can be estimated by maximum likelihood method:

$$\prod_{i=1}^{n} \{p(\bar{x})^{y_i}[1 - p(\bar{x}_i)]^{1-y_i}\}$$

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

(The MLE's are determined numerically, by maximizing the log likelihood.) by taking the first derivative of the log maximum likelihood equation and then equivalents by zero we get the equations are nonlinear in the parameters the solution can be estimated numerically and therefore resort to the use of Newton Rafson iterative method and after a few cycles of succession produced appropriate estimates of the parameters. (Mawlood, 2019: 708)

2.1.2. Evaluating the Performance of the models: In linear regression analysis ordinary least square used to fit a model, t. F tests, and residuals are used to the coefficient's and the model. The situation is different with in logistic regression, the approximate chi-square and z tests and likelihood ratio test are used.

To test the parameters rather to determine are they equal to zero or not:

$H_0: \beta_0 = \cdots = \beta_p = 0$

$H_1:$ at least two of them are not equal to zero

chi-square test is used which is based on difference between the estimated log likelihoods corresponding to the two models, the test statistics is given by:

$$LR(p) = -2[LnL(\alpha) - LnL(\alpha, \beta)] \quad \ldots(4)$$

Where: $LnL(\alpha)$ represents the logarithm of likelihood function of the Reduced Model, which contains only the Intercept parameter. (Archer & Lemeshow, 2006: 98)

And for testing that if the explanatory variables are included in the model or not:

$H_0:$ Explanatory variables are included in the model.

$H_1:$ Explanatory variables are not included in the model.

The Hosmer-Lemeshow goodness-of-fit measure, on the other hand, is useful for unreplicated datasets or datasets with just a few repeated observations. The observations are grouped in this test based on their approximate probabilities. The test statistic that results is approximately chi-square distributed with n - 2 degrees of freedom, where n is the number of groups (generally chosen to be between 5 and 10, depending on the sample size). (Hosmer & Lemeshow, 2000: 327)

Similar to McFadden's R squared, Cox-Snell's R squared uses the likelihood of the selected model and an intercept-only model fit to the same data (McFadden's R squared uses the log likelihood). In this case,

Cox-Snell's R Squared = 1 – [(Likelihood (Intercept-only Model)/ (Likelihood(Specified Model)]$^{2/n}$

$$R^2_{C\&S} = 1 - \left[\frac{L_O}{L_M}\right]^{2/n} \quad \ldots(5)$$

Where: n is the number of observations.

Wald χ^2 statistics is used to test the significance of individual coefficients in the model and are calculated as follows:

$$\left(\frac{coefficient}{SEcoefficient}\right)^2 = \left(\frac{\beta}{SE(\beta)}\right)^2 \quad \ldots(6)$$

To test the coefficients are they equal to zero or not:

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

Each Wald statistic is compared to a 2-degree-of-freedom distribution. Wald statistics are simple to compute, but their accuracy is debatable, particularly for small samples. The standard error is often exaggerated for data that yield large estimates of the coefficient, resulting in a lower Wald statistic, and therefore the explanatory variable may be wrongly assumed to be unimportant in the model. The use of likelihood ratio tests is commonly thought to be preferable.

2-2. Survival analysis: The study of time-to-event data is known as survival analysis. Such information describes the length of time between a time origin and a desired endpoint. Individuals could be tracked from birth until the onset of a disease, or the recovery period after a disease diagnosis could be studied. Data collected prospectively in time, such as data from a prospective cohort study or data collected for a clinical trial, is typically analyzed using survival analysis techniques. (Xin, 2011: 68)

Survival analysis can be used to analyze health-care use in the field of public health. Since the health-care system represents a society's political and economic structure and is concerned with fundamental philosophical issues such as life, death, and quality of life, such an analysis is particularly important for both planners and scholars. (Liu, 2012: 12)

Survival analysis is usually deal with the analysis of data in times of events in the history of individual life. The survival analysis and modeling the time it takes events occur; this typical event is death, which is derived from the name ' survival ' analysis. Let T be a random variable represents survival time of an event, with probability density function f (t) and cumulative function $F(t) = Pr(T \leq t)$, the survival function S(t) is defined as: (Fox, 2014: 66)

$$S(t) = Pr(T > t) = 1 - F(t) \quad \dots (7)$$

The hazard function is the function that symbolized as h(t) and gives the failure rate for the survival time, which is defined as the probability of a failure during a small period of time (conditional failure rate) assuming that the individual might have remained alive until the beginning of the period, as well as the individuals fail in the so small time per unit time given that the individual have remained alive until time (t): (Wienke, 2011: 88)

$$h_{(t)} = \lim_{\Delta t \to 0} \frac{pr(t \leq T \leq t+\Delta t | T \geq t)}{\Delta t} \quad \dots (8)$$

2-2-1. Censoring: Censoring is a type of missing data problem in which the time to event is not recorded for a variety of reasons, such as the research being terminated before all recruited participants have demonstrated the event of interest, or the subject leaving the study before witnessing an event. In survival research, censorship is popular. When knowledge about a subject's survival time is lacking, observations are censored. Interval censoring is applied to the data in the sense that certain transition times are not observed but are assumed to fall within a given time interval. The onset of dementia, for example, is latent, but when longitudinal evidence is available, the onset can be determined to be during the time period specified by 2 sequential observations. (Hout, 2017: 67)

2-2-2. Kaplan-Meier estimator to estimate hazard function: The Kaplan-Meier (KM) method is a non-parametric method used to estimate the survival probability from observed survival times (Kaplan and Meier, 1958).

The survival probability at time t_i, $S(t_i)$, is calculated as follow:

$$S(t_i) = \prod_{t_i \leq} (1 - \frac{d_i}{n_i}) \quad \dots (9)$$

Where,

t_i Is duration of study at point i

d_i Is number of deaths up to point i

n_i Is number of individuals at risk just prior to t_i

S is the likelihood of an individual surviving at the end of a time period assuming that the individual was alive at the beginning of the time interval. The hazard rate at time t conditional on surviving up to or beyond time t is described as the instantaneous hazard function h(t) [also known as the hazard rate, conditional failure rate, or force of mortality].

Since h (t) is a rate rather than a chance, its units are 1/t. The cumulative hazard function H hat (t) is the integral of the hazard rates from time 0 to t, which reflects the sum of the hazard over time-mathematically, this quantifies the number of times the failure occurrence would be expected to occur in a given time span if the event was repeatable. As a result, thinking of hazards in terms of rates rather than probabilities is more reliable. The cumulative hazard is calculated using Peterson's (1977) method as follows: (Korosteleva, 2008: 125)

$$\hat{H}_t = -\ln(\hat{S}_t) \qquad \dots (10)$$

The estimated survivor function $\hat{S}_t$ and hazard function $\hat{h}_t$ from the life-table method and the estimated survivor function $\hat{S}_t$ from the Kaplan-Meier method can be further plotted to produce graphics. In these graphics, each estimated statistic is used as the vertical axis, and the study time is used as the horizontal axis. Based on equations about the cumulative hazard function, the analyst can further produce a cumulative hazard plot. (Guo, 2010: 88)

2-3. The cox proportional hazards model: The Cox proportional hazards regression model is the most convenient way to build regression models for survival data, time to-event outcome, based upon the values of given covariates. The Cox (1972) proportional hazards (PH) model has been an extremely popular regression model in the analysis of survival data during the last decades. The corresponding survival functions are related as follows: (Balakrishman & Rao, 2004: 186)

$$S(t|x) = S_{o\,(t)}{}^{\exp\,(\sum_{i=1}^{p} B_i X_i\ )} \qquad \dots (11)$$

where $h_0(t)$ is an unspecified baseline hazard function free of the covariates x. The covariates act multiplicatively on the hazard. Clearly, the exponential and Weibull are special cases. (Ekman, 2017: 86)

One subject's hazard is a multiplicative replica of another's; comparing subject j to subject m, the model is stated as: (Mawlood, 2019: 711)

$$\frac{h(t|x_j)}{h(t|x_m)} = \frac{\exp(x_j B_x)}{\exp(x_m B_x)}$$

A parametric regression model based on the exponential distribution: (Fox, 2014: 69)

$$\log h_{i(t)} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots\cdots + \beta_k x_{ik} \qquad \dots (12)$$

Let $h(t|x_t)$ denote the hazard rate at time for an individual have covariate value x_t

$$h(t|x_t) = h_0(t) * \exp(\beta' x) \qquad \dots(13)$$

Here $x_t = (x_{1t}, x_{2t}, \dots, x_{kt}), \beta = (\beta_1, \beta_2, \dots, \beta_k)$

k is the total number of the covariates.

β_t Is the constant Proportional effect of treatment.

$h_{0\,(t)}$ is called the baseline hazard; it is the hazard for the respective individual when all independent variable values are equal to zero. (Schmidt & Witte, 1998: 210)

Fitting the Cox proportional hazards model, the estimation of the bias line (h_0 (t)) and β is needed to attempt to maximize the likelihood function for the observed data simultaneously with respect to h_0 (t). Similarly, a more popular approach is proposed by Cox in which the partial likelihood function that does not depend on h_0 (t) is obtained. (Lokeshmaran, 2013: 23) (فتيحة، ٢٠١٥: ٣٣)

2-4. Measures of the Model Selection: In this study two measures for selecting the best model have been used by comparing the accuracy and performance of methods for comparing models simply involves calculating the measures for each model; the model with the lowest value is chosen as the best model. (Lee & Wang, 2003: 230)

2-4-1. Akaike's Information Criterion: The Akaike Information Criterion (AIC) compares the quality of a set of statistical models to each other. For example, Akaike's Information Criterion is calculated as follows:

$$AIC = -2\text{loglikelihood} + 2K \quad \dots (14)$$

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. (Moore, 2016: 81)

2-4-2. The Bayesian Information Criterion: The Bayesian information criterion (BIC) is one of the most widely known and pervasively used equipment in statistical model selection. BIC is computed for each of the models corresponding to the minimum value of BIC is selected.

$$BIC = -2\text{loglikelihood} + 2 * \log N * k \quad \dots (15)$$

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

Comparing models with the Bayesian information criterion simply involves calculating the BIC for each model; the model with the lowest BIC is chosen as the best model. (Lee & Wang, 2003: 231)

3. Application part:

3-1. Introduction: In this section we are dealing with the collected data and the analysis of our data. In the medicine and health situations the most important analysis to use is the analyzing the time to event, which is the time from entry of any decease into a study until a decease has a particular outcome. Survival analysis techniques have been presented to analyze

tuberculosis data. The Kaplan-Meier estimator have been used to estimate the hazard function. And for two models (Logistic model and Cox PH model) were used for the survival analysis data. Moreover, all the corresponding results have been given and a comparison between the main methods: Cox model and Logistic model has been done.

To evaluate the best survival model to our tuberculosis data two statistical measures (AIC and BIC) were used.

3-2. Data collection: The data for this study of tuberculosis have been collected from Kurdistan Regional Government, Iraq/Ministry of health/General Directorate of Health, Hawler/Chest and Respiratory disease Center, Hawler. The data consisted of 788 cases which are collected during 5 years period; beginning from 11th January 2015 through 23th November 2019 of all tuberculosis patients followed up by the hospital until 14th April 2020. Out of those patients there are 159 patients died during the study and 629 patients are survived or still alive. The survival time are measured in months.

Table (1): Classification table

Variable names	Classifications	N	No. of Alive	No. of Death
Age grouped	1=0-4	20	19	1
	2=5-14	28	20	8
	3=15-24	147	127	20
	4=25-34	158	133	25
	5=35-44	89	71	18
	6=45-54	87	70	17
	7=55-64	96	75	21
	8=65+	163	114	49
Gender	1=male	363	289	74
	2=female	425	340	85
Chest Symptoms	0=no	598	473	125
	1=yes	190	158	34
Type of patient	1=N(EP)	383	307	76
	2=N(S+)	172	144	28
	3=N(S-)(ND)	132	105	27
	4=R(EP)	37	27	10
	5=R(S+)	37	26	11
	6=R(S-)(ND)	21	15	6
	7=D(S+)	6	5	1

Variable names	Classifications	N	No. of Alive	No. of Death
Site of TB	1=PTB	370	296	74
	2=EP	418	333	85
X-ray result	0=NA	683	546	137
	1=Normal	8	4	4
	2=TB	97	79	18
TTT-outcome	1=complete	396	396	0
	2=cure	99	99	0
	3=to	124	124	0
	4=Death	159	0	159
	5=Fail	10	10	0
Condition	0=Death	159		
	1=Alive	629		
Treatment Duration	the number of months of treatment duration	788		

3-3. Application of Kaplan-Meier: The Kaplan-Meier method is a nonparametric technique for estimating time-to events (the survivorship function). Ordinarily it is used to analyze death as an outcome. It may be used effectively to analyze time to an endpoint. Also, used to estimate the hazard function and for comparing two different study populations.

Table (2): Means for survival time

Factor	Means for survival time				
		Estimate	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Age group	0-4	14.556	.432	13.709	15.402
	5-14	12.438	.686	11.094	13.781
	15-24	16.116	.757	14.631	17.600
	25-34	17.386	.779	15.858	18.913
	35-44	15.298	.810	13.710	16.886
	45-54	14.384	.697	13.019	15.750
	55-64	14.531	.653	13.251	15.811
	65+	16.349	.909	14.568	18.131
chest symptoms	No	16.160	.440	15.298	17.022
	Yes	17.703	1.247	15.259	20.146

Factor		Means for survival time			
		Estimate	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Type of patient	N(EP)	16.476	.528	15.440	17.512
	N(S+)	13.642	.407	12.844	14.439
	N(S-)(ND)	18.267	.971	16.364	20.170
	R(EP)	11.543	.810	9.955	13.132
	R(S+)	12.332	.761	10.840	13.823
	R(S-)(ND)	12.152	1.218	9.765	14.539
	D(S+)	8.250	.650	6.977	9.523
Gender	Male	16.732	.493	15.765	17.699
	Female	16.997	.690	15.645	18.348
Site of TB	PTB	17.428	.752	15.955	18.902
	EP	16.283	.514	15.277	17.290
X-ray result	NA	16.297	.410	15.493	17.100
	Normal	12.625	2.345	8.029	17.221
	TB	19.130	.903	17.360	20.901
Overall		17.335	.509	16.338	18.332

Table (2) explains the results of KM for all factors applied to data set of 788 cases. The results of KM for the Age group factor the Ages between 25-34 have the highest estimated mean time until death and the ages between 5-14 have the lowest estimated mean time until death. Moreover, for the chest symptoms the estimated mean time until death for those patients have the chest symptoms is greater than those don't have chest symptoms which are (17.703) for they have chest symptoms with confidence interval (15.259-20.146) and (16.160) for those don't have chest symptoms with confidence interval (15.298-17.022) under probability of 95%. The largest estimated mean time until death for the factor type of patient is for the N(S-)N(D) which is (18.267) with confidence interval (16.364-20.170) and the lowest estimated mean time until death is for the D(S+) and equals to (8.250) with confidence interval (6.977-9.523) under probability of 95%. And about the Gender the estimated mean time until death for Female is greater than Male which the estimated mean time until death for Female is (16.997) and for the Male is (16.732) with the confidence interval (15.645-18.348) for Female and (15.765-17.699) for

male under probability of 95%. And about the site of TB the estimated mean time until death for the PTB is (17.248) while for the EP is (16.283). the estimated mean time until death for the X-ray result factor is (16.297) for NA, (12.625) for normal and (19.130) for TB with confidence intervals (15.493-17.100) for NA, (8.029-17.221) for normal and (17.360-20.901) for TB under 95% probability.

3-3-1. Kaplan Meier Curve: The plot of K M curves is an important part of survival analysis for each group of interest. In our study, in our study the most important curve is the curve of hazard function.

Interpreting our results is usually with the plot of the cumulative hazard functions for the different groups of treatment (i.e., the two types of questions regarding condition), "Alive", and "Death".

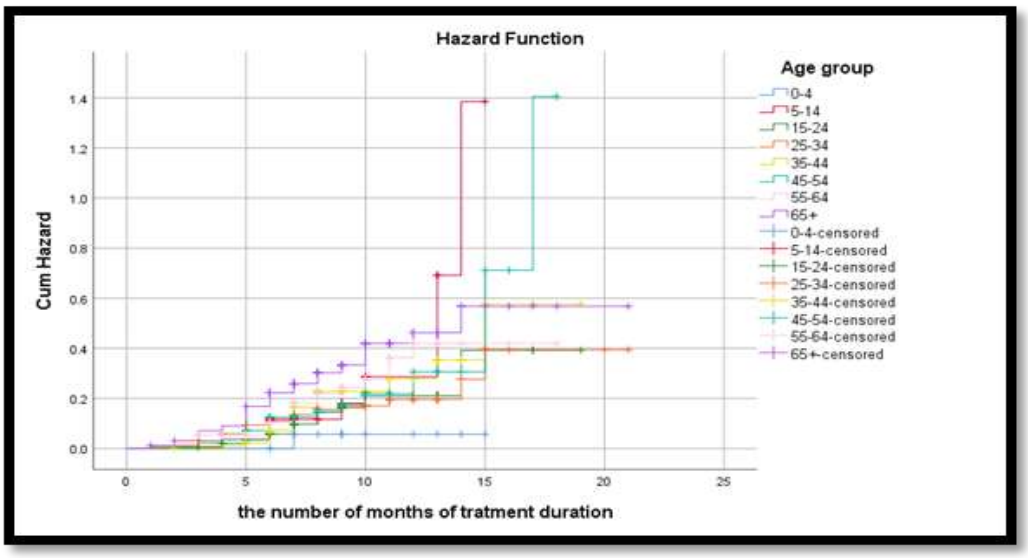

Figure (1): Kaplan Meier curve of age group

In Figure (1) the vertical axis represents the cumulative of hazard and the horizontal axis represent the time to event, of Hazard curve is clear from the plot that the risk of dying increases with time and sometimes stabled then increased again especially after 15 months which increases dramatically. We can see clearly there is not decreasing for the risk of dying. For the ages between 5-14 and 45-54 the risk of dying is greater than another age groups, especially the ages between 45-54 have the highest risk of dying.

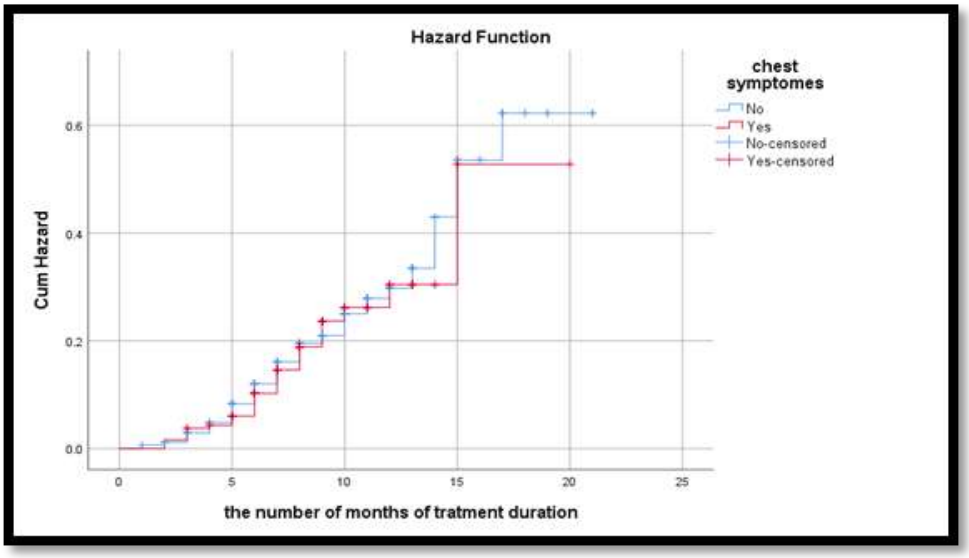

Figure (2): Kaplan Meier curve of chest symptoms

In Figure (2) the vertical axis represents the cumulative of hazard and the horizontal axis represent the time to event, of Hazard curve is clear from the plot that the risk of dying increases with time and sometimes increased dramatically. We can see clearly there is not decreasing for the risk of dying for the both type of patients who have the chest symptoms and those don't have chest symptoms.

3-4. Application of Logistic Regression: Logistic regression is a classification algorithm. It is used to predict a dichotomous outcome based on a set of independent variables (nominal, ordinal, interval or ratio-level independent variables

Table (3): Model summary

Model Summary			
Step	**-2 Log likelihood**	**Cox & Snell R Square**	**Nagelkerke R Square**
1	310.097[a]	.629	.839

Table (3) the table of The Model Summary provides the -2Log Likelihood and pseudo-R^2 values for the full model. The result of Cox & Snell R^2 which suggests that the model explains roughly 62.9% of the variation in the outcome. And the results of Nagelkerke R^2 the variations in the outcome is 83.9%.

3-4-1. Hosmer-Lemeshow test: The Hosmer-Lemeshow test is a goodness of fit test for logistic regression, especially for risk prediction models. The test is only used for binary response variables (a variable with two outcomes like alive or dead, yes or no).

Table (4): Hosmer and Lemeshow test

Hosmer and Lemeshow Test			
Step	**Chi-square**	**Df**	**p-value**
1	35.759	8	.000

In table (4) we can see that the result of the Hosmer and Lemeshow test is statistically significant because p-value is less than 0.05, moreover, this means that our data fits the model.

3-4-2. Variables in the equation for the logistic regression: However, the most important of all outputs of logistic regression analysis is the Variables in the Equation table.

Table (5): Variables in the equation

Variables in the Equation									
		B	S.E.	Wald	df	p-value	Exp (B)	95%C.I.for Exp(B)	
								Lower	Upper
Step 1	Age group	.020	.073	.071	1	.790	1.020	.883	1.177
	Gender	1.159	.275	17.734	1	.000	3.187	1.858	5.466
	chest Symptoms	2.218	.503	19.488	1	.000	9.193	3.433	24.617
	Type of patient	.682	.122	31.022	1	.000	1.978	1.556	2.515
	Site of TB	3.616	.391	85.446	1	.000	37.205	17.281	80.097
	X-ray result	-.300	.285	1.104	1	.293	.741	.424	1.296
	TTT-Outcome	-2.712	.235	132.785	1	.000	.066	.042	.105
	Constant	15.817	1.992	63.038	1	.000	7398875.084		

Table (5) of variables in the equation, provides the parameter estimates (also known as the coefficients of the model) (B), their standard error (S.E.), the Wald statistic (to test the statistical significance) related p-values that are less than α at level (0.05) are statistically significant, otherwise are not, degree of freedoms, and the important Odds Ratio (Exp (B)) for each variable category.

B: These are the values for the logistic regression equation for predicting the dependent variable from the independent variable. They are in log-odds units. Similar to OLS regression, the prediction equation is

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

where p is the probability of being in honors composition. Expressed in terms of the variables used in this example, the logistic regression equation is

$$\log\left(\frac{p(\bar{x})}{1-p(\bar{x})}\right) = 15.817 + 0.02\ \text{Age group} + 1.159\ \text{Gender} + 2.218\ \text{Chest symptoms} + 0.682\ \text{Type of patient} + 3.616\ \text{Site of TB} - 0.3\ \text{X\_ray result} - 2.712\ \text{TTT\_outcome}$$

Furthermore, we can write the logistic regression equation with just significant variables:

$$\log\left(\frac{p(\bar{x})}{1-p(\bar{x})}\right) = 15.817 + 1.159\ \text{Gender} + 2.218\ \text{Chest symptoms} + 0.682\ \text{Type of patient} + 3.616\ \text{Site of TB} - 2.712\ \text{TTT\_outcome}$$

The interpretation of B values:

- Age group-For every one-unit increase in Age group score (so, for every additional point on the Age group), we expect a 0.020 increase in the log-odds of statue1, holding all other independent variables constant.
- Gender- is one of the affecting factors to the risk or death in Tuberculosis diseases. For an increase by 1.159 which is an increase in the risk of the death for patient with (male or female), holding all other independent variables constant.
- Chest Symptoms - For every one-unit increase in Chest Symptoms score (so, for every additional point on the Chest Symptoms), we expect a 2.218 increase in the log-odds of statue1, holding all other independent variables constant.
- Type of patient - For every one-unit increase in Type of patient score (so, for every additional point on the Type of patient), we expect a 0.682 increase in the log-odds of statue1, holding all other independent variables constant.
- Site of TB - For every one-unit increase in Site of TB score (so, for every additional point on the Site of TB), we expect a 3.616 increase in the log-odds of statue1, holding all other independent variables constant.
- X-ray result - For every one-unit increase in X-ray result score (so, for every additional point on the X-ray result), we expect a -0.300 decrease in the log-odds of statue1, holding all other independent variables constant.
- TTT-outcome- For every one-unit increase in TTT-outcome score (so, for every additional point on the TTT-outcome), we expect a -2.717 decrease in the log-odds of statue1, holding all other independent variables constant.

3-5. The application of Cox Regression: The Cox PH model provides an estimate of the effect of treatment on survival after modification for other explanatory variables.

The model building process occurs in seven treatments (Age group, Gender, chest Symptoms, Type of patient, Site of TB, X-ray result, TTT-outcome). In our study, we have 159 Event cases, which its number of deaths and 629 censored; cases, which are patients that still alive. Moreover, if the event has not occurred then the case is said to be censored. The -2 Log Likelihood equals to 1940.362 for the omnibus tests of model coefficients before adding the explanatory variables to the model

3-5-1. Omnibus Tests of Cox Model Coefficients: Omnibus Tests of Cox Model Coefficients are used to verify that the new model (with explanatory variables included) is an improvement over the baseline model.

Table (6): omnibus tests of model coefficients

Omnibus Tests of Model Coefficients									
-2 Log	Overall (score)			Change From Previous Step			Change From Previous Block		
Likelihood	Chi-square	df	p-value	Chi-square	df	p-value	Chi-square	df	p-value
1438.661	556.180	7	.000	501.701	7	.000	501.701	7	.000

The table (6) of omnibus tests of model coefficients there is a statistically significant of the results of chi-square and, this means that the explanatory variables are included in the model. Furthermore, and about the -2 Log Likelihood there is a decreasing with the result of the -2 Log Likelihood before adding the explanatory variables by 501.701 and our -2 Log Likelihood after adding the explanatory variables is 1438.661.

Table (7): Variables in the equation

Variables in the Equation								
	B	SE	Wald	df	p-value	Exp (B)	95% CI for Exp (B)	
							Lower	Upper
Age group	.103	.042	6.178	1	.013	1.109	1.022	1.203
Gender	-.337	.167	4.091	1	.043	.714	.515	.990
chest Symptoms	-.414	.288	2.068	1	.150	.661	.376	1.162
Type of patient	-.013	.069	.037	1	.848	.987	.863	1.129
Site of TB	-.452	.224	4.084	1	.043	.636	.411	.986
X-ray result	-.108	.170	.405	1	.525	.898	.644	1.252
TTT-Outcome	1.828	.126	209.552	1	.000	6.220	4.857	7.967

We can see in the table (7) which shows the coefficients (B), standard errors (SE), value of Wald test, degree of freedom. The quantities E (B) are called hazard ratios (HR). A value of B greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases.

The interpretation of E (B):

- Age is one of the affecting factors to the risk in tuberculosis decease increase by Exp (0.103) = 1.109 which is increase in the risk of the death for patient with (Age group). The significant value 0.013 is less and equals to ($\alpha = 0.05$) so there is significant effect on tuberculosis. And this factor increases in the hazard.
- Although, Gender is one of the affecting factors to the risk or death in tuberculosis diseases. For a decease by Exp (-0.337) equal 0.714 which is a decrease in the risk of the death for patient with (male or female). The p-value is 0.043 so there is evidence of a greater risk of death following tuberculosis in either sex.
- The value of Exp (B) for Chest Symptoms means that the tuberculosis hazard for all patients who had chest Symptoms are 0.661 which is a decrease in the risk of death for patient to have or haven't chest Symptoms. The p-value is 0.150 which there is not significant effect on tuberculosis.
- The estimated hazard in the Type of patient is, Exp (-0.013) equals to 0.987, which is a 98.7% decrease in the risk after adjustment for the other explanatory variables in the model of the death for patient. furthermore, the p-value is 0.848 so, it is not statistically significant.
- The estimation of hazard decease by Exp (-0.452) equals to 0.636 for Site of TB, with p-value is 0.043 which is statistically significant.
- The estimated hazard in the X-ray result is Exp (-0.108) equals to 0.898, which is an 89.9% decrease in the risk. However, the p-value equals to 0.525 is not statistically significant.
- The estimation of hazard increases by Exp (1.828) equals to 6.220 for TTT-outcome, with p-value (0.000) which is statistically significant and the 95% confidence interval for the hazard ratio included.

When x is the vector of all the fixed covariates (Age group, Gender, Chest Symptoms, Type of patient, Site of TB, X-ray result, TTT-outcome)

and β is the corresponding vector of the regression coefficient for the fixed covariates.

$$h_i(t) = h_0(t) * \exp(\beta' x)$$

$$h_i(t) = h_0(t) \exp(0.103\ Age\ group - 0.337\ Gender - 0.414\ Chest\ Symptoms - 0.013\ Type\ of\ patient - 0.452\ Site\ of\ TB - 0.108\ X - ray\ result + 1.828\ TTT - outcome)$$

There are variables which not accepted by the above model because the score statistics with significance values greater than 0.05, which is three factors (Chest Symptoms, Type of patient and X-ray result), has not significant.

The Cox-PH model with significant factor as follows:

$$h_i(t) = h_0(t) \exp(0.103\ Age\ group - 0.337\ Gender - 0.452\ Site\ of\ TB + 1.82\ TTToutcome)$$

3-6. Comparing models: There are many measures to comparing between two or among models in survival functions, it could also be worthwhile to consider the model comparison using the Akaike information criterion and the Bayesian information criterion. In survival analysis, Akaike information criterion and the Bayesian information criterion are the most widely used for comparing models.

Table (8): comparing models with AIC and BIC

Models	No. of parameters	Log Likelihood	AIC	BIC
Logistic regression	7	-546.200	1106.4	1139.1
Cox regression	7	-970.181	1954.4	1987.0

Table (8) indicates the results for the AIC and BIC values which are used to comparing between two models (Cox regression model and Logistic regression model) for selecting the most suitable model to our data of tuberculosis. For each of the models based on two measures; the AIC and BIC were computed; the minimum value of AIC and BIC are selected.

The results shows that Logistic regression model is the best model for our study data of tuberculosis because, it's AIC equals to 1106.4 and BIC equals to 1139.1 are the lowest values in comparison with AIC equal to 1954.4 and Bic equal to 1987 for the Cox regression model.

Conclusions: During analyzing the tuberculosis data and as indicated by the outcomes from the practical part, the following conclusions have been drawn:

1. The comparison tests of Kaplan Meier estimator indicates that in the factor X-ray result TB has the highest value of estimated mean time until death, the Kaplan Meier curves are clearly indicated that the risk of dying increased with the time especially after 15 months.
2. According to the results of the logistic regression of this study indicates the most popular factors that affecting on tuberculosis disease are (Gender, Chest symptoms, Type of patient, Site of TB, TTT-outcome).
3. Depending on the cox regression results of this thesis denotes that the most common factors that have an impact on tuberculosis are (Age group, Gender, Site of TB, TTT-outcome).
4. After comparing the results of AIC and BIC it is concludes that, the Logistic regression model is the most suitable model for our study data set.

References:

1. ARCHER, K. J. & LEMESHOW, S., 2006, Goodness-of-fit test for a logistic regression model fitted using survey sample data. The Stata journal, 6(No.1), pp. 97-105.
2. BALAKRISHNAN, N. & RAO, C. R., 2004, Handbook of Statistics 23- Advances in Survival Analysis. north holand: s.n.
3. EKMAN, A., 2017, Variable selection for the Cox proportional hazards model. Umea university, 21 January.p. 84.
4. FOX, J., 2014, Introduction to Survival analysis. sociology 761.
5. FRANK E. HARRELL, I., 2001, regression modeling strategies. 1 ed. New York: Springer Science.
6. GUO, S., 2010, Survival analysis. 1st ed. New York: Oxford university press, Inc.
7. HARRELL, F. E., 2001, Regression Modeling With Applications to Linear Models, Logistic Regression, and Survival Analysis. Nashville, TN 37232-2637: Springer Science+Business Media New York.
8. HEAGERTY, P., 2005, Survival Analysis. new work: Va/Uw Summer.
9. HOSMER, D. W. & LEMESHOW, S., 2000, applied logistic regression. 2nd ed. canada: Wiley & Sons, inc..
10. HOSMER, D. W., LEMESHOW, S. & MAY, S., 2008, Applied Survival Analsis: Regression Modeling of Time to Event Data. 2nd ed. canada: John wiley & sons, Inc.
11. HOUT, A. V. D., 2017, multi-state survival models for interval censored data. U.S.: taylor & francis group.
12. JONY, 2016. how to. 4 ed. iraq: afsana.
13. KOROSTELEVA, O., 2008, Clinical Statistics-Introducing Clinical Trials, Survival Analysis, and Longitudinal Data Analysis. 1st ed. USA: Jones and Bartlett Series in Mathematics.
14. LEE, E. T. & WANG, J. W., 2003, Statistical Methods for survival data analysis. 2nd ed. NEW YORK: Wiley & Sons, Inc.

15. LIU, X., 2012, Survival analysis 'models and applications'. 1st ed. united kingdom: John Wiley & Sons Ltd.
16. LOKESHMARAN A, &. ,. R. E., 2013, BAYESIAN VARIABLE SELECTION FOR COX'S REGRESSION MODEL. Asia Pacific Journal of Research, pp. 11 - 23.
17. MARK, S., 2007, An Introduction to Survival Analysis. EpiCentre, IVABS, pp. 2-31.
18. MAWLOOD, K. I., 2019, Using Logistic Regression and Cox Regression Models to Studying the Most Prognostic Factors for Leukemia patients. QALAAI ZANIST SCIENTIFIC JOURNAL, 4(2), pp. 705-724.
19. MCDONALD, J. H., 2014, Handbook of Biological Statistics. 3rd ed. USA: Sparky house publishing.
20. MOORE, D. F., 2016, Applied Survival Analysis Using R. Switzerland: springer international publishing.
21. SCHMIDT, P. & WITTE, A. D., 1998, Predicting Recidivism Using Survival Models. 1st ed. London: Springer verlag.
22. WIENKE, A., 2011, Frailty models in survival analysis. 1st ed. USA: Taylor & Francis group, LLC.
23. XIN, X., 2011, A Study of ties and Time-Varying Covariates in Cox Proportional Hazards Model. University of Guelph,, pp. 1-43.

المصادر العربية:

١. التلباني، ش، ٢٠١١، دراسة مقارنة بين نموذج الانحدار اللوجيستي ونموذج انحدار كوكس لدراسة أهم العوامل الاقتصادية والديموغرافية المؤثرة على معرفة واتجاهات الشباب نحو قضايا الصحة الإنجابية، رسالة دكتوراه غير منشوره، جامعة أبو بكر بلقايد، تلمسان، الجزائر.
٢. فتيحة، ب.، ٢٠١٥، COX تقدير مدة البحث عن الشغل لحاملي شهادات تكوين المهني باستعمال نموذج الأخطار النسبية، مجلة العلوم الاقتصاد والتسيير والتجارة (جامعة بغداد)، ص١١-٣٣.