




A comparison of parametric and semi-parametric survival models with artificial neural networks

Reza Mokarram, Mahdi Emadi, Arezou Habibi Rad, and Mahdi Jabbari Nooghabi 

Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran

ABSTRACT

Survival models are used to examine data in the event of an occurrence. These are discussed in various types including parametric, non-parametric and semi-parametric models. Parametric models require a clear distribution of survival time, and semi-parametric models assume proportional hazards. Among these models, the non-parametric model of artificial neural network has the fewest assumptions and can be often replaced by other models. Given the importance of distribution Weibull survival models in this study of simulation shape parameter of the Weibull distribution have been assumed as 1, 2 and 3, and also the average rate at levels of 0%–75% have been censored. The values predicted by the neural network forecasting model with parametric survival and Cox regression models were compared. This comparison considering levels of complexity due to the hazard model using the ROC curve and the corresponding tests have been carried out.

ARTICLE HISTORY

Received 14 March 2016
Accepted 25 January 2017

KEYWORDS

Artificial Neural Networks;
Cox model; Parametric
model; Proportional Hazard



MATHEMATICS SUBJECT CLASSIFICATION

62N01

1. Introduction

Many different semi-parametric, non-parametric, and parametric regression methods are evaluated to determine a relation between a related variable and a set of covariates. The choice of a proper method for modeling depends on the methodology of the study and the nature of the results and explanatory variables.

A usual research context in medical and industrial studies is a survival analysis in order to determine if a set of covariates is correlated with the survival time. Survival analysis are methods for analyzing longitudinal data on the occurrence of events. The events in medical area may include death, injury, onset of illness, or recovery from illness. Two basic features of survival data are censoring and lack of normal assumption. Therefore, these are reasons why multiple regression techniques cannot be applied to this kind of data (Biglarian and Bakhshi, 2013). Subjects are said to be censored if they are lost to follow up or drop out of the study, or if the study ends before they die or exits from the area of interest. The generated data used in this study contain right censored data, which is the most common form of censoring. It occurs if the event is not observed before the prespecified study term unit or competing event that causes the interruption to follow from the individual experiments. Considering the survival time distribution, we can use a parametric model. If we have a proportional hazard (PH) assumption, we can use the Cox model which is known as semiparametric model. Artificial

CONTACT Mahdi Emadi  Emadi@math.um.ac.ir  Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

Neural Network (ANN) is one of the most accurate and widely used forecasting models. ANNs are flexible computing frameworks and universal approximations. They can be applied with a high degree of accuracy to a wide range of statistical problems such as a survival model. In particular, a state that violates proportional hazard ANNs can be used to determine predictions of survival data (Ciampi and Lechevallier, 1997; Hastie et al., 2012). In this article, simulated data, covariates and survival time are taken from three models of hazard with different levels of complexity and many rate of censoring were used to predict the outcome variable using ANN, parametric regression, and Cox's regression models. Then, the results of the prediction of ANN are compared with parametric and semiparametric models. This comparison by using the area under ROC curve and the ROC test have been carried out.

2. Methods

2.1. Parametric survival model

In a parametric survival model, it is supposed that the survival time has known distribution. Examples of distributions used for survival time are the exponential, Weibull, log-normal, and log-logistic. The Weibull distribution is the most important and flexible model for survival data because it provides a sufficient fit in many situations, even when the data do not follow an exact Weibull distribution (Carroll, 2003).

Suppose that the Weibull distribution including shape parameter α and scale parameter γ indicates $T \sim W(\gamma, \alpha)$, we know $T^\alpha \sim E(\gamma)$, Kleinbaum and Klein (2012). The survival function is

$$S(t) = P(T > t) = \exp\{-\gamma^\alpha\} \quad \gamma > 0, \alpha > 0.$$

The hazard function is

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \alpha \gamma^\alpha t^{\alpha-1} \end{aligned}$$

The Weibull model provides a wide range of monotonic hazard rates. The hazard rate reduces when shape parameter α is less than 1. The hazard is constant when $\alpha = 1$ (exponential distribution) and it increases for $\alpha > 1$.

The Weibull model can be fitted to the survival data (t_i, δ_i) , if the observation is censored then $\delta_i = 0$ and $\delta_i = 1$ when it is complete.

The parameters of the model are estimated by using the maximum likelihood method.

2.2. Cox regression model (semi-parametric survival model)

The survival time of an event is a continuous, non-negative random variable with survival function $S(t)$ and density function $f(t)$. The related hazard function $h(t)$ shows the probability density of an event occurring around time t , provided that it has not occurred before time t .

Cox regression, which is also called the proportional hazards regression model (PH), is shown as

$$h_x(t, \beta) = h_0(t) \exp(\beta' X) \quad \beta' = (\beta_1, \beta_2, \dots, \beta_p) \quad (1)$$

So that β' is the regression coefficient vector, $h_0(t)$ is the baseline hazard which is related to time and an unspecified function (this is the property that makes the Cox model a semiparametric model, Lee and Wang 2003). Also, X is a vector of x_i elements that are predictors.

This model has properties as follows (Lee and Wang, 2003; Kleinbaum and Klein, 2012):

- The covariates, x_1, x_2, \dots, x_p are assumed to act additively on $\log_x(t, \beta)$.
- $\log_x(t, \beta)$ has a linear relation with β_i .
- The hazard ratio (HR) for a subject with a set of predictors x_i compared with a set of predictors x_j is

$$\begin{aligned} \text{HR} &= \frac{h_{x_i}(t, \beta)}{h_{x_j}(t, \beta)} \\ &= \frac{\exp(\beta'x_i)}{\exp(\beta'x_j)} \\ &= \exp\{\beta'(x_i - x_j)\} \end{aligned}$$

which is free of time event.

Given the covariates \mathbf{X} , the survival function for (1) is

$$S_x(t, \beta) = S_0(t)e^{\beta'X}$$

where $S_0(t) = e^{-\int_0^t h_0(v)dv}$ denotes the baseline survival function.

The Likelihood function is given by Lee and Wang (2003)

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n h_{x_i}(t_i, \beta)^{\delta_i} S_{x_i}(t_i, \beta) \\ &= \prod_{i=1}^n (h_0(t_i) \exp(\beta'x_i))^{\delta_i} S_0(t_i)^{\exp(\beta'x_i)} \end{aligned}$$

Parameters estimate in the Cox(PH) model is attained by maximizing the partial likelihood with the Newton-Raphson method. The partial likelihood is given by

$$\prod_{i=1}^d \frac{\exp(\beta'x_i)}{\sum_{j \in R(t_i)} \exp(\beta'x_j)}, \quad d = \sum_{i=1}^n \delta_i \quad (2)$$

where $R(t_i)$ is the set of risk at time t_i . In a particular case (e.g., high-dimensional), this cannot be used to estimate β and another method need to be used (Kleinbaum and Klein, 2012).

2.3. Neural network model (non-parametric survival model)

ANNs, as an interconnected group of artificial neurons, consist of several layers. Besides, each layer has a weight indicating the amount of the effect of neurons on one another.

Typically, an ANN model has three layers called the input, hidden, and output layers. The input layer contains the predictors; the hidden layer contains unobservable nodes and is used to apply a nonlinear transformation into the linear combination of the input layer. The number of hidden nodes depends on factors such as the predictors and model complexity. The output layer contains the outcome which is the functions of the hidden units (Ciampi and Lechevallier, 1997; Haykin, 1994).

In fact, the ANN model is a nonlinear model including a large number of parameters in comparison with the corresponding statistical model. Various methods to learn are postulated

in ANN. The most usual method is minimizing the sum of squares error by back-propagation learning algorithm (Ripley et al., 2004; Ripley and Ripley, 2001).

In this study, we take the activation function (φ_h) to be sigmoid function in hidden and output layers (φ_o). A Multiple Layer Perceptron (MLP) is given by

$$t_i = \varphi_o \left[\omega_0 + \sum_{n=1}^{N-1} \omega_n \varphi_h(\check{X}_i \alpha_n) \right] + \varepsilon_i$$

where X_i is i th row of the input data matrix, ω_i is the weights of the hidden to the output units, and α_i is the weights of the input to the hidden units. N is the number of units of the hidden layer and the sigmoid activation function is

$$\varphi(\tau) = \frac{1}{1 + \exp(-a\tau)}$$

where a is the slope parameter of the sigmoid function. We take the slope parameter as equaling unit and so

$$t_i = \left[1 + \exp \left(-\omega_0 - \sum_{n=1}^{N-1} \omega_n [1 + \exp(-\check{X}_i \alpha_n)]^{-1} \right) \right]^{-1} + \varepsilon_i \tag{3}$$

Also, we can write

$$t_i = \varphi(X_i, \alpha, \omega) + \varepsilon_i$$

where α, ω are unknown parameter vectors, X_i is a vector of predictors for the i th case, and ε_i is residuals. The weights (α, ω) are determined by minimizing some error function, the most common of which is distance between the target value of the output variables and those given by (3).

The partial derivatives of the error function regarding the weights can usually be calculated repeatedly from output to input in the network (a process known as back-propagation) then we use quasi-Newton method to find a local minimum. MLP allows fitting of very general nonlinear function relationships between inputs and outputs. The results show a sufficient number of nodes in the hidden layer can be estimated to be an arbitrary relation function.

The MLP process is a major concern of over-fitting. Commonly to control it, a penalty term is added to criteria optimization. Therefore, least-squares criterion is given by

$$\sum_{i=1} (t_i - \hat{t}_i)^2 + \text{Pen}_\lambda(\alpha, \omega) = \sum_{i=1} (t_i - \varphi(X_i, \alpha, \omega))^2 + \text{Pen}_\lambda(\alpha, \omega)$$

where the penalty term is Hastie et al. (2012)

$$\text{Pen}_\lambda(\alpha, \omega) = \lambda \left(\sum \alpha_{ij}^2 + \sum \omega_{jk}^2 \right) \quad i = 1, \dots, 5 \quad j = 1, \dots, 5 \quad k = 1, 2$$

where i, j, k are the number of input, hidden, and output units, respectively. The penalty weight λ (weight decay) controls the amount of over-fitting (a better value of λ is between 0.001 and 0.1)

Unfortunately, no reliable law to determine the number of nodes in the hidden layer of a neural network exists. In most cases, the number of nodes 5–20 in the hidden layer network can provide good results. If the number of nodes in the hidden layer is considered low, neural network model cannot describe the behavior of nonlinear data, where the number of nodes in the hidden layer is increased, the training process slows down and most of the weight is concentrated around zero. Usually by increasing the number of inputs and training cases the

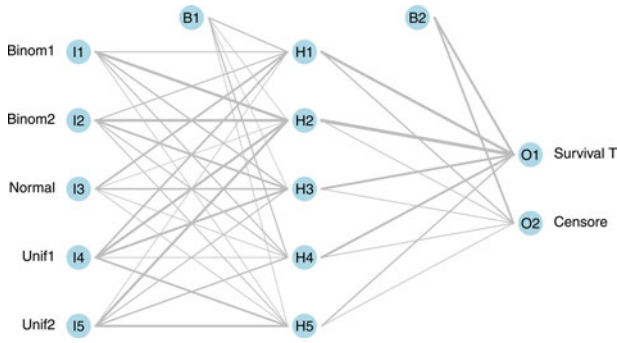


Figure 1. Neural network for nonlinear hazard model.

number of nodes in the hidden layer increases. The number of nodes in the hidden layer of the neural network is determined by previous experiences in any academic area (Hastie et al., 2012). A criterion for selecting the number of nodes in the hidden layer is using of root-mean-square error (RMSE). The RMSE of a model prediction with respect to the estimated variable X_m is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (X_{o,i} - X_{m,i})^2}{n}}$$

where $X_{o,i}$ is observed values and $X_{m,i}$ is modeled values at time i . In this article, to determine the number of nodes in the hidden layer we used the average values RMSE in deferent rate of censorship for each of hazard function and select the number of nodes that this quantity was the minimum. According to this, the number of nodes in the hidden layer for linear and non-linear hazard models 3 and 5 in order be considered (Figure 1).

2.4. Simulation

To compare the accuracy of the predictions in the ANN model with Cox regression and parametric survival regression model, three different simulation schemes, based on Mont-Carlo simulation, were applied. Simulated data were considered with five covariates as inputs (two binomials, one normal, and two uniforms). We assume the hazard model as follows (Xiang et al., 2000)

$$h(t, x) = h(t) \exp \left\{ \sum_{i=1}^p \gamma_i x_i + \sum_{i \neq j} \gamma_{ij} x_i x_j + \sum_{i \neq j \neq k} \gamma_{ijk} x_i x_j x_k + \sum_{i \neq j \neq k \neq l} \gamma_{ijkl} x_i x_j x_k x_l \right\}$$

where $h(t)$ is baseline hazard from the Weibull distribution.

The following three schemes were considered which include the main effects without or with any interaction as simple and complicate model, respectively.

(1)

$$\gamma_1 = \gamma_4 = 0.25, \quad \gamma_2 = \gamma_3 = \gamma_5 = 0.5 \tag{H1}$$

(2)

$$\gamma_1 = \gamma_4 = 0.25, \quad \gamma_2 = \gamma_3 = \gamma_{14} = \gamma_{25} = \gamma_{34} = \gamma_{35} = 0.5 \tag{H2}$$

(3)

$$\gamma_1 = \gamma_4 = \gamma_{12345} = 0.25, \quad \gamma_2 = \gamma_3 = \gamma_{14} = \gamma_{25} = \gamma_{134} = \gamma_{235} = 0.5 \quad (H3)$$

For each scheme, 1000 independent random observations were generated and then survival times were simulated using the Weibull model with shape parameter $\alpha = 1, 2, 3$ and based on the relation between hazard and survival time (Carroll, 2003).

Then, the survival times were transformed as right censored. In case, the generated time was greater than the quantile of Weibull, it was considered as censored observation. The average rates of censoring reviewed are equal to 0%, 15%, 30%, 45%, 60%, and 75%. In addition, we assume 75% of each sample is for learning model and the remaining for testing. This process was repeated 200 times and average results was considered.

In the next step, to compare ANN predictions with Cox regression and survival regression prediction, we used Receive Operating Characteristics (ROC) curve and concordance index. This index is the area under the curve of ROC (AUC) and shows that the proportion of the cases is classified correctly (Fawcett, 2006). Likewise, we used ROC test to compare ANN models with parametric and semiparametric survival models. It is based on the Delong method that tests AUC for two correlated or uncorrelated ROC curves (Robin et al., 2011).

Simulations, fitting models, and comparison process results are implemented by R 3.3.2 software (Fox and Weisberg, 2011).

3. Results and discussion

In this article, two comparisons for modeling of survival simulated data with different hazard ratio and different censoring (ANN model with Cox and parametric models) were studied. A Monte-Carlo simulation from the Weibull distribution was performed to compare the predictive accuracy of ANN models with the other models.

In this study, three different hazard models were considered. They include main effects model, and a model with maximum of two interactions as well as one with a maximum of four interaction terms. The rate of censorship in all of the models was considered from 0% up to 75%.

Table 1. Results of simulation for Weibull model with Shape Parameter = 1.

Hazard Model	Censored Rate	ANN - AUC	COX - AUC	SURVReg - AUC	P-value ANN-COX	P-value ANN-SURVReg
H1	0%	0.702	0.603	0.719	0.029	0.339
	15%	0.703	0.601	0.719	0.044	0.346
	30%	0.701	0.603	0.718	0.035	0.298
	45%	0.702	0.601	0.719	0.021	0.312
	60%	0.702	0.602	0.718	0.053	0.311
	75%	0.701	0.600	0.718	0.066	0.336
H2	0%	0.819	0.710	0.824	0.005	0.449
	15%	0.819	0.714	0.824	0.008	0.489
	30%	0.817	0.717	0.822	0.019	0.466
	45%	0.816	0.718	0.822	0.021	0.494
	60%	0.818	0.719	0.821	0.042	0.495
	75%	0.809	0.719	0.821	0.007	0.415
H3	0%	0.771	0.663	0.779	0.018	0.456
	15%	0.772	0.667	0.778	0.016	0.453
	30%	0.771	0.668	0.778	0.023	0.446
	45%	0.770	0.672	0.777	0.017	0.452
	60%	0.772	0.679	0.777	0.026	0.398
	75%	0.773	0.682	0.776	0.033	0.403

Table 2. Results of simulation for Weibull model with Shape Parameter = 2.

Hazard Model	Censored Rate	ANN - AUC	COX - AUC	SURVReg - AUC	P-value ANN-COX	P-value ANN-SURVReg
H1	0%	0.836	0.805	0.845	0.167	0.404
	15%	0.836	0.803	0.844	0.192	0.410
	30%	0.834	0.804	0.844	0.198	0.422
	45%	0.834	0.803	0.842	0.241	0.498
	60%	0.833	0.801	0.841	0.212	0.360
	75%	0.835	0.801	0.840	0.191	0.322
H2	0%	0.921	0.896	0.921	0.142	0.493
	15%	0.921	0.898	0.921	0.138	0.556
	30%	0.920	0.900	0.920	0.128	0.485
	45%	0.919	0.900	0.919	0.126	0.489
	60%	0.918	0.899	0.918	0.110	0.429
	75%	0.917	0.897	0.918	0.120	0.360
H3	0%	0.891	0.858	0.888	0.102	0.443
	15%	0.889	0.860	0.889	0.118	0.439
	30%	0.893	0.861	0.890	0.109	0.478
	45%	0.893	0.864	0.887	0.115	0.479
	60%	0.881	0.863	0.885	0.112	0.442
	75%	0.890	0.864	0.889	0.134	0.358

Concordance index from the ROC curve was computed and comparison process was performed by the ROC test.

Calculation results are presented in Tables 1–3 with respect to shape parameter $\alpha = 1, 2, 3$

Key results of the study are as follows:

1. In the Weibull distribution with shape parameter $\alpha = 1$ (exponential distribution), the ANN model predicts better than the Cox model and this difference is significant in all three types of hazard functions.
2. In the Weibull distribution, no significant difference was found between predictions based on the ANN model and predictions based on the parametric survival model.
3. Different levels of censorship have no significant effect on the accuracy of prediction of three models studied.

Table 3. Results of simulation for Weibull model with Shape Parameter = 3.

Hazard Model	Censored Rate	ANN - AUC	COX - AUC	SURVReg - AUC	P-value ANN-COX	P-value ANN-SURVReg
H1	0%	0.901	0.893	0.903	0.293	0.396
	15%	0.900	0.892	0.903	0.367	0.367
	30%	0.901	0.891	0.902	0.392	0.353
	45%	0.902	0.890	0.902	0.352	0.351
	60%	0.903	0.891	0.901	0.399	0.350
	75%	0.901	0.892	0.901	0.311	0.396
H2	0%	0.963	0.954	0.961	0.245	0.439
	15%	0.961	0.955	0.964	0.279	0.456
	30%	0.961	0.953	0.962	0.266	0.460
	45%	0.966	0.950	0.962	0.236	0.393
	60%	0.965	0.947	0.957	0.222	0.319
	75%	0.969	0.943	0.958	0.199	0.203
H3	0%	0.939	0.923	0.933	0.166	0.328
	15%	0.937	0.923	0.932	0.195	0.413
	30%	0.930	0.924	0.933	0.259	0.443
	45%	0.931	0.921	0.930	0.238	0.466
	60%	0.932	0.918	0.928	0.184	0.340
	75%	0.930	0.911	0.930	0.128	0.272

4. In the hazard function, using terms that include double interaction gives better results than when using simple linear hazard function or nonlinear including at least double interactions.
5. Increasing shape parameter of the Weibull distribution improves the accuracy of predictions in each of the three models. This increase in the accuracy of predictions is more evident in the state of increasing the shape parameter from $\alpha = 1$ to $\alpha = 2$ than in other states.

4. Discussion and conclusion

In this article, three parametric, semi-parametric, and non-parametric approaches were considered for modeling survival data on different censor levels. Data were produced by simulation and were used in the three approaches to compare predictions accuracy. In this simulation, three linear hazard models with different complexity levels were considered, and the predictions strength of this models were compared using ROC-test and concordance index.

The obtained results show when it is known that the survival time distribution is Weibull, it is preferable to use survival parametric model instead of Cox semi-parametric or ANN non-parametric models. In the case, when (1) there is doubt about survival time distribution, and (2) nonlinear hazard models with high complexity are used, then using ANNs seems suitable. Accuracy in predictions based on the survival parametric model in the Weibull distribution is increased compared to the exponential distribution with increasing shape parameter. Also, using the nonlinear hazard function with a double interaction gives better results compared to the linear hazard function and the nonlinear hazard function, including at least double interactions. In addition the obtained results show that censor levels difference has no significant effect on the accuracy of predictions from the three hazard models (H_1, H_2, H_3).

Almost these results with findings obtained by A. Biglarian (Biglarian and Bakhshi, 2013) are in a direction but more informative due to more general case Weibull distribution and considering parametric survival models. These results in various fields related to the production of survival data are useful. This is due to determine the condition use of Neural Network model instead of parametric and semi-parametric survival models. In most cases, the use of the ANN model rather than the classic model of survival which has a less restrictive and better results. At the end, it is recommended that other distributions survival such as lognormal, log-logistic, and gamma similarly be compared with the ANN model.

Acknowledgment

The authors feel much obligated for the time spent to review this article by highly informed referees.

ORCID

Mahdi Jabbari Nooghabi  <http://orcid.org/0000-0002-5636-2209>

References

- Bender, R., Augustin, T., Blettner, M. (2003). Generating survival times to simulate cox proportional hazards models. *Sonderforschungsbereich* 386, 338.
- Biglarian, A., Bakhshi, E. (2013). Nonlinear survival regression using artificial neural network. *Journal of Probability and Statistics* 2013, 753930.

- Carroll, K. J. (2003). On the use and utility of the Weibull model in the analysis of survival data, see front matter Elsevier Inc. All rights reserved.
- Ciampi, A., Lechevallier, Y. (1997). *Statistical Models and Artificial Neural Networks*. McGill University, Montreal, QC, Canada INRIA-Rocquencourt.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27:861–874.
- Fox, J., Weisberg, S. (2011). Cox proportional-hazards regression for survival data in R. An Appendix to “A R Companion to Applied Regression,” 2nd ed.
- Kleinbaum, D. G., Klein, M. (2012). *Survival Analysis a Self Learning Text*. 3rd ed. New York: Springer.
- Hastie, T., Tibshirani, R., Friedman, J. (2012). *The Elements of Statistical Learning*. 2nd ed. Springer series in statistics. New York: Springer.
- Haykin, S. (1994). *Neural Networks. A Comprehensive Foundation*. New York: Macmillan College Publishing.
- Lee, E. T., Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis*. Wiley Series in Probability and Statistics, Hoboken, NJ, USA: Wiley-Interscience.
- Ripley, M., Harris, A. L., Tarassenko, L. (2004). Non-linear survival analysis using neural networks. *Statistics in Medicine* 23:825–842.
- Ripley, B. D., Ripley, R. M. (2001). Neural networks as statistical methods in survival analysis. In: *Clinical Applications of Artificial Neural Networks*, Cambridge, UK: Cambridge University Press, pp. 237–255.
- Robin, X., Turk, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., Muller, M. (2011). pROC: An open-source package for R to analyze and compare ROC curves. *BMC Bioinformatics* 12:77.
- Xiang, A., Lapuertab, P., Ryutova, A., Buckley, J., Azena, S. (2000). Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational Statistics & Data Analysis* 34:243–257.