Parametric regression model for survival data: Weibull regression model as an example

Zhongheng Zhang

Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University, Jinhua 321000, China

Correspondence to: Zhongheng Zhang, MMed. 351#, Mingyue Road, Jinhua 321000, China. Email: zh_zhang1984@hotmail.com.

Author's introduction: Zhongheng Zhang, MMed. Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University. Dr. Zhongheng Zhang is a fellow physician of the Jinhua Municipal Central Hospital. He graduated from School of Medicine, Zhejiang University in 2009, receiving Master Degree. He has published more than 35 academic papers (science citation indexed) that have been cited for over 200 times. He has been appointed as reviewer for 10 journals, including *Journal of Cardiovascular Medicine*, *Hemodialysis International*, *Journal of Translational Medicine*, *Critical Care*, *International Journal of Clinical Practice*, *Journal of Critical Care*. His major research interests include hemodynamic monitoring in sepsis and septic shock, delirium, and outcome study for critically ill patients. He is experienced in data management and statistical analysis by using R and STATA, big data exploration, systematic review and meta-analysis.

Zhongheng Zhang, MMed.

Abstract: Weibull regression model is one of the most popular forms of parametric regression model that it provides estimate of baseline hazard function, as well as coefficients for covariates. Because of technical difficulties, Weibull regression model is seldom used in medical literature as compared to the semi-parametric proportional hazard model. To make clinical investigators familiar with Weibull regression model, this article introduces some basic knowledge on Weibull regression model and then illustrates how to fit the model with R software. The *SurvRegCensCov* package is useful in converting estimated coefficients to clinical relevant statistics such as hazard ratio (HR) and event time ratio (ETR). Model adequacy can be assessed by inspecting Kaplan-Meier curves stratified by categorical variable. The *eha* package provides an alternative method to model Weibull regression model. The check.dist() function helps to assess goodness-of-fit of the model. Variable selection is based on the importance of a covariate, which can be tested using anova() function. Alternatively, backward elimination starting from a full model is an efficient way for model development. Visualization of Weibull regression model after model development is interesting that it provides another way to report your findings.

Keywords: Survival analysis; parametric model; Weibull regression model

Submitted May 20, 2016. Accepted for publication Jun 23, 2016.
doi: 10.21037/atm.2016.08.45
View this article at: http://dx.doi.org/10.21037/atm.2016.08.45

Introduction

While semi-parametric model focuses on the influence of covariates on hazard, fully parametric model can also calculate the distribution form of survival time. Advantages of parametric model in survival analysis include: (I) the distribution of survival time can be estimated; (II) full maximum likelihood can be used to estimate parameters; (III) residuals can represent the difference between observed and estimated values of time; (IV) estimated parameters provide clinically meaningful estimates of effect (1). There are a variety of models to be specified for accelerated failure time model including exponential, Weibull and log-logistic regression models. In this article, Weibull regression model is employed as an example to illustrate parametric model development and visualization.

Weibull regression model

Before exploring R for Weibull model fit, we first need to review the basic structure of the Weibull regression model. The distribution of time to event, T, as a function of single covariate is written as (1):

$$In(T) = \beta_0 + \beta_1 x + \sigma\varepsilon \qquad [1]$$

where β_1 is the coefficient for corresponding covariate, ε follows extreme minimum value distribution G(0, σ)and σ is the shape parameter. This is also called the accelerated failure-time model because the effect of the covariate is multiplicative on time scale and it is said to "accelerate" survival time. In contrast, the effect of covariate is multiplicative on hazard scale in the proportional hazard model. The hazard function of Weibull regression model in proportional hazards form is:

$$\begin{aligned} & h(t,x,\beta,\lambda) \\ & = \lambda t^{\lambda-1} e^{-1(\beta_0+\beta_1 x)} \\ & = \lambda t^{\lambda-1} e^{-\lambda\beta_0} e^{-\lambda\beta_1 x} \\ & = \lambda\gamma t^{\lambda-1} e^{-\lambda\beta_1 x} \\ & = h_0(t) e^{\theta_1 x} \end{aligned} \qquad [2]$$

where $\gamma = e^{\frac{-\beta_0}{\sigma}} = e^{\theta_0}$, $\theta_1 = -\beta_1 / \sigma$, and the baseline hazard function is $h_0(t) = \lambda\gamma t^{\lambda-1}$. σ is a variance-like parameter on log-time scale. $\gamma = 1/\sigma$ is usually called a scale parameter. Parameter λ is a shape parameter. Parameter θ_1 has a hazard ratio (HR) interpretation for subject-matter audience.

The accelerated failure-time form of the hazard function can be written as:

$$\begin{aligned} & h(t,x,\beta,\lambda) \\ & = \lambda t^{\lambda-1} e^{-\lambda(\beta_0+\beta_1 x)} \\ & = \lambda\gamma \left(te^{-\beta_1 x}\right)^{\lambda-1} e^{-\beta_1 x} \end{aligned}$$

Weibull regression model can be written in both accelerated and proportional forms, allowing for simultaneous description of treatment effect in terms of HR and relative change in survival time [event time ratio (ETR)] (2).

Fitting Weibull regression model with R

The survreg() function contained in *survival* package is able to fit parametric regression model. Let's first load the package into the workspace. To build a Weibull regression model, the *dist* argument should be set to a string value "weibull", indicating the distribution of response variable follows Weibull distribution. The summary() function is to print content of the returned object of class *survreg*.

```
> library(survival)
> wei.lung<-survreg(Surv(time, status)~ph.ecog+sex+age,lung,
dist='weibull')
> summary(wei.lung)

Call:
survreg(formula = Surv(time, status) ~ ph.ecog + sex + age, data
= lung,
 dist = "weibull")
                    Value        Std. Error    z          p
```

```
(Intercept)   6.27344    0.45358    13.83    1.66e-43
ph.ecog      -0.33964    0.08348    -4.07    4.73e-05
sex           0.40109    0.12373     3.24    1.19e-03
age          -0.00748    0.00676    -1.11    2.69e-01
Log(scale)   -0.31319    0.06135    -5.11    3.30e-07

Scale= 0.731

Weibull distribution
Loglik(model)= -1132.4 Loglik(intercept only)= -1147.4
         Chisq= 29.98 on 3 degrees of freedom, p= 1.4e-06
Number of Newton-Raphson Iterations: 5
n=227 (1 observation deleted due to missingness)
```

The output first recalls the structure of the Weibull regression model, including the covariates. Next, the coefficients of each covariate are shown, together with standard error and P values. Scale is an important parameter in Weibull regression model and is shown in the following line. A log likelihood test shows that the model is significantly better than null model (P=1.4e–06). However, the estimated coefficients are not clinically meaningful. That is why Weibull regression model is not widely used in medical literature. Since Weibull regression model allows for simultaneous description of treatment effect in terms of HR and relative change in survival time, ConvertWeibull() function is used to convert output from survreg() to more clinically relevant parameterization. The function is contained in *SurvRegCensCov* package and we need to install it first.

```
> install.packages("SurvRegCensCov")
> library(SurvRegCensCov)
> ConvertWeibull(wei.lung,conf.level = 0.95)
$vars
                Estimate          SE
lambda          0.0001876914      0.0001506884
gamma           1.3677851193      0.0839087686
ph.ecog         0.4645519368      0.1136759822
sex            -0.5486056737      0.1673299432
age             0.0102247948      0.0092298732

$HR
             HR           LB           UB
ph.ecog      1.5913010    1.2734772    1.9884447
sex          0.5777548    0.4162096    0.8020013
age          1.0102772    0.9921654    1.0287197

$ETR
             ETR          LB           UB
ph.ecog      0.7120280    0.6045610    0.8385984
sex          1.4934525    1.1718447    1.9033242
age          0.9925524    0.9794818    1.0057975
```

The first table of the output displays parameters of the Weibull regression model. Lambda and gamma are scale and shape parameters of Weibull distribution. The estimate for each covariate is different from that displayed in the value column of the summary() output. The relationship can be described by an equation $\beta = -\alpha/\sigma$, where α is parameter for each of the covariate and σ is the scale (2). In our example, β is the estimate in the first table of the ConvertWeibull() output and α is displayed in the output of summary(wei.lung). The second table shows the HR and corresponding 95% confidence interval. The last table displays the ETR and its 95% confidence interval. Female reduces the risk of death compared to male by 42% (HR =0.58), and female significantly increases the survival time by approximately 50% (ETR =1.49). Although HR is more widely reported in medical literature and is familiar to clinicians, ETR may be easier to understand.

Alternatively, the Weibull regression model can be fit with WeibullReg() function. In essence, it is the combination of survreg() and ConvertWeibull().

```
> wei.lung.alt<-WeibullReg(Surv(time,status)~ph.
ecog+sex+age,data=lung,conf.level=0.95)
```

Adequacy of Weibull model

Weibull model with categorical variables can be checked for its adequacy by stratified Kaplan-Meier curves. A plot of log survival time versus log[–log(KM)] will show linear and parallel lines if the model is adequate (3).

```
> WeibullDiag(Surv(time,status)~sex,data=lung)
```

Figure 1 is the Weibull regression diagnostic plot showing that the lines for male and female are generally parallel and linear in its scale.

Weibull regression model with eha package

An alternative way to model Weibull regression model is via

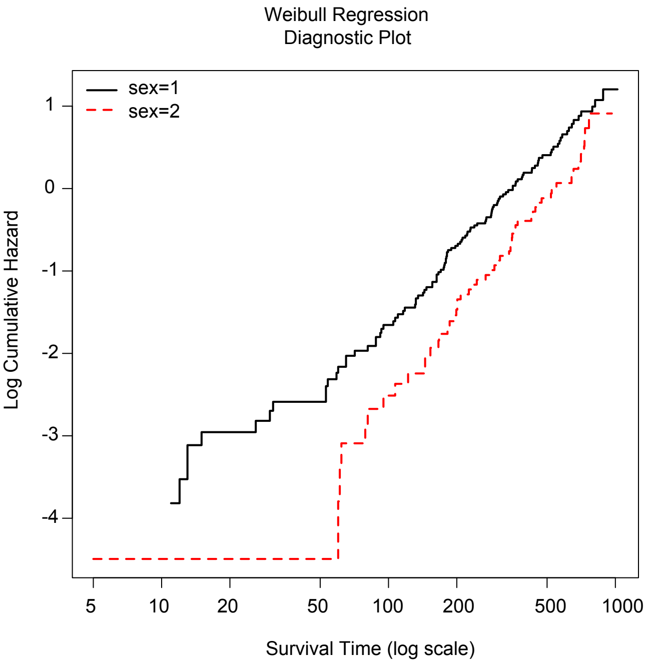

Figure 1 Weibull regression diagnostic plot showing that the lines for male and female are generally parallel and linear in its scale.

eha package. This package provides a variety of functions for Weibull regression model. Let's now first install the package and load it into the workspace.

```
> install.packages("eha")
> library(eha)
> lung.alt<-weibreg( Surv(time, status)~age+sex+ph.ecog, data
= lung)
> lung.alt
Call:
weibreg(formula = Surv(time, status) ~ age + sex + ph.ecog,
data = lung)

Covariate    Mean     Coef     Exp(Coef)  se(Coef)  Wald p
age          61.963   0.010    1.010      0.009     0.268
sex          1.439    -0.549   0.578      0.167     0.001
ph.ecog      0.853    0.465    1.591      0.114     0.000

log(scale)            6.273    530.296    0.454     0.000
log(shape)            0.313    1.368      0.061     0.000


Events      164
Total time at risk    69522
Max. log. likelihood   -1132.4
LR test statistic    30
Degrees of freedom    3
Overall p-value    1.3944e-06
```

The argument of weibreg() function is similar to that of the survreg(). The coefficient of covariates in the above output is the HR in log scale. Thus, the exponentiation of coefficient gives the HR.

Hazard, cumulative hazard, density and survivor functions can be plotted from the output of a Weibull regression model.

```
> par(mfrow=c(2,2))
> plot(lung.alt, fn=c("sur"),new.data=c(80,2,3))
> plot(lung.alt, fn=c("sur"),new.data=c(60,2,3))
> plot(lung.alt, fn=c("sur"),new.data=c(40,2,3))
> plot(lung.alt, fn=c("sur"),new.data=c(20,2,3))
```

Figure 2 is the graphical display of the output of Weibull regression model. The *fn* argument specifies the functions to be plotted. It receives a vector of string values, choosing from "haz", "cum", "den" and "sur". The *newdata* argument specifies covariate values at which to plot the function. If covariates are left unspecified, the default value is the mean of the covariate in the training dataset. In the example, four plots were drawn at age of 80, 60, 40 and 20 years old (in the order from left to right and from top to bottom). The *sex* and *ph.ecog* variables were set at values of 2 and 3, respectively.

Graphical goodness-of-fit test

The *eha* package has a function check.dist() to test the goodness-of-fit by graphical visualization. It compares the cumulative hazards functions for non-parametric and parametric model, requiring objects of "coxreg" and "phreg" as the first and second argument.

```
> phreg.lung<-phreg(Surv(time, status)~ph.
ecog+sex+age,lung, dist='weibull')
> coxreg.lung<-coxreg(Surv(time, status)~ph.
ecog+sex+age,lung)
> check.dist(coxreg.lung,phreg.lung)
```

The solid line is the parametric Weibull cumulative hazard function and the dashed line is non-parametric

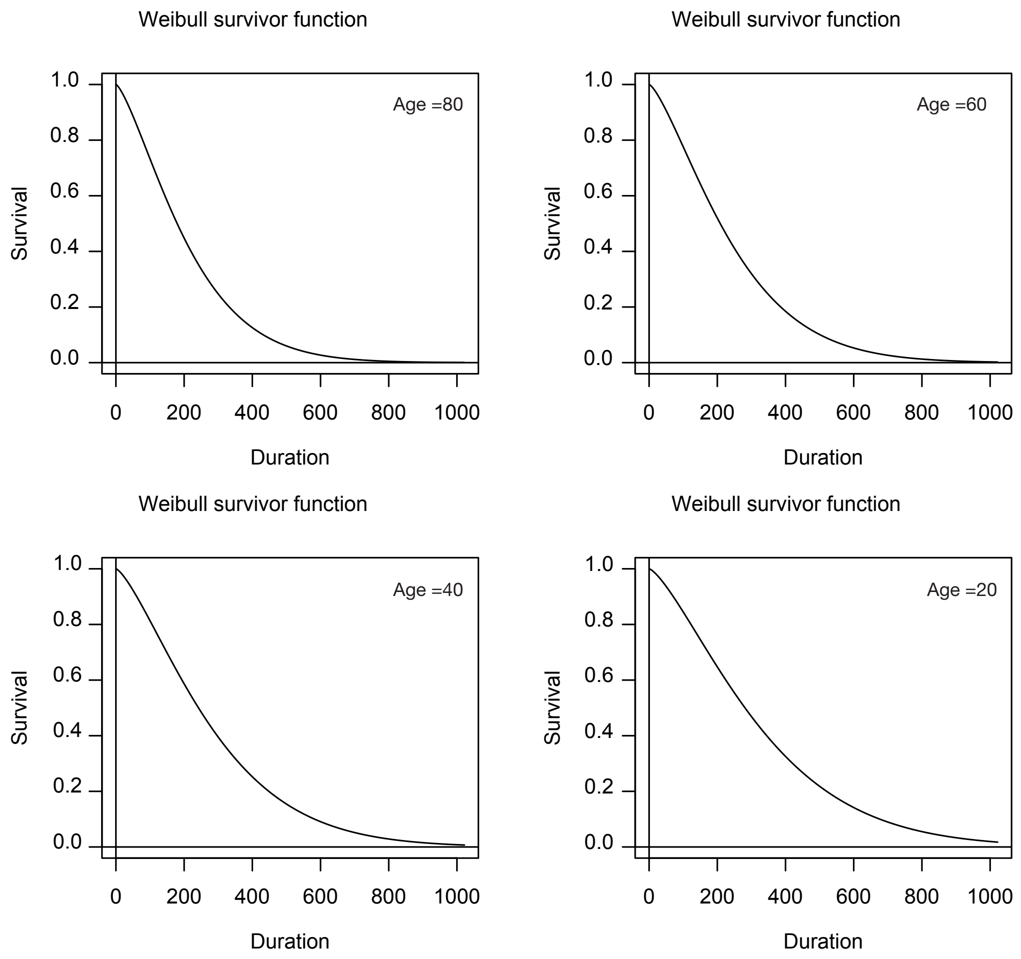

Figure 2 Graphical display of the output of Weibull regression model. The four survivor function plots correspond to ages of 80, 60, 40 and 20. Variables sex and ph.ecog are set to values of 2 and 3, respectively.

function. It appears that the parametric function fits well to the semi-parametric function (*Figure 3*). Note that non-parametric model is closer to the observed data because no function is assumed for the baseline hazard function.

Variable selection and model development

Like generalized linear model development (4), it is essential to include statistically important and clinically relevant covariates into the model in fitting parametric regression model. While clinical relevance is judged by clinical expertise, the statistical importance is determined by software. The anova() function tests the statistical importance of a covariate, interaction and non-linear terms. The function reports Chi-square statistics and associated P value. Also, it provides dot charts depicting the importance of variables in the model.

```
> psm.lung<-psm(Surv(time, status)~ph.ecog+sex*age+ph.
karno+pat.karno+meal.cal+wt.loss,lung, dist='weibull')
> anova(psm.lung)
      Wald Statistics     Response: Surv(time, status)

Factor                                    Chi-     d.f.  P
                                          Square
ph.ecog                                   13.86    1     0.0002
sex (Factor+Higher Order Factors)         10.24    2     0.0060
All Interactions                          3.22     1     0.0728
age (Factor+Higher Order Factors)         3.75     2     0.1532
All Interactions                          3.22     1     0.0728
ph.karno                                  5.86     1     0.0155
```

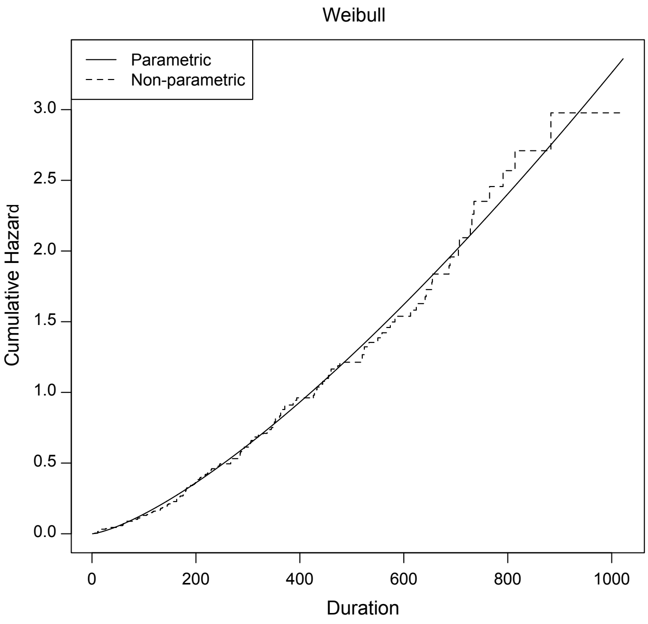

Figure 3 Goodness-of-fit test by graphical comparison between parametric and non-parametric regression models. It appears that the parametric function fits well to the non-parametric function.

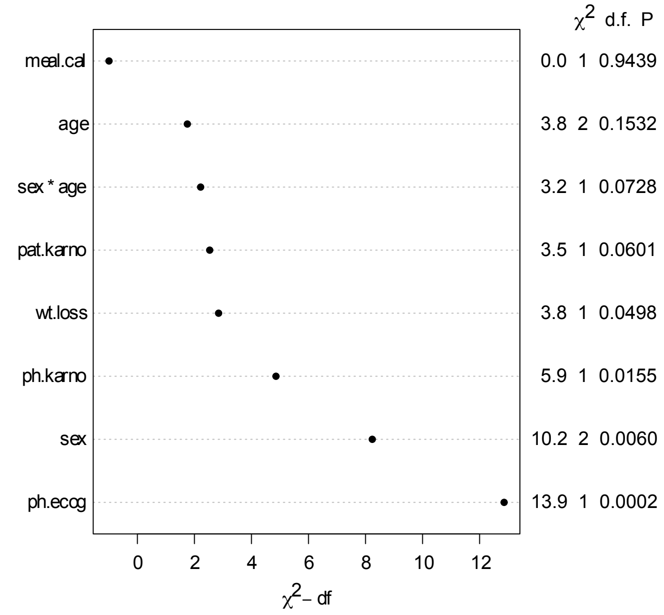

Figure 4 Dot chart showing relative importance of covariates.

```
pat.karno                              3.54    1    0.0601
meal.cal                               0.00    1    0.9439
wt.loss                                3.85    1    0.0498
sex * age(Factor+HigherOrderFactors)   3.22    1    0.0728
TOTAL                                  33.18   8    0.0001
> plot(anova(psm.lung),margin=c("chisq", "d.f.", "P"))
```

In the example, we included all available covariates into the model to rank their statistical importance. This is often the case in real research setting that researchers have no prior knowledge on which variable should be included. The first argument of psm() function is a formula describing the response variable and covariates, as well as interaction between predictors. The output of anova() includes variable names, Chi-square statistics, degree of freedom and p-value. Dot chart is drawn with generic function plot(). It appears that meal.cal is the least important variable and ph.ecog is the most important one (*Figure 4*). Some pre-specified rules can be applied to inclusion/exclusion of variables (4).

Alternatively, model development can be done with backward elimination on covariates. This method starts with a full model that included all available covariates and then applies Wald test to examine the relative importance of each one. Statistical significance level for a covariate to stay in a model can be specified. R provides a function fastbw() to perform fast backward variable selection.

```
> fastbw(psm.lung,rule="aic")

Deleted    Chi-  d.f.   P        Residual  d.f.P          AIC
           Sq
meal.cal   0.00  1      0.9439   0.00      1   0.9439    -2.00
sex        1.94  1      0.1634   1.95      2   0.3777    -2.05
pat.karno  2.75  1      0.0970   4.70      3   0.1950    -1.30
wt.loss    2.36  1      0.1248   7.06      4   0.1328    -0.94

Approximate Estimates after Deleting Factors

             Coef        S.E.       Wald Z    P
(Intercept)  8.276947    0.936299   8.840     0.000e+00
ph.ecog      -0.546884   0.137424   -3.980    6.905e-05
age          -0.015444   0.007976   -1.936    5.283e-02
ph.karno     -0.015375   0.007394   -2.080    3.757e-02
sex*age      0.005949    0.002182   2.727     6.395e-03

Factors in Final Model

[1] ph.ecog age   ph.karno sex * age
```

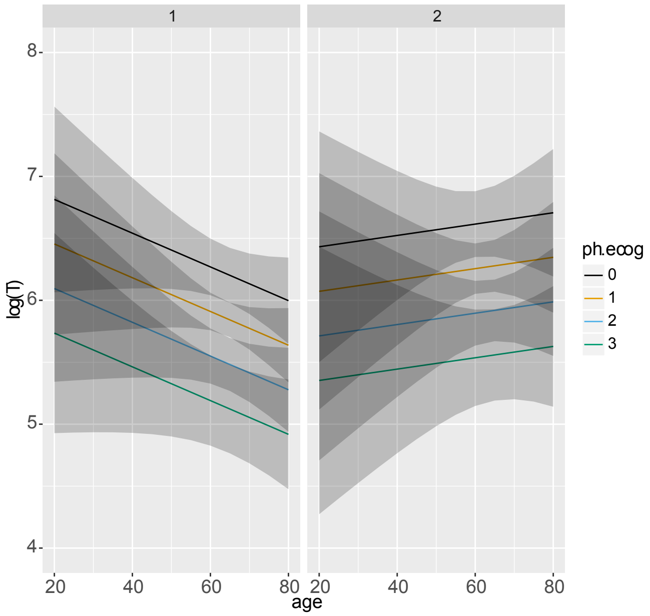

Figure 5 Graphical presentation of the relationship between covariates and survival time on log scale. The effect of age on survival time is dependent on sex. While older patients have shorter survival time in the male, older patients show longer survival time in the female.

The first argument of fastbw() receives an object fit by psm(). The *rule* argument defines stopping rule for backward elimination. The default is Akaike's information criterion (AIC). If P value is used as the stopping rule (rule="p"), the significance level for staying in a model can be modified using *sls* argument (sls =0.1 for example). The output shows that variables meal.cal, sex, pat.karno and wt.loss are eliminated from the model based on AIC. Sometimes if you want to retain a covariate in the model based on clinical judgment, the *force* argument can be employed. It passes a vector of integers specifying covariates to be retained in the model. Intercept is not counted.

Visualization of Weibull regression model

Weibull model can be used to predict outcomes of new subjects, allowing predictors to vary. In Weibull regression model, the outcome is median survival time for a given combination of covariates. We first use Predict() to calculate median survival time in log scale, then use ggplot() function to draw plots.

```
> psm.lung1<-psm(Surv(time, status)~ph.ecog+sex*age,lung,
dist='weibull')
> ggplot(Predict(psm.lung1, age=seq(20,80,by=5),ph.
ecog=c(0,1,2,3),sex=c(1,2)))
```

In the example, an interaction term sex*age is specified. We let variable age to vary between 20 to 80 years old. Both male and female, and all four levels of ph.ecog are considered. *Figure 5* shows the output of ggplot() function. The effect of age on survival time is dependent on sex. While older age is associated with shorter survival time in the male, it is associated with longer survival time in the female.

Figure 5 visualizes relationship between covariates. Occasionally, investigators may be interested in survivor and/or hazard functions of individuals with given covariate patterns. The *smoothSurv* package provides functions for this purpose. Similarly to the previous model building strategy, we first fit a model including interaction terms between sex and age.

```
> install.packages("smoothSurv")
> library(smoothSurv)
> smooth.lung <- smoothSurvReg(Surv(time, status)~ph.
ecog+sex*age,data=lung, init.dist='weibull')
> cov<-matrix(c(0,1,2,3,1,2,2,2,20,30,40,70,20,60,80,140),ncol=
4,byrow=FALSE)
> cov
        [,1]    [,2]    [,3]    [,4]
[1,]    0       1       20      20
[2,]    1       2       30      60
[3,]    2       2       40      80
[4,]    3       2       70      140
```

A matrix object of *cov* is created representing 4 patients whose survival time is unknown and the treating physician wants to make a prediction based on Weibull regression model. The number of columns of the matrix should be equal to the number of covariates in the model, including interaction terms.

```
> par(mfrow=c(2,2))
> survfit(smooth.lung,cov=cov)
> survfit(smooth.lung, cdf = TRUE,cov=cov)
> hazard(smooth.lung,cov=cov)
> fdensity(smooth.lung,cov=cov)
```

The output is a series of plots showing survivor, cumulative distribution, hazard and density functions

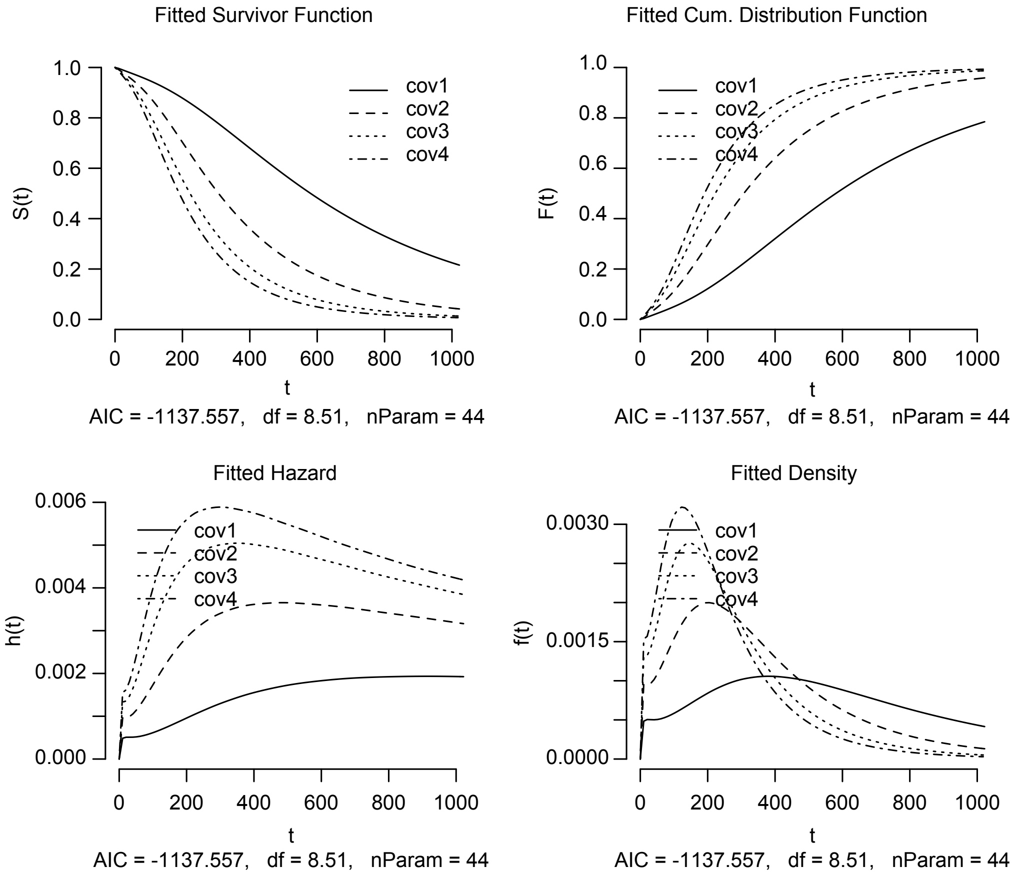

Figure 6 Survivor, cumulative distribution, hazard and density functions of four subjects. Cov1 to cov4 are indicators of four patients with given covariate patterns.

(*Figure 6*). Cov1 to cov4 are indicators of four patients with given covariate patterns. Important parameters of the model are displayed at the bottom of each plot.

Acknowledgements

None.

Footnote

Conflicts of Interest: The author has no conflicts of interest to declare.

References

1. Hosmer DW Jr, Lemeshow S, May S. editors. Applied Survival Analysis, 2nd ed. New York: John Wiley & Sons, Inc., 2008:1.
2. Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. Control Clin Trials 2003;24:682-701.
3. Klein JP, Moeschberger ML. editors. Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed. New York: Springer, 2005:1.
4. Zhang Z. Model building strategy for logistic regression: purposeful selection. Ann Transl Med 2016;4:111.

Cite this article as: Zhang Z. Parametric regression model for survival data: Weibull regression model as an example. Ann Transl Med 2016;4(24):484. doi: 10.21037/atm.2016.08.45