

Investigate the Optimum Model for Length of Stay and Mortality Prediction in the Intensive Care Unit

Al Houssainy A. Rady¹ and Sabah Lotfy Mohamed El Sayed^{2*}

¹Professor of Applied Statistics, Institute of Statistical Studies and Research, Cairo University, Egypt

²Assistant Professor of Obstetrics & Gynecology Nursing, Faculty of Nursing, Zagazig University, Egypt

Abstract

Modeling can be a useful tool to find out how the distributions of hospital length of stay and mortality. The aim of the paper is to investigate the optimum predictive statistical model for the length of stay and mortality in the intensive care unit at Zagazig University hospital, Egypt. The paper presents cross-sectional study of 340 patients who were admitted to the intensive care unit of Zagazig University hospital in Zagazig city, Egypt. The data was collected through the medical records of all patients who admitted to intensive care unit of Hospital in 2016. Several statistical analysis and uni-variate regression models were applied using of SPSS and R Package software.

The maximum and minimum lengths of stay belonged to patients were 60 days and 1 day, respectively. The analysis of results showed that both the Log-normal model and binary logistic model are the optimum techniques for length of stay and mortality prediction because they gave the best results for fitting and predicting the length of stay and mortality in the hospital intensive care unit.

Keywords: Length of stay; Mortality; statistical modeling; ICU

Introduction

Intensive care unit (ICU) is one of the critical parts of a hospital which can reduce the rate of mortality and side effects of hospitalization [1]. An increase in the length of stay and the hospital costs are the main problems of admitted patients in intensive care units [2]. The LOS is the key indicators that use to plan for hospital resources in making plans for patients [3]. Since the optimum use of resources plays an important role in enhancing the efficiency of hospitals, reducing the length of stay can lead to increase hospital resources efficiency and effectiveness [4]. To manage hospital beds and patients length of stay and mortality in the intensive care unit need to model hospital data [5].

Modeling is one of the basic tools used in explaining the medical and health phenomena. These tools determine the distribution of specific variables and their relations with other variables by means of regression statistical methods. Modeling the length of stay is a valuable way to know about the status of distribution of LOS [3]. The effecting factors in the length of stay cannot be predicted by using the common statistic methods such as linear regression method because the distribution of LOS data is normal [6].

LOS is a numerical discrete variable; therefore, numerical data distribution should be used to analyze those [7]. Statistical models with different distribution can predict the length of stay, as well as the factors influencing it [8]. The present study aims to determine the optimum statistical predictive model of the hospital length of stay and mortality.

Methods

Study Area and Data collection

This study consisted of 340 patients who are attended in the intensive care unit of Zagazig University Hospital through the period of data collection (from January 2016 to December 2016). The hospital is considered the largest facility in the Al-Sharkia governorate, Egypt. The intensive care unit composed of three wards and each ward about 15 beds. Data applied in this study was obtained as primary data from medical records of all patients.

The collected variables included the following for patient and were

coded: The number of days each patient stay at the hospital (in days) was recorded as count, Patients Age (by years), Gender (0=Male, 1=Female), Education Level (1=Illiterate, 2=Primary educate, 3=Medium educated, 4=Higher), The 1st ICU admission (1=Yes, 0=No), Referral Setting (1=Emergency Department, 2=Medicine Department, 3=Surgery Department, 4=Others), Reasons for ICU admission (1=Medical, 2=Surgical, 3=Others), Number of Nurses (patients per nurse) (1=two or more nurse, 0=one nurse), Number of beds in ICU (1=>10, 0= ≤ 10), Level of Care Provide (1=adequate, 0=inadequate), Blood Transfusion (1=Yes, 0=No), Dialysis (1=Yes, 0=No), Procedures in ICU (1=Yes, 0=No) and Mechanical Ventilation (1=Yes, 0=No), and Mortality (1=Dead, Alive=0).

In this study, the Log-normal distribution was appropriate for predicting the length of stay. Let Y be the random variable representing the length of stay. Have N observations ($N=340$), denoted $(y_i) i \leq N$. Each length of stay is linked to a hospital admission record, providing additional information that can help in understanding the random behaviour of Y . Let denote these complementary observations X . The Lognormal fitting option described by their μ and σ . A variable Y is lognormal distributed if the logarithm of the random variable is normally distributed. The data must be greater than zero.

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

where, $0 \leq x$ or $-\infty < \mu < \infty$ for; $0 < \sigma$

Logistic regression is a class of generalized linear models and a

*Corresponding author: Lotfy Mohamed El Sayed, Assistant Professor, Zagazig University, Egypt, Tel: 01128182813; E-mail: sl_200878@yahoo.com

Received March 12, 2018; Accepted June 04, 2018; Published June 14, 2018

Citation: Rady AHA, El Sayed SLM (2018) Investigate the Optimum Model for Length of Stay and Mortality Prediction in the Intensive Care Unit. J Perioper Crit Intensive Care Nurs 4: 143. doi:10.4172/2471-9870.10000143

Copyright: © 2018 Rady AHA, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

type of regression model used when the outcome variable is qualitative and has binary indicators. In the current study, the outcome event is mortality. For multiple regressions a model of the following form can be used to predict the value of a response variable y using the values of a number of explanatory variables:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_q X_q \quad (2)$$

Where: y is the response variable (patients mortality) with a status 1 if the patient died and 0 if otherwise. $\beta_0 = \text{Constant/ intercept}$, $\beta_1 \rightarrow \beta_q = \text{coefficients for } q \text{ explanatory variables } X_1 \rightarrow X_q$. The expected value of y is the probability that $y=1$ which makes the range of y to be limited between 0 through 1.

The regression process finds the coefficients which minimise the squared differences between the observed and expected values of y (the residuals). As the outcome of logistic regression is binary, y needs to be transformed so that the regression process can be used. The logit transformation gives the following:

The logit function is defined as the natural logarithm (ln) of the odds of death [9]. That is,

$$\text{Logit}(p) = \ln \left[\frac{p}{1-p} \right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_q X_q \quad (3)$$

Therefore y is expressed as logit (p) where p is the probability of death e.g. patient dies,

$$\frac{p}{1-p} = \text{odds ratio}$$

If probabilities of the event of interest happening for patients are needed, the logistic regression equation can be written as:

$$\text{probability of death} = p = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_q X_q)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_q X_q)} \quad (4)$$

where $0 < p < 1$

The process was carried out according to the following steps:

Step (1): Tests of Normality

The data were tested for normality distribution using SPSS package. Initial analyses indicated that log-normal distribution was suitable for predicting the LOS.

Step (2): Fit of distributions

The data was fitted using R package, the Cramer-von Mises test confirmed that the log normal and log logistic models have the accuracy to predict patient LOS. Additionally, the binary logistic model was fit for predicting the mortality.

Step (3): Model Evaluation

The final step is to test the best of these models and compare them in order to choose the optimum predictive models. To select the optimum model use the following methods:

- (1) Akaike information criterion (AIC). It is a penalized-likelihood criterion which is used for choosing best predictor subsets in regression. AIC is well known method for the identification of an optimum model in a class of competing models. In the present study, the Akaike information criterion (AIC) was employed to compare the performance of the models. In statistical models, the AIC denotes which model shows a better fit to the data [10]. A lower AIC value denotes a better fit. The Formula for the statistical estimation of the AIC was as follows:

$$\text{AIC} = -2 \times \log(\text{likelihood}) + 2(P+K)$$

Where P is the number of parameters, and K (constant factor) equals 2.

AIC for a model is usually illustrated as $[-2\log l + kp]$, where l is the likelihood function, p is the number of parameters in the model, and $k=2$

Where P is the number of parameters, and K (constant factor) equals 2.

AIC for a model is usually illustrated as $[-2\log l + kp]$, where l is the likelihood function, p is the number of parameters in the model, and $k=2$

- (2) Mean Square Error (MSE) to measure the average magnitude of the errors in the models.

Taking all these measures into account, the optimum model among all models specified for the available data at hand is the one with the lowest (AIC), and lowest MES. The implementation of all model fits is done through R package.

Results and analysis

In this section, the results of test trials are presented and its analysis to get the optimum predictive model. A prediction model is made according to the analytic results, which shows a good consistency with the original observed data.

Part (I): Tests of Normality

Visual inspection of the distribution may be used for assessing normality. The normality tests are supplementary to the graphical assessment of normality. The frequency distribution (histogram), Kolmogorov-Smirnov (K-S) test, Shapiro-Wilk test, P-P plot (probability-probability plot), and Q-Q plot (quantile-quantile plot) are used for checking normality visually.

Table 1 shows the descriptive statistics of length of stay in intensive care unit. Average length of stay in ICU was (7.20±9.19) days. The maximum and minimum lengths of stay belonged to patients were 60 days and 1 day, respectively. Additionally, the high values of skewness (3.123), kurtosis (12.510) and the presence of outliers suggest that the distribution do not follow a normal distribution. As known, significance value must be greater than 0.05 in a normal distribution. Significance value, Sig.=0.00. So, the data is not normally distributed (Table 2).

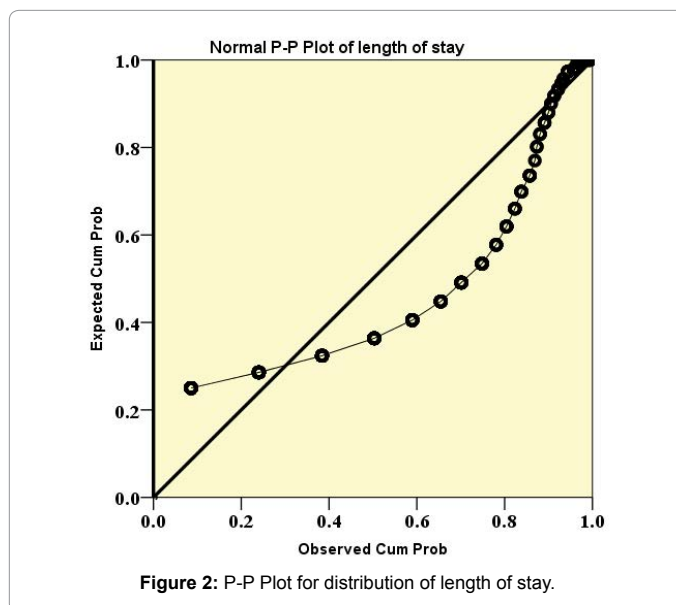
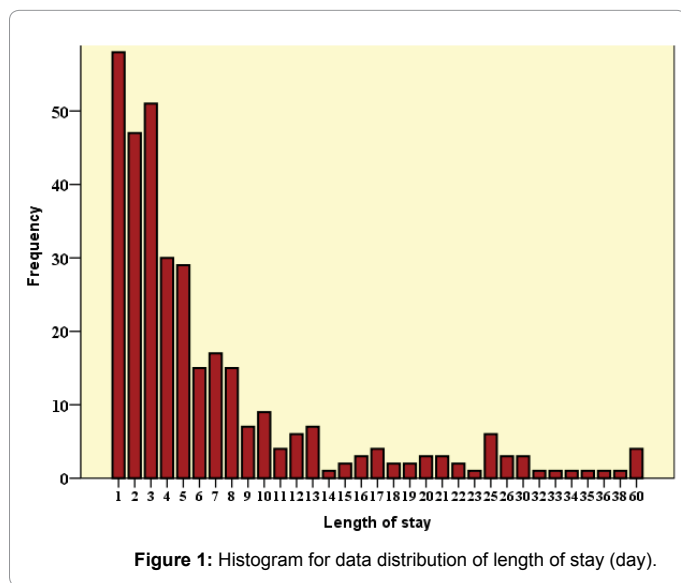
Figure 1 shows the histogram of the patient's length of stay in the intensive care unit. The distributions contains outliers (which are truncated at LOS=60 days). The distribution of the patients length of stay was from 1-60 days. Most patients 58 (17.1%) stayed in intensive care unit for 1 day. Additionally, this diagram is right skewed, so the

No (patients)	340
Mean of Length of stay (day)	7.20
Median of Length of stay (day)	4.00
Std. Deviation of Length of stay (day)	9.197
Variance of Length of stay (day)	84.581
Skewness of Length of stay (day)	3.123
Kurtosis of Length of stay (day)	12.510
Minimum of Length of stay (day)	1
Maximum of Length of stay (day)	60

Table 1: Findings related to descriptive statistics for the length of stay in ICU.

Distribution	Cramer-von Mises test	Kolmogorov-Smirnov test	Anderson-Darling test
1. Exponential	P-value = 0.006657	P-value = 1.953e-05	P-value = 0.0002297
2. Uniform	P-value < 2.2e-16	P-value < 2.2e-16	P-value = 1.765e-06
3. Logistic	P-value = 2.432e-05	P-value = 0.01421	-
4. Log normal	P-value = 0.07909	-	P-value = 0.01833
5. Log-Logistic	P-value = 0.08276	P-value = 0.01421	P-value = 0.019
6. Cauchy	P-value = 8.649e-05	P-value = 1.993e-07	P-value = 1.765e-06
7. Poisson	P-value = 6.055e-09	P-value < 2.2e-16	P-value = 1.765e-06
8. Geometric	P-value = 4.287e-06	p-value = 6.661e-16	P-value = 1.765e-06
9. Negative binomial	P-value = 6.726e-05	-	P-value = 1.793e-06
10. Gumbel	P-value = 0.0001134	P-value = 2.611e-10	-
11. Pareto	P-value = 0.006657	P-value = 1.971e-05	P-value = 0.0004011
12. Gamma	P-value = 0.04299	-	P-value = 0.003761
13. Weibull	P-value = 0.0363	P-value = 0.008263	P-value = 0.002922

Table 2: Distribution of the length of stay in ICU according to the fit distributions.



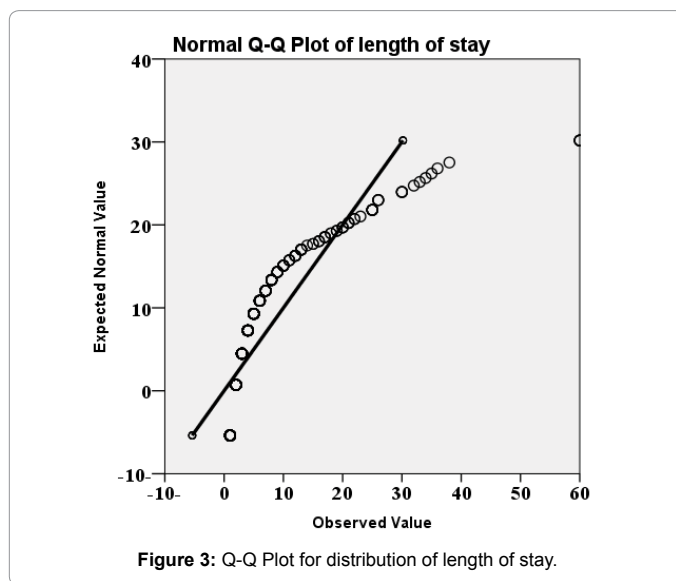
data is not normally distributed. Most of the classical statistical tests rely on the assumption of normality of the outcome variable. By visual inspection, the histograms look far from normal. The P-P Plot & Q-Q plot for distribution of length of staying. Figures 2 & 3 confirm the lack of normality (i.e. in a P-P Plot & Q-Q plot generated from normally distributed data it is expected that observed values would follow the straight line).

Part (II): Fit of distributions

In assessing whether a given distribution was suited to a data-set, the following tests and appropriate models of fit were used:

Table 3, shows the distribution of the length of stay in ICU according to the fit distributions. According to the Cramer-von Mises test, it was shown that the lognormal and log-logistic distributions were suited among the chosen distributions (P=0.079 & P=0.083, respectively).

The Q-Q plot histogram for the length of stay in the intensive care unit shows that the distribution is near from a log-normal distribution. As regard to the P-P Plot of the length of stay distribution in the intensive care unit (Figure 4). Figure 5 illustrates that, distribution is closed to the lognormal distribution.



Models	AIC	MSE	X ²	P value
Lognormal	1953.218348	93.78138997	32.78449081	0.018
Log logistic	1960.68635	95.2044254	39.2723833	0.0026
Exponential	2019.473287	81.94018254	41.17953133	0.0014
Weibull	2020.561237	82.22775339	41.36721612	0.0014
Gaussian	2483.071095	82.1256816	29.62736803	0.041
Poisson	3468.5	-	-	0.000

AIC: Akaike information criterion
 MSE: Mean Square Error

Table 3: Comparison of the regression models according to the evaluation criteria of the best model of length of stay in intensive care unit.

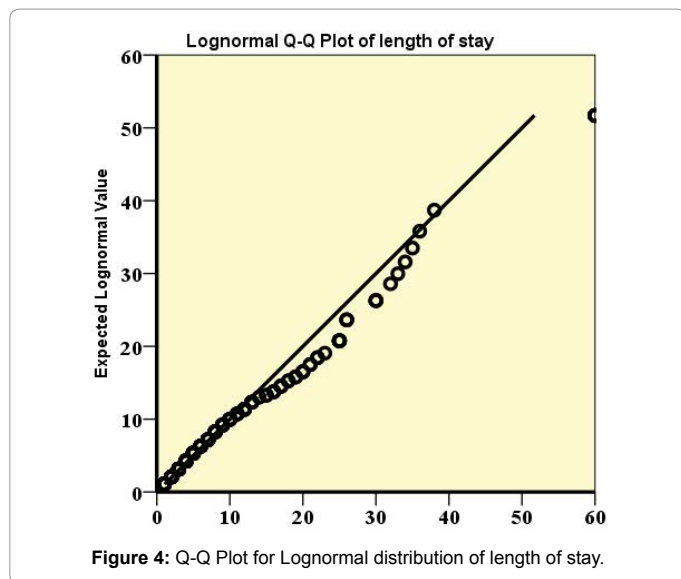


Figure 4: Q-Q Plot for Lognormal distribution of length of stay.

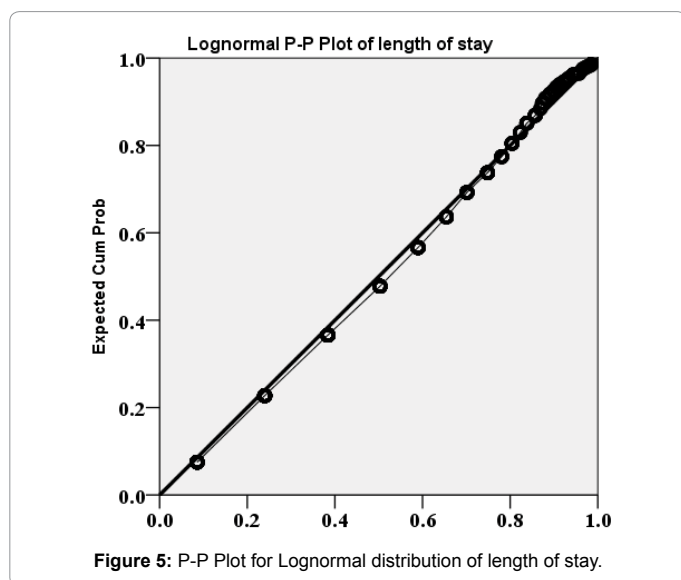


Figure 5: P-P Plot for Lognormal distribution of length of stay.

Part (III): Fit of models

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the difference between observed values and the values expected under the model in question. These measures can be used to test whether the dependent variable follows a specified distribution. With regard to the Akaike information criterion which is used to evaluate the best model.

Patient Status at Discharge	Internal Value	Frequency	Percent	Valid Percent
Alive	0	176	51.8	51.8
Dead	1	164	48.2	48.2
Total	-	340	100.0	100.0

Table 4: The frequency distribution of dependent (patient's status at discharge) variable according to their encoding and descriptive statistics.

	Beta (β) coefficient	Exp(B)	p-value
Intercept (Constant)	.964	2.622	0.564
Age	.025	1.026	0.010
Gender (female)	-.747-	.474	0.011
Education of patient	-	-	0.643
Primary	.424	1.528	0.424
Secondary	.572	1.771	0.286
Higher	.623	1.864	0.221
First ICU admission (yes)	-.317-	.728	0.353
Referral Setting	-	-	0.166
Emergency department	.524	1.689	0.436
Medicine department	.206	1.229	0.781
Surgical department	-.632-	.531	0.463
Reasons of Admission	-	-	0.314
Medical	-.232-	.793	0.856
Surgical	.318	1.374	0.806
Nurses No (patients per nurse) (≥2)	-.307-	.735	0.334
Number of beds in ICU (> 10)	.170	1.185	0.616
Care Provide level (adequate)	.301	1.351	0.436
Blood Transfusion (yes)	-.176-	.838	0.550
Dialysis(yes)	-.686-	.503	0.071
Procedure(yes)	-.088-	.915	0.790
Ventilation(yes)	-2.943-	.053	0.000

Table 5: Parameter estimates of final logistic regression model for ICU mortality.

Table 4 shows that the lognormal model has the lowest value (1953.2), while the Poisson model had the highest value (3468.5).

For this study, data from the intensive care unit was used to identify significant variables affecting mortality by using Binary Logistic Regression Analysis. The sample consisted of 340 patients. The dependent variable was dichotomous, i.e. whether the patient was alive or not at discharge.

Regarding SPSS output for logistic regression, Table 5 shows that the coding of the binary variable is coded as 0 and 1. It noticed from Table 5, less than half (48.2%) of the patients died. In Table 6, the probability of the chi-square statistic was 135.7; the model chi-square has 18 degrees of freedom, and a probability of p<0.001. Thus, the indication is that the model has a poor fit. From the table above, the model summary for logistic regression reveals that two values of the R2 values were between 32.9% and 43.9% of the variation in mortality is explained by the model. Cox and Snell's R2 indicates that, 32.9% of the

N	Mortality = Alive		Mortality = Dead		Total
	Observed	Expected	Observed	Expected	
1	32	30.920	2	3.080	34
2	23	29.109	11	4.891	34
3	31	27.065	3	6.935	34
4	29	24.480	5	9.520	34
5	18	21.255	16	12.745	34
6	15	17.386	19	16.614	34
7	15	11.561	19	22.439	34
8	8	7.345	26	26.655	34
9	1	4.437	33	29.563	34
10	4	2.441	30	31.559	34

Table 6: Hosmer and Lemeshow and goodness-of-fit test for the logistic regression model.

variation in the dependent variable is explained by the logistic model. The Nagelkerke's R² was 0.439, indicating a weak relationship of 43.9% between the predictors and the prediction

The final formula equation for model is:

$$\text{Log [Mortality]} = 0.964 + 0.025 [\text{age}] - 0.747 [\text{Gender (female = 1)}] + 0.424 [\text{Education of patient (primary)}] + 0.572 [\text{Education of patient (secondary)}] + 0.623 [\text{Education of patient (higher)}] - 0.317 [\text{1st admission}] + 0.524 [\text{Referral Setting (Emergency department)}] + 0.206 [\text{Reasons of Admission (Medicine department)}] - 0.632 [\text{Reasons of Admission (Surgical department)}] - 0.232 [\text{Reasons of Admission (Medical)}] + 0.318 [\text{Reasons of Admission (Surgical)}] - 0.307 [\text{Nurses No } (\geq 2)] + 0.170 [\text{Number of beds in ICU } (>10)] + 0.301 [\text{Care Provide level (adequate)}] - 0.176 [\text{Blood Transfusion (yes)}] - 0.686 [\text{Dialysis(yes)}] - 0.088 [\text{Procedure (yes)}] - 2.943 [\text{Ventilation (yes)}].$$

Table 5 revealed that, data from the intensive care unit was used to identify significant factors affecting mortality by using binary logistic regression model. The dependent variable was dichotomous, i.e. whether the patient was alive or not at discharge. As a measure of calibration from the model, the Hosmer and Lemeshow goodness-of-fit statistics reveals $\chi^2=22.873$ ($P=0.004$), which indicating poor agreement between observed and expected ICU mortality (Table 6).

Conclusion

This study is a contribution to the development of the statistical analysis of LOS distributions in the intensive care unit. Further, there are outliers with the long right tail were observed in LOS distributions. This study assumed that there is a variety of distributions patterns for LOS but, it is far from the normal distribution. This finding was comparable with Malehi et al. [8] who reported that LOS data are right skewness to model skew data, use log-normal regression to model skewed data. R analysis confirmed that the LOS was followed the lognormal model and have the accuracy to predict patient LOS.

The current study considered the modeling evaluation tools and the associated procedures for model fitting. These findings showed that the

lognormal model had the lowest value of AIC, consequently log-normal model is a fit model for predicting the length of stay in the intensive care unit. The study results are contradictory with Gharacheh et al. [11] mentioned that goodness of fit for gamma model was more suitable and powerful than the log-normal model to predict the factors affecting the patients length of stay in the intensive care units of the hospital.

The results of the study showed that logistic regression model had the ability in predicting the mortality of ICU patients. These results are accordance to Bera & Nayak [12] reported that logistic regression can be used in prediction of hospital mortality of ICU patients. Moreover, Austin et al. [13] stated that logistic regression predicted in-hospital mortality in patients hospitalized with heart failure more accurately. So, the results of this study demonstrated that in comparison to statistic models, log-normal model and logistic model had the ability in predicting the length of stay and mortality, respectively.

References

1. Marzban S, Maleki MR, Nasiri pour AA, Jahangiri K (2013) Assessment of patient safety management system in ICU. *J Qazvin Univ Med Sci* 17: 47-55.
2. Besharati R, Sadeghian A, Mamori GA, Lashkardoust H, Gholami S (2013) Sources of bacteria causing nosocomial infections At NICU of Ghaem Hospital in Mashhad, Iran. *JNKUMS* 5: 25-29.
3. Abolghasem P, Sedigheh S, Somayeh SD, Mahdiyeh SN, Hamed T, et al. (2015) Factors influencing the length of stay in infectious ward of Razi hospital in Ahvaz: Iran. *Health Inf Manage* 11: 779-788.
4. Nasiripour A, Riahi L, Gholamipour A (2010) Presence of fulltime medicos in hospital and its effect on length of stay of Yazd hospital. *Medical Military Organization of Islamic Republic of Iran Journal* 28: 169-175.
5. Soltani MH, Sahaf R, Shahbolaghi FM, Ghaffari S, Khosravi A, et al. (2015) Factors influencing the length of stay in infectious ward of Razi hospital in Ahvaz: Iran. *Health Inf Manage* 11: 779-788.
6. Mihaylova B, Briggs A, O'hagan A, Thompson SG (2011) Review of statistical methods for analyzing healthcare resources and costs. *Health Econ* 20: 897-916.
7. Golaghaie F, Sarmadian H, Raffie M, Nejat N (2008) study on waiting time and length of stay of attendants to emergency department of Vali-e-Asr Hospital, Arak-Iran. *J Arak Uni Med Sci* 11: 74-83.
8. Malehi AS, Pourmotehari F, Angali KA (2015) Statistical models for the analysis of skewed healthcare cost data: a simulation study. *Health Economics Review* 5: 1-16.
9. Kirkwood BR, Sterne JAC (2003) *Essential medical statistics*. Malden: Blackwell, 2. *Int J Epidemiol* 6: 1418-1419.
10. Klein JP, Moeschberger ML (2003) Semiparametric proportional hazards regression with fixed covariates. *Survival analysis* 243-293.
11. Gharacheh L, Torabipour A, Khiavi FF, Malehi AS, Haddadzadeh M (2017) Comparison of Statistical Models of Predict the Factors Affecting the Length of Stay (LOS) in the Intensive Care Unit (ICU) of a Teaching Hospital. *Mater Sociomed* 29: 88-91.
12. Bera D, Nayak MM (2012) Mortality risk assessment for ICU patients using logistic regression. *Computing in Cardiology (CinC)* 39: 493-496.
13. Austin PC, Tu JV, Lee DS (2010) Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *J Clin Epidemiol* 63: 1145-1155.