Statistical analysis of COVID-19 in Erbil-Kurdistan/Iraq: Using parametric survival Models

Kurdistan Ibrahim Mawlood [1], Sarween Asaad othman [2]

[1, 2] Salahaddin University, Erbil College of Administration and Economics - Statistics & Informatics Department

***Abstract*—The basic idea of this study focused on using three parametric survival models (Weibull Model, Lognormal Model, Log Logistic Model) instead of nonparametric ones for modeling and estimating affecting factor parameters of Covid-19 patient's.**

The data set of this study was obtained from Arzheen private hospital in Erbil city.

The results indicated that, the models have not reached to the same variables that have an impact on the Covid-19 patient's data in Erbil city. Moreover, the results indicated that the Lognormal Model describes the data well or give the best fit for our data of Covid-19 survival patients in Erbil city. Comparison among models were done based on two model selecting criterion; Akaike Information Criterion (AIC) and Bayesian information criterion (BIC). The results obtained by utilizing the statistical packages (Mat-lab V. 14, Stata V. 16 and STATGRAPHICS V. 19).

***Keywords*—Survival Analysis, parametric survival models, Akaike Information Criterion (AIC), Covid-19**

I. Introduction

The duration of a subject's survival from a beginning point to an ending points measured by their survival time. The concept of survival need not be taken literally. Here, survival indicates that a person is in a situation that corresponds to the default situation. The situation won't change until an interesting thing happens. Failure is the important occurrence that signifies the end of the period of survival. Failure usually involves death or going through a bad experience. Failure typically results in death or a bad experience. Failure, however, can sometimes have a positive outcome, such a disease cure. Failure may also be known as the event or, death if death represents the failure (PINTO, (2015))

2. Background Information

This section reviews the foundation of survival data analysis with an essential issue in health, which is covid-19 disease including the fundamental concepts and basic methods in modeling survival data in the presence of censored observations. In addition, exploration and description of parametric modeling (Weibull model, Log-normal model, and Log-Logistic model) explained. Also, (AIC) and (BIC) criterions were used to select the best model between parametric models.

2.1: Covid-19 disease

Covid-19 is the disease caused by the emerging coronavirus also known as SARS-CoV-2. This novel virus was first detected by who on December 31, 2019, after a cluster of cases of viral pneumonia were reported in Wuhan, People's Republic of China.

Coronaviruses are a widespread family known to cause illnesses ranging from the common cold to more severe illnesses such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS).

Fever, a dry cough, and tension are the most typical covid-19 symptoms. Various rash patterns, loss of taste and smell, nasal congestion, conjunctivitis, sore throat, headache, muscular or joint discomfort, nausea or vomiting, diarrhea, tremors, or disorientation are other, less frequent symptoms that some people may have.

2.2: Survival analysis

In the last few decades, applications of the statistical methods for survival data analysis have been extended beyond biomedical research. Survival analyses have been used for data involving time to a certain event such as death, the occurrence of a disease, or the relapse of a condition. Survival analysis dates back to the 17th century with the first life table ever produced by John Graunt in 1662.(CAMILLERI, 2019). Additional fields like engineering, political science, corporate management, and economics have made extensive use of survival analysis. (LIU, 2012).

In clinical research, the survival time is used. Depending on the sector of application, survival time may also be referred to as time to event, life time, duration time, or failure time. These techniques are widely used in a variety of sectors, including public health, epidemiology, the social sciences, economics, and engineering, in addition to medical research. (LAWLESS,

Received 30 july 2022; Accepted 4 September 2022; Regular research paper: Published 15 August 2023

2003). The subject will remain in that state until an event of interest occurs. Failure is the event of interest that marks the end of the survival time. (PINTO, 2015)

2.3: Survival Function

Survival analysis involves with the implementation of certain statistical methods to model and analyses survival time data. The probability density function (pdf) defined by,$f(t)$ which is written as: (LEE & WANG, 2003)

$$\boldsymbol{f(t)=\frac{dF(t)}{dt}} \qquad \dots(\mathbf{2.1})$$

and cumulative distribution function (C.D.F.) define by $F(t)$ describes the probability the time to event (T) is lesser or equal compared to a fixed time (t) and is shown as:

$$\boldsymbol{F}\ \mathbf{(t)} = \boldsymbol{p(T \leq \mathrm{t})} \qquad \mathbf{\dots\ (2.2)}$$

From as, the survival function, the probability the time to event (T) is greater compared to a fixed time (t), that can be represented as:

$$\mathbf{S\ (t) = Pr}\ (\mathbf{T} > t) = \int_{\boldsymbol{t}}^{\infty} \boldsymbol{f(u)du} = \mathbf{1 - F(t), t \geq 0} \qquad \mathbf{\dots\ (2.3)}$$

Which means, the probability that an individual survives beyond time (t). Note that the survival function S(t) is a monotonic non-creasing continuous function with S(t) = 1 and $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

2.4: Hazard Function

That represents an individual the probability condition of death at time t after survival time, the hazard function that is defined by $h(t)$, which is written as:

$$\boldsymbol{h(t)} = \lim_{\boldsymbol{\Delta t \to 0}} \frac{\boldsymbol{p(t \leq T + \Delta t | T > t)}}{\boldsymbol{\Delta t}} \qquad \mathbf{\dots\ (2.4)}$$

Representing the probability that an individual fails within a small interval $(t, t + \Delta t)$, given that the individual survived to the beginning of the interval.

That relationship between $s(t)$ and $h(t)$ is shown as:

$$\boldsymbol{h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{-d \log s(t)}{dt}} \qquad \dots(\mathbf{2.5})$$

$f(t)$ is the density function which is the fraction of the initial group for whom the event will occur during the time interval at t adjusted for the width of the time interval (LAWLESS , (2002))

2.6: Censoring Data

Another critical feature of survival analysis is that it can state censoring, which occurs when there is some information about a patient's survival time is known, but the accurate survival time is unknown. Data can be censored to the right or left censored. The most common form of censoring is right censoring. In survival analysis right, censored data are a common problem in estimating survival and hazard function (MAN, (2014)).

2.8: Parametric distribution models for time to event

The study of times, events, or lifetimes is called survival analysis. Parametric models, including the Weibull distribution, lognormal distribution, log-logistic distribution, and associated methods of inference, are used to depict lifespan distributions and their relationship to explanatory variables, or covariates. It is described and illustrated how to use linear models for log lifetimes in parametric regression analysis (accelerated failure time models). (LAWLESS, 2003).

The accelerated failure time model (AFT model) is a parametric model that offers an alternative to the popular proportional hazards models. In contrast to a proportional hazards model, which assumes that the effect of a covariate is to multiply the risk by a constant, an AFT model assumes that the effect of a covariate is to either accelerate or decelerate the life course of a disease by a constant.

For $i = 1, \dots, n$ let T_i be the failure time for the ith individual and let X_i be the associated p-vector of covariates. The accelerated failure time model denotes that

$$\mathbf{Log}\, \boldsymbol{T_i} = \boldsymbol{B_0 + B_i X_i + \varepsilon_i} \qquad \dots(\mathbf{2.6})$$

where B_0 is a p-vector of unspecified regression parameters and ε_i $(i = 1, \dots, n)$ are independent error terms with the common, but completely undefined, distribution. (Qi, (2009))

Four AFT model classifications were included in this research; the Weibull model, log-logistic model, and log-normal model.

2.8.2: Weibull distribution

An expansion of the exponential distribution is the Weibull distribution. It offers a wider range of applications since, unlike the exponential distribution, it does not presuppose a constant hazard rate. The distribution was first introduced by Weibull (1939), and Weibull (1951) further examined its relevance to many failure situations. It has since been utilized in numerous studies of dependability and human disease mortality. (LEE & WANG, 2003)

Let the survival time T a random variable following the Weibull distribution, then its probability density function (pdf) defined by $f(t)$,

$$\boldsymbol{f(t) = \alpha\beta(\beta t)^{\alpha - 1}}\mathbf{exp}\{\boldsymbol{-(\beta t^{\alpha})}\} \qquad \dots(\mathbf{2.7})$$

the cumulative distribution denoted by F (t),

$$\boldsymbol{F}\ \mathbf{(t) = 1\text{-}\ exp}(\boldsymbol{\beta t^{\alpha}}) \qquad \dots(\mathbf{2.8})$$

Then, the survival function denoted by S (t),

$$\boldsymbol{S}\ \mathbf{(t) = exp}\{\boldsymbol{-(\beta t^{\alpha})}\} \qquad \dots(\mathbf{2.9})$$

So that, the hazard function denoted by $h(t)$ is defined as

$$\boldsymbol{h(t) = \alpha\beta(\beta t)^{k-1}} \qquad \dots(\mathbf{2.10})$$

The Weibull distribution is defined where t > 0, is the support of the distribution by two parameters, α > 0, which determines the distribution is shape and its known as the shape parameter, and β > 0 that also determines the distribution is scaling and is also known as the scale parameter (WEBULL, (1951)).

The exponential case is when the shape parameter α =1, the hazard rate remains constant as time increases, when α <1, the hazard rate decreases with time and when α >1, the hazard rate increases with time. So, the Weibull distribution can be used to model survival data of individuals with increasing and decreasing as well as constant risk (LEE & WANG, 2003).

2.8.3: Lognormal distribution

The distribution of a random variable whose logarithm is normally distributed is known as the lognormal distribution. A number of diseases can be roughly approximated by the distribution since it is very positively skewed. (LEE & WANG, 2003).

The lognormal distribution is frequently used to explain occurrences where the rate initially climbs and then consistently decreases afterwards. However, this distribution only functions properly when there is no censoring. when the survival data contain a significant amount of censored observations.

The probability density function and the survival function are, respectively,

$$f(x) = \frac{1}{t\sigma\sqrt{2\pi}} exp\left[- \frac{1}{2\sigma^2}(log\, t - \mu)^2\right] \quad t > 0 \quad ...(2.11)$$

And $$S(t) = \frac{1}{\sigma\sqrt{2\pi}}\int_t^{\infty} \frac{1}{x}\ exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right] dx$$

Let *a* = exp (- *μ*). Then - *μ* = log *a* can be written as,

$$S(t) = -\ \phi\ \left(log \frac{at}{\sigma}\right) \quad ...(2.12)$$

The cumulative distribution denoted by *F* (t),

$$F\ (t) = \phi\ \left(log \frac{at}{\sigma}\right) \quad ...(2.13)$$

So, then the hazard function denoted by $h(t)$ is define as,

$$h(t) = \frac{f(t)}{S(t)} = \frac{(1/t\sigma\sqrt{2\pi})\ \exp[-\ (\log at)^2/2\sigma^2]}{1 -\ \phi\ (log\ at/\sigma)} \quad ...(2.14)$$

Let the survival time *T* be a random variable, where t > 0 is the support of the distribution, Φ (·) is the standard normal c.d.f, and $\sigma^2 > 0$ and $\mu > 0$ are the parameters. Note that if T is log-normal with parameters *μ* and σ^2, then Y = log T is normal with mean *μ* and variance σ^2 (LAWLESS, 2003).

2.8.4: Log-logistic distribution

An appropriate alternative for describing a lifetime event that the lognormal function can capture is the log- logistic distribution. Generally, the log-logistic distribution has heavier tails than the lognormal distribution. The log-logistic distribution further differs from the lognormal model in having closed forms for the cumulative distribution, survival, and hazard functions, making it a better distributional form to describe a lifetime event that reverses direction. (LIU, (2012)).

The probability density function of the log- logistic distribution is as follows,

$$f(t) = \frac{\lambda k t^{k-1}}{(1 + \lambda t^k)^2} \quad ...(2.15)$$

And the cumulative distribution is,

$$F\ (t) = 1 - \frac{1}{1+\lambda t^k} \quad ...(2.16)$$

survival function of the log- logistic distribution is,

$$S(t) = 1 - F(t) = \left[1 + \lambda t^k\right]^{-1} \quad ...(2.17)$$

where t > 0 is the support of the distribution, and $\lambda > 0$ and κ > 0 are the parameters, where λ is the rate parameter and κ is the shape parameter (LAWLESS, 2003).

2.9: Maximum Likelihood Estimation for a Parametric Distribution

Maximum likelihood methods are very important and basic for analysis of accelerated test data. They are frequently utilized with many sorts of data and models. Estimates and confidence intervals for model parameters and other intriguing values are provided using these methods. (HOUT, (2017)).

If $x_1, x_2, \dots, x_n$ are independent and identically distributed observations from a distribution that depends on the unknown parameters $\theta_1, \theta_2, \dots, \theta_m$ the likelihood function is defined as (RODRIGUEZ, (2010)):

$$L(\theta|x) = Pr(X_1 = x_1, x_2, \dots, X_n = x_n) = f(x|\theta) = f(x_1|\theta)\ \times \dots \times\ f(x_n|\theta)$$

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i\,|\theta) \quad ...(2.18)$$

2.10: Model Selection

It is common to choose a model that performs the best on a dataset or to estimate model performance using Akaike and Bayesian Information Criterion.

2.10.1: Akaike's Information Criterion

The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data and an estimator predictor of prediction error. AIC calculates the quality of each model in comparison to the other models given a set of data models; the lower the information a model loses, the greater its quality. (BURNHAM & ANDERSON, (2002))

The AIC is a practical method of comparing an estimated model's complexity with how well it fits the data. The AIC is determined utilizing:

$$\text{AIC} = -2\ log\ (likelihood) + 2\ (p + k) \quad ...(2.19)$$

where p is the number of parameters, k = 2 for the Weibull, log logistic, and log normal models.

2.10.2: Bayesian Information Criterion

The criteria were formulated by Schwartz in 1978 to perform as an asymptotic approximation to a transformation of the Bayesian posterior probability of a proposed model. (NEATH & CAVANAUGH, (2012)).

the model corresponding to the minimum value of BIC is selected.

$$\text{BIC} = -2\text{logLikelihood}\ + 2 * \text{logN} * \text{k} \quad ...(2.20)$$

This criterion is based on the log-likelihood L, the number of parameters in the distribution (k), and the total number of observations (N). For each selected distribution, compute.

3. Results and Discussions:

Is in this section four parametric models were used for survival analysis data. Also; all the corresponding results have been given and a comparison between the two models has been done. Two statistical measures (AIC and BIC) were used to select the best model fit for our data. The following programs were used to analyze the data:

1. Mat-lab V. 14.
2. Stata V. 16.
3. STATGRAPHICS V, 19.

3.1 Data Collection

The data for this study of covid-19 have been collected from Arzheen private hospital in Erbil city. The data consisted of 350 cases for all patients with covid-19 who were registries and treated at Arzheen private hospital, corona department, during 1st September 2020 through 30th June 2021, of those patients 44 died during the study and 306 survival alive. The survival time are measured in days from the first day that patient admitted to hospital to the date of death or the last visit to the hospital

The following covariate were included as prognostic factors in the study had been collected for all patient:

The patient related variables (Age, Gender, Smoker).

Clinical related variables (Peripheral oxygen saturation (SPO2), White blood cell (WBC), Lymphocyte, Monocyte,

Hemoglobin (Hb), Red blood cell (RBC), Platelet (PLT), C reactive protein (CRP), Ferritin, Lactate dehydrogenase (LDH), Heart beat (HR), Blood Pressure, D Dimer).

Chronic diseases (Hypertension, Diabetes mellitus, Chronic lung disease, Cardiovascular disease).

Dependent variable, this is the outcome of treatment of a patient enrolled at a corona department in Arzheen hospital, these outcomes were either died or completed treatment and survival alive. Specific variables used and their categorization is shown in table 3.1 below.

Table 3.1 variable categorization

Variable names	Categorization	N	No. of Alive	No. of Death
Age grouped	<=18	2 (0.6%)	2	0
	19 - 39	61 (17.4%)	56	5
	40 - 59	163 (46.6%)	138	25
	60- 79	121 (34.6%)	107	14
	80+	3 (0.9%)	3	0
Gender	1=male	196 (56%)	174	22
	2=female	154 (44%)	132	22
Smoker	0=non-smoker	261 (74.6%)	241	20
	1=smoker	89 (25.4%)	65	24
SPO2	1=Low	232 (66.3%)	193	39
	2=Normal	118 (33.7%)	113	5
	3=High	0 (0%)	0	0
WBC	1=Low	9 (2.6%)	4	5
	2=Normal	192 (54.9%)	183	9
	3=High	149 (42.6%)	119	30
Lymphocyte	1=Low	137 (39.1%)	117	20
	2=Normal	213 (60.9%)	189	24
	3=High	0 (0%)	0	0
Monocyte	1=Low	13 (3.7%)	7	6
	2=Normal	332 (94.9%)	296	36
	3=High	5 (1.4%)	3	2
Hb	1=Low	341 (97.4%)	297	44
	2=Normal	8 (2.3%)	8	0
	3=High	1 (0.3%)	1	0
RBC	1=Low	103 (29.4%)	91	12
	2=Normal	195 (55.7%)	164	31
	3=High	52 (14.9%)	51	1
PLT	1=Low	15 (4.3%)	8	7
	2=Normal	333 (95.1%)	296	37
	3=High	2 (0.6%)	2	0
CRP	1=Normal	3 (0.9%)	3	0
	2=High	347 (99.1%)	303	44
Ferritin	1=Low	12 (3.4%)	12	0
	2=Normal	100 (28.6%)	99	1
	3=High	238 (68%)	195	43
LDH	1=Low	27 (7.7%)	27	0
	2=Normal	93 (26.6%)	85	8
	3=High	230 (65.7%)	194	36
HR	1=Low	55 (15.7%)	50	5
	2=Normal	73 (20.9%)	67	6

	3=High	222 (63.4%)	189	33
Blood Pressure	1=Low	122 (34.9%)	109	13
	2=Normal	213 (60.9%)	188	25
	3=High	15 (4.3%)	9	6
D Dimer	1=Normal	116 (33.1%)	116	0
	2=High	234 (66.9%)	190	44
hypertension	0=No	146 (41.7%)	132	14
	1=Yes	204 (58.3%)	174	30
diabetes mellitus	0=No	214 (61.1%)	195	19
	1=Yes	135 (38.9%)	110	25
chronic lung disease	0=No	206 (58.9%)	192	14
	1=Yes	144 (41.1%)	114	30
cardiovascular disease	0=No	290 (82.9%)	265	25
	1=Yes	60 (17.1%)	41	19
Status	0= Alive (censored)	306 (87.4%)		
	1= Death	44 (12.6%)		
Treatment Duration	The number of days of treatment duration			

Table 3.1 shows the age of diagnosis ranged from 15 to 85 years, most of the patients (46.6%) were at the age group of 40 to 59, out of a total of 163 cases including 25 patients died and 138 cases remained alive. A total of 196 patients (56%) were male and 154 patients (44%) were female. Only 3 patients were recorded to have normal CRP and non-of them had died from the disease, all the death cases that have been recorded have had high blood inflammation, which means CRP elevated in patients with covid-19. In total of 44 death cases in the study 43 of them had elevated Ferritin, 36 had elevated LDH and 33 had abnormally high HR. Patients that had a normal blood pressure were 213 patients (60.9%) and 25 are died out of 44 cases. Regarding HR and D Dimer the results show all the dead cases were 44 patients all had low Hb and high D Dimer. The results show that all the dead cases that were 44 patients all had a high D Dimer in a total of 234 patients (66.9%). The majority of the patients that had cardiovascular disease were 290 patients (82.94%) and 25 patients died out of them.

3.6.1: Fitting Model

These models (Weibull Model, Lognormal Model, and Log-Logistic Model) are being used to illustrate the effects of prognostic factors on survival in the study. The fit of the models was verified using the survival function of the fitted data measured.

Model fitting is a function that takes a set of parameters and outputs predicted data sets and a "error function" that outputs a number indicating the discrepancy between the anticipated values and the actual values for every given set of model parameters. Using 20 treatments, we applied the models to our data (patient related variables, clinical related variables and chronic diseases). The outputs are shown in tables (3.13, 3.14, 3.15, 3.16).

β: is a Regression coefficient explains the amount and direction of the association between a predictor and a response variable; coefficients are the amounts that are multiplied by the term values in a regression equation.

The sign of a regression coefficient indicates whether each independent variable and the dependent variable are positively or negatively correlated; a positive coefficient means that as the independent variable's value rises, the dependent variable's mean tends to increase as well, and a negative coefficient means that as the independent variable increases, the dependent variable tends to decrease.

Table 3.14 Analysis of Fitting Weibull Model

Parameter Estimates in Weibull Model

Parameter	β	Standard Error	95% Confidence Limits		Chi-Square	Df	P-Value
			Lower	Upper			
Constant	2.469	0.911	0.683	4.256			
Age	-0.209	0.052	-0.311	-0.107	15.414	1	0.000
Gender	0.126	0.076	-0.024	0.276	2.577	1	0.108
Smoker	0.247	0.088	0.074	0.420	7.877	1	0.005
SPO2	0.031	0.072	-0.110	0.171	0.109	1	0.047
WBC	0.129	0.065	0.001	0.257	3.484	1	0.062
Lymphocyte	-0.090	0.071	-0.229	0.049	1.401	1	0.237
Monocyte	-0.606	0.167	-0.934	-0.279	13.573	1	0.000
Hb	-0.146	0.170	-0.479	0.187	0.657	1	0.417
RBC	0.041	0.051	-0.059	0.141	0.480	1	0.488
PLT	-0.394	0.165	-0.718	-0.070	5.540	1	0.019
CRP	0.424	0.337	-0.237	1.085	1.086	1	0.029
Ferritin	-0.083	0.067	-0.214	0.047	1.456	1	0.022
LDH	0.097	0.052	-0.005	0.199	3.169	1	0.075
HR	0.028	0.046	-0.061	0.117	0.231	1	0.631
Blood Pressure	0.256	0.065	0.129	0.384	14.286	1	0.000
D Dimer	0.331	0.079	0.176	0.486	15.723	1	0.000
Hypertension	0.267	0.073	0.123	0.410	12.286	1	0.001
diabetes mellitus	0.353	0.073	0.210	0.496	22.751	1	0.000
chronic lung disease	0.345	0.073	0.201	0.488	21.366	1	0.000
cardiovascular disease	0.289	0.101	0.091	0.487	8.671	1	0.003

Scale	0.564	0.024	0.519	0.613

The survival function for Weibull model is:

$S\ (t;\ \mathrm{X}) = \exp\ (-t^{k}\ [\exp\ (-\ b_0 - b_1x_1 - b_2x_2\ \dots\ -\ b_n\ x_n)])$

When the linear regression coefficient $\boldsymbol{\beta}$ associated with predictor $\boldsymbol{X}$ is the vector of all the fixed variables and $\boldsymbol{k}$ is the scale parameter. Hence, it can write the Weibull Distribution equation with just significant variables:

$S\ (t;\ \mathrm{X}) = \exp\ (-t^{0.564}$ [exp (– 2.469 + 0.209 Age – 0.247 Smoker – 0.031 SPO_2 + 0.606 Monocyte + 0.394 PLT – 0.424 CRP + 0.083 Ferritin – 0.256 Blood Pressure – 0.331 D Dimer– 0.267 Hypertension – 0.353 diabetes mellitus – 0.345 chronic lung disease – 0.289 cardiovascular disease)]).

And the survival model can be written as:

$\mathrm{Log}\ T_i = B_0 + B_iX_i + \varepsilon_i$

We fit the survival model above to the data in covid-19 disease.

Log T_i = 2.469 – 0.209 Age + 0.126 Gender + 0.247 Smoker + 0.031 SPO_2 + 0.129 WBC – 0.090 Lymphocyte – 0.606 Monocyte – 0.146 Hb + 0.041 RBC – 0.394 PLT + 0.424 CRP – 0.083 Ferritin + 0.097 LDH + 0.028 HR + 0.256 Blood Pressure + 0.331 D Dimer + 0.267 Hypertension + 0.353 diabetes mellitus + 0.345 chronic lung disease + 0.289 cardiovascular disease + ε_i

we take a look at how to interpret each regression coefficient.

- In the result, in table (3.14) the second column presents the regression coefficient. The sign of the coefficients is an important issue to consider.

In patient related variables (Age and Smoker)

The age variable coefficient is negative (β = - 0.209), which means the risk of the death will decrease for covid-19 diseases, and chi-Square test value is equal (15.414) and the smoker coefficient is positive (β = 0.247), so the risk of death which is an increase, the chi-Square test value is equal to (7.877).

Hence, p-value is showing the significance of the explanatory variables. The significant values for (Age and Smoker) variables are (0.000 < = 0.05 and 0.005 < = 0.05), it means that these treatments are significant.

- (Blood Pressure, SPO_2, PLT, CRP, Monocyte and D Dimer) are variables in clinical related variables that their coefficients are statistically significant.

The coefficient values of (Blood Pressure, CRP and D Dimer), equal to (β =0.256, 0.424 and 0.331), the factors affect in disease at diagnosis had an increased risk for death in term of hazard ratio, with chi-Square test values equal (14.286, 1.086 and 15.723) respectively.

The regression coefficient for three variables (Monocyte, PLT and Ferritin) equal to (β = - 0.606, - 0.394 and -0.083), which are decrease in the risk after adjustment for the other explanatory variables in the model of the survival for patient, and their chi-Square test values are (13.573, 5.540 and 1.456).

- In this model even though the results showed for four variables in chronic diseases (Hypertension, diabetes mellitus, chronic lung disease and cardiovascular disease), all these variables are increase in the risk of the death for patient, because the estimate of the coefficient is sign positive (β = 0.267, 0.353, 0.345 and 0.289), with chi-square test values equal to (12.286, 22.751, 21.366 and 8.671). However, the resulting p- value for all in chronic diseases are smaller than common significance, it means that these treatments are significant in the covid-19 disease.

Also, in the table 3.14 if the value of chi-square column is considered as a significant factor; then, diabetes mellitus will be one of the significant factors in our study; because it has a greater value in chi-square test column (22.751) with significant value of (0.000<=0.05).

- For the remained variables such as (Gender, WBC, Lymphocyte, Hb, RBC, LDH, HR), P-value and hence are non-significant as in the Weibull Model.

Table 3.15 Analysis of Fitting Log-normal Model

Parameter Estimates in Lognormal Model

Parameter	β	Standard Error	95% Confidence Limits		Chi-Square	Df	P-Value
			Lower	Upper			
Constant	2.099	0.970	0.197	3.999			
Age	-0.168	0.054	-0.274	-0.063	9.681	1	0.002
Gender	0.116	0.075	-0.031	0.263	2.390	1	0.122
Smoker	0.213	0.088	0.041	0.385	5.861	1	0.016
SPO2	-0.055	0.075	-0.202	0.093	0.532	1	0.041
WBC	0.041	0.070	-0.096	0.178	0.349	1	0.555
Lymphocyte	-0.090	0.074	-0.234	0.055	1.476	1	0.224
Monocyte	-0.392	0.170	-0.726	-0.059	5.328	1	0.021
Hb	-0.007	0.189	-0.377	0.362	0.002	1	0.963
RBC	0.061	0.056	-0.048	0.170	1.195	1	0.274
PLT	-0.255	0.179	-0.604	0.095	2.054	1	0.152
CRP	0.178	0.372	-0.552	0.908	0.228	1	0.633
Ferritin	-0.038	0.069	-0.173	0.098	0.296	1	0.587

LDH	0.108	0.056	-0.001	0.218	3.740	1	0.053
HR	0.049	0.048	-0.044	0.142	1.057	1	0.304
Blood Pressure	0.191	0.071	0.051	0.331	7.088	1	0.008
D Dimer	0.253	0.080	0.097	0.410	9.974	1	0.000
Hypertension	0.188	0.080	0.0316	0.344	5.528	1	0.019
diabetes mellitus	0.302	0.076	0.154	0.450	15.735	1	0.000
chronic lung disease	0.336	0.078	0.183	0.489	18.202	1	0.000
cardiovascular disease	0.282	0.102	0.083	0.481	7.733	1	0.005
Scale parameter	0.630	0.026	0.582	0.683			

The survival function for log-normal model is:

$S\ (t;\ \mathrm{X}) = \phi\ [b_0 + b_1 x_1 + b_2 x_2 + \cdots - k \log(t)]$

Where **β** is termed as the regression coefficients of predictor variables ***X***, and ***k*** is the scale parameter. Hence, the survival function for the log-normal Distribution equation with only significant variables can be written as follows:

$S\ (t;\ \mathrm{X})$ = ϕ [2.099 – 0.168 Age + 0.213 Smoker – 0.055 SPO$_2$
– 0.392 Monocyte + 0.191 Blood Pressure +0.253 D Dimer
+0.188 Hypertension +0.302 diabetes mellitus
+ 0.336 chronic lung disease + 0.282 cardiovascular disease
– 0.630 log(t)].

The survival model is:

$$\mathrm{Log}\ T_i = B_0 + B_i X_i + \varepsilon_i$$

We fit the survival model above to the data in covid-19 disease

Log T_i = 2.099 – 0.168 Age + 0.116 Gender + 0.213 Smoker
– 0.055 SPO$_2$ + 0.041 WBC – 0.090 Lymphocyte – 0.392 Monocyte
– 0.007 Hb + 0.061 RBC – 0.255 PLT + 0.178 CRP – 0.038 Ferritin
+ 0.108 LDH + 0.049 HR + 0.191 Blood Pressure + 0.253 D Dimer
+ 0.188 Hypertension + 0.302 diabetes mellitus + 0.336 chronic lung
disease + 0.282 cardiovascular disease + ε_i

In table 3.15 the interpretation of coefficient (β) values:

- The results showed that the patient related variables (Age and Smoker) have significant effects on the disease, with the p-values equal to (0.002 < = 0.05 and 0.016 < = 0.05),

the chi-Square test value for age variable is equal (9.681) with regression coefficient equal to (β = - 0.168). This indicates that the variable has high risk of death.

Although, the coefficient value for Smoker equal to (β = 0.213), The estimation of coefficient increases in the risk of the death, and its chi-Square test value equal to (5.861).

- in clinical related variables (Blood Pressure, SPO$_2$, Monocyte and D Dimer) are statistically significant factors in the model because their p-values are less than (0.05), This means they have an effect on disease.

it seems like the result of variables (SPO$_2$ and Monocyte) are affecting in covid-19 diseases by coefficients (β = - 0.055 and - 0.392), which is decrease in the risk of the death for patient, with chi-Square test values equal to (0.532 and 5.328).

However, (Blood Pressure and D Dimer) from the same in clinical related variables are two factors affect in the disease increase by coefficient equal to (β = 0.191 and 0.253), which are increase in the risk of the death for patient, and chi-Square test values are equal (7.088 and 9.974), respectively.

- The result shows all four variables in chronic diseases (Hypertension, diabetes mellitus, chronic lung disease and cardiovascular disease), have p-value that are less than the significance level of (0.05), these results indicate that they are statistically significant, and increase the risk of death, the coefficients for chronic diseases variables equal to (β = 0.188, 0.302, 0.336 and 0.282), and their chi-square test values are equal to (5.528, 15.735, 18.202 and 7.733), respectively.
- The remained variables that do not affect in the covid-19 disease are (Gender, WBC, Lymphocyte, Hb, RBC, PLT, CRP, Ferritin, LDH, HR), because their p-value are greater than (0.05) which indicates that there is not enough evidence to conclude that they do not increase or decrease.

Table 3.16 Analysis of Fitting Log-logistic Model

Parameter Estimates in Log-logistic Model

Parameter	β	Standard Error	95% Confidence Limits		Chi-Square	Df	P-Value
			Lower	Upper			
Constant	2.276	0.977	0.362	4.191			
Age	-0.164	0.054	-0.269	-0.059	9.207	1	0.002
Gender	0.121	0.076	-0.028	0.271	2.505	1	0.114
Smoker	0.271	0.093	0.089	0.454	8.258	1	0.004
SPO2	-0.067	0.078	-0.219	0.086	0.726	1	0.394
WBC	0.022	0.070	-0.116	0.159	0.099	1	0.753
Lymphocyte	-0.075	0.075	-0.222	0.071	1.022	1	0.312
Monocyte	-0.426	0.199	-0.817	-0.035	4.590	1	0.032
Hb	-0.023	0.172	-0.361	0.315	0.017	1	0.896
RBC	0.075	0.056	-0.035	0.184	1.760	1	0.185
PLT	-0.306	0.203	-0.703	0.091	2.320	1	0.128
CRP	0.159	0.338	-0.503	0.822	0.221	1	0.638
Ferritin	-0.038	0.070	-0.175	0.099	0.284	1	0.594
LDH	0.116	0.057	0.005	0.227	4.122	1	0.042
HR	0.043	0.049	-0.052	0.138	0.	1	0.

					795		373
Blood Pressure	0.202	0.074	0.056	0.348	7.262	1	0.007
D Dimer	0.255	0.079	0.100	0.410	10.198	1	0.001
Hypertension	0.174	0.082	0.013	0.335	4.491	1	0.034
diabetes mellitus	0.310	0.077	0.159	0.461	15.946	1	0.000
chronic lung disease	0.337	0.081	0.179	0.495	17.171	1	0.000
cardiovascular disease	0.247	0.105	0.042	0.452	5.632	1	0.018
Scale	0.368	0.017	0.336	0.403			

The survival function for Log-logistic model is:

$$S\ (t;\ \mathrm{X}) = [1 + t^{k} * exp(-b_0 - b_1x_1 - b_2x_2 \ldots - b_nx_n)]^{-1}$$

Where **T** is the time, $\boldsymbol{X}$ is the vector of covariate and $\boldsymbol{\beta}$ the vector of regression coefficient and $\boldsymbol{k}$ is the scale parameter. Furthermore, we can write the Log-logistic Distribution equation with just significant parameters:

S (t; X) = [1+ $t^{0.368} * exp$ (– 2.276 + 0.164 Age – 0.271 Smoker
+ 0.426 Monocyte – 0.116 LDH
– 0.202 Blood Pressure – 0.255 D Dimer
– 0.174 Hypertension – 0.310 diabetes mellitus
– 0.337 chronic lung disease – 0.247 cardiovascular
disease) $]^{-1}$

The survival model is:

$$\text{Log } T_i = B_0 + B_iX_i + \varepsilon_i$$

We fit the survival model above to the data in covid-19 disease

Log T_i = 2.276 – 0.164 Age + 0.121 Gender + 0.271 Smoker
– 0.067 SPO_2 + 0.022 WBC – 0.075 Lymphocyte – 0.426 Monocyte
– 0.023 Hb + 0.075 RBC – 0.306 PLT + 0.159 CRP – 0.038 Ferritin
+ 0.116 LDH + 0.043 HR + 0.202 Blood Pressure + 0.255 D Dimer
+ 0.174 Hypertension + 0.310 diabetes mellitus + 0.337 chronic lung
disease + 0.247 cardiovascular disease + ε_i

In table (3.16), it showed be noted that.

- For (Age and Smoker), in patient related variables, we can see that their p-values are (0.002 < = 0.05, 0.004 < = 0.05), which means that they are significant risk that variables that affect the survival of covid-19 patients, with regression coefficient for age is (β = - 0.164), this means that the risk of death is lower, with chi-Square test value is equal to (9.207). The smoker coefficient value is (β = 0.271), which means that risk of death is higher, with chi-Square test value is equal to (8.258).
- the results show in clinical related variables (Blood Pressure, LDH, Monocyte and D Dimer). The significant values equal (0.007 < = 0.05, 0.042 < = 0.05, 0.032 < = 0.05, 0.001 < = 0.05), it means that these treatments are significant.

Monocyte only variables in clinical related variables affecting in decrease covid-19 disease by coefficient (β = - 0.426), which indicates decrease in the risk of the death for patient, and it is chi-Square test value equal to (4.590).

However, in remaining clinical related variables three variables (Blood Pressure, LDH and D Dimer) are affect increasing in disease by coefficients (β = 0.202, 0.116 and 0.255), and chi-Square test value equal (7.262, 4.122 and 10.198), respectively.

- All variables in chronic disease (Hypertension, diabetes mellitus, chronic lung disease and cardiovascular disease), are significant because their p-values are less than (0.05), by coefficients (β = 0.174, 0.310, 0.337 and 0.247), so all variables which are increase in the risk of the death for patient with the chi-square test values equal to (4.491, 15.946, 17.171 and 5.632).

In addition, chronic lung disease will be one of the significant factors in our study; because it has a greater value in chi-square test column (17.171) with significant value of (0.000 <= 0.05).

- The p-value from the regression table tells us whether or not this regression coefficient is actually statistically significant for (Gender, $SPO_{2,}$ WBC, Lymphocyte, Hb, RBC, PLT, CRP, Ferritin, HR) non-significant p-values were observed, which are not statistically significant at an alpha level of (0.05).

3.7: Selection of best fit parametric model

There are the many ways to compare two or more survival function models, when comparing parametric models, the Akaike Information Criterion (AIC) and the Bayesian information criterion (BIC). Can be used to select the best parametric model. Once the model is identified we will perform survival analysis check that lets us assess the absolute goodness of fit of the identified parametric model. We used MATLAB software to find the value of each the Akaike Information Criteria (AIC) and the Bayesian information criterial (BIC).

Table 3.17 comparing models with AIC and BIC

Models	Number of parameters	Log Likelihood	AIC	BIC
Weibull Model	20	-1001.54	2043.1	2237.4
Log-normal Model	20	-983.792	2007.6	2201.9
Log-Logistic Model	20	-989.126	2018.6	2212.6

The result in Table 3.12 shows the AIC and BIC values which are used to comparing between four models (Weibull Model, Log-normal Model, Log Logistic Model) for selecting the most suitable model to our data of covid-19, the minimum value of AIC and BIC are selected.

By performing the survival analysis for the fitted parametric models log-normal parametric model performed better than other models. Also, we identify that the log-normal model has the lowest AIC and BIC values, it's AIC equals to 2007.6 and BIC equals to 2201.9, in comparison with AIC and BIC in two models (Weibull Model, Log-Logistic Model).

The Log-normal Model with significant variables as follows:

Log T_i = 2.099 – 0.168 Age + 0.213 Smoker – 0.055 SPO_2 – 0.392 Monocyte
+ 0.191 Blood Pressure + 0.253 D Dimer + 0.188 Hypertension
+ 0.302 diabetes mellitus + 0.336 chronic lung disease
+ 0.282 cardiovascular disease

Conclusions

The following conclusions have been reached after studying the data on covid-19 in Erbil city

and as shown by the results from the practical part:

1. Only three patients were found to have normal CRP levels, and none of them had died from the disease. All other death cases were discovered to have high blood inflammation, which indicates that patients with COVID-19 had raised CRP levels.
2. In the study's 44 death cases overall, 33 had abnormally high HR, 36 had raised LDH, and 43 had elevated Ferritin. 213 patients (60.9%) had normal blood pressure, while 25 of 44 cases resulted in deaths.
3. The results for HR and D Dimer indicate that 44 patients who died all had low hemoglobin and high D Dimer. The results indicate that, out of all 44 deceased cases had high D Dimers.
4. The Log-Normal model identified nine prognostic factors that influenced in covid-19 patient's survival, which are (Age, Blood Pressure, Smoker, Monocyte, D Dimer, Hypertension, diabetes mellitus, chronic luge disease, cardiovascular disease).
5. The results of Weibull model Shows that the variables that effecting on the covid-19 disease in our data are (Age, Blood Pressure, Smoker, SPO2, Monocyte, PLT, CRP, Ferritin, D Dimer, Hypertension, diabetes mellitus, chronic luge disease, cardiovascular disease).
6. According to the results of the Log-Logistic model, the most significant variables that have an impact on covid-19 disease are (Age, Blood Pressure, Smoker, Monocyte, LDH, D Dimer, Hypertension, diabetes mellitus, chronic luge disease, cardiovascular disease).
7. The performance of the models in analyzing the covid-19 data in Erbil city was evaluated using the Akaike Information Criterion and the Bayesian Information Criterion. log-normal model appears to be most suitable model according to AIC and BIC compared to other models, and the two other parametric models did not differ significantly from one another.

2. References

1 - LAWLESS , J. F., (2002). Statistical Models and Methods for Lifetime Data. 2nd ed. Canada: A JOHN WILE Y & SONS, INC.

2 - LEE, E. T. & WANG, J. W., (2003). Statistical Methods for Survival Data Analysis. 3rd ed. Canada: John Wiley & Sons, Inc.

3 - BURNHAM, K. P. & ANDERSON, D. R., (2002). Model Selection and Multimodel Inference. 2nd ed. New York: Springer-Verlag New York, Inc.

4 - HOUT, A. V. D., (2017). Multi-State Survival Models for Interval Censored Data. New York: Taylor & Francis.

5 - LAWLESS, J. F., (2003). Statistical Models and Methods for Lifetime Data. 2nd ed. Canada: A JOH N WILE Y & SONS, INC.

6 - LIU, X., (2012). Survival Analysis Models and Applications. 1st ed. United Kingdom: A John Wiley & Sons, Ltd.

7 - MAN, R., (2014). Survival analysis in credit scoring. University of Twente.

8 - NEATH, A. A. & CAVANAUGH, J. E., (2012). The Bayesian information criterion: background, derivation, and applications. Wires Comput Stat 2, Volume 4, pp. 199-203.

9 - PINTO, J. D., (2015). Outlier Detection in Survival Analysis. Tecnico Lisboa.

10 - Qi, J., (2009). Comparison of Proportional Hazards and Accelerated Failure Time Models. Saskatchewan.

11 - RODRIGUEZ, G., (2010). Parametric Survival Models. University, Rapport technique.

12 - WEBULL, W., (1951). A Statistical Distribution Function of Wide Applicability. Journal of Applied Mechanics, American Society of Mechanical Engineers.