Stratified Weibull Regression Model for Interval-Censored Data

by Xiangdong Gu, David Shapiro, Michael D. Hughes and Raji Balasubramanian

Abstract Interval censored outcomes arise when a silent event of interest is known to have occurred within a specific time period determined by the times of the last negative and first positive diagnostic tests. There is a rich literature on parametric and non-parametric approaches for the analysis of interval-censored outcomes. A commonly used strategy is to use a proportional hazards (PH) model with the baseline hazard function parameterized. The proportional hazards assumption can be relaxed in stratified models by allowing the baseline hazard function to vary across strata defined by a subset of explanatory variables. In this paper, we describe and implement a new R package **straweib**, for fitting a stratified Weibull model appropriate for interval censored outcomes. We illustrate the R package **straweib** by analyzing data from a longitudinal oral health study on the timing of the emergence of permanent teeth in 4430 children.

Introduction

In many clinical studies, the time to a silent event is known only up to an interval defined by the times of the last negative and first positive diagnostic test. Event times arising from such studies are referred to as 'interval-censored' data. For example, in pediatric HIV clinical studies, the timing of HIV infection is known only up to the interval from the last negative to the first positive HIV diagnostic test (Dunn et al., 2000). Examples of interval-censored outcomes can also be found in many other medical studies (Gomez et al., 2009).

A rich literature exists on the analysis of interval-censored outcomes. Non-parametric approaches include the self-consistency algorithm for the estimation of the survival function (Turnbull, 1976). A semi-parametric approach based on the proportional hazards model has been developed for interval-censored data (Finkelstein, 1986; Goetghebeur and Ryan, 2000). A variety of parametric models can also be used to estimate the distribution of the time to the event of interest, in the presence of interval-censoring (Lindsey and Ryan, 1998). An often used parametric approach for the analysis of interval-censored data is based on the assumption of a Weibull distribution for the event times (Lindsey and Ryan, 1998). The Weibull distribution is appropriate for modeling event times when the hazard function can be reliably assumed to be monotone. Covariate effects can be modeled through the assumption of proportional hazards (PH), which assumes that the ratio of hazard functions when comparing individuals in different strata defined by explanatory variables is time-invariant. The article by Gomez et al. (2009) presents a comprehensive review of the state-of-the-art techniques available for the analysis of interval-censored data.

In this paper, we implement a parametric approach for modeling covariates applicable to interval-censored outcomes, but where the assumption of proportional hazards may be questionable for a certain subset of explanatory variables. For this setting, we implement a stratified Weibull model by relaxing the PH assumption across levels of a subset of explanatory variables. We compare the proposed model to an alternative stratified Weibull regression model that is currently implemented in the R package **survival** (Therneau, 2012). We illustrate the difference between these two models analytically and through simulation.

The paper is organized as follows: In Section 2, we present and compare two models for relaxing the PH assumption, based on the assumption of a Weibull distribution for the time to event of interest. In this section, we discuss estimation of the unknown parameters of interest, hazard ratios comparing different groups of subjects based on specific values of explanatory covariates and tests of the PH assumption. These methods are implemented in a new R package, **straweib** (Gu and Balasubramanian, 2013). In Section 3, we perform simulation studies to compare two stratified Weibull models implemented in R packages **straweib** and **survival**. In Section 4, we illustrate the use of the R package **straweib** by analyzing data from a longitudinal oral health study on the timing of the emergence of permanent teeth in 4430 children in Belgium (Leroy et al., 2003; Gomez et al., 2009). In Section 5, we discuss the models implemented in this paper and present concluding remarks.

Weibull regression models

Let T denote the continuous, non-negative random variable corresponding to the time to event of interest, with corresponding probability distribution function (pdf) and cumulative distribution func-

tion (cdf), denoted by $f(t)$ and $F(t)$, respectively. We let $S(t) = 1 - F(t)$ to denote the corresponding survival function and $h(t) = lim_{\delta t \to 0} \frac{P(t \leq T < t+\delta t | T \geq t)}{\delta t}$ to denote the hazard function. We let $\mathbf{Z}$ denote the $p \times 1$ vector of explanatory variables or covariates.

We assume that the random variable $T \mid \mathbf{Z} = \mathbf{0}$ is distributed according to a Weibull distribution, with scale and shape parameters denoted by λ and γ, respectively. The well known PH model to accommodate the effect of covariates on T is expressed as:

$$h(t \mid \mathbf{Z}) = h(t \mid \mathbf{Z} = \mathbf{0}) \times \exp(\boldsymbol{\beta}'\mathbf{Z}),$$

where $\boldsymbol{\beta}$ denotes the $p \times 1$ vector of regression coefficients corresponding to the vector of explanatory variables, $\mathbf{Z}$.

Thus, under the Weibull PH model, the survival and hazard functions corresponding to T can be expressed as

$$S(t \mid \mathbf{Z}) = \exp\left(-\lambda \exp\left(\boldsymbol{\beta}'\mathbf{Z}\right) t^{\gamma}\right) \tag{1}$$

$$h(t \mid \mathbf{Z}) = \lambda \exp(\boldsymbol{\beta}'\mathbf{Z}) \gamma t^{\gamma - 1} \tag{2}$$

where, $\lambda > 0$ and $\gamma > 0$ correspond to the scale and shape parameters corresponding to T when $\mathbf{Z} = \mathbf{0}$. The hazard ratio comparing two individuals with covariate vectors $\mathbf{Z}$ and $\mathbf{Z}^*$ is equal to $\exp(\boldsymbol{\beta}'(\mathbf{Z} - \mathbf{Z}^*))$.

Stratified Weibull regression model implemented in the R package survival

In this section, we describe the stratified Weibull PH regression model implemented in the the R package **survival** (Therneau, 2012).

Consider the following log-linear model for the random variable T:

$$log(T \mid \mathbf{Z}) = \mu + \alpha_1 Z_1 + \cdots \alpha_p Z_p + \sigma\epsilon$$

where, $\alpha_1, \cdots, \alpha_p$ denote unknown regression coefficients corresponding to the p dimensional vector of explanatory variables, μ denotes the intercept, and σ denotes the scale parameter. The random variable ϵ captures the random deviation of event times on the natural logarithm scale (i. e. $\log(T)$) from the linear model as a function of the covariate vector $\mathbf{Z}$. In general, the log-linear form of the model for T can be shown to be equivalent to the accelerated failure time (AFT) model (Collett, 2003).

The assumption of a standard Gumbel distribution with location and scale parameters equal to 0 and 1, respectively, implies that the random variable T follows a Weibull distribution. Moreover, in this case, both the PH and AFT assumptions (or equivalently, the log-linear model) lead to identical models with different parameterizations (Collett, 2003). The survival and hazard functions can be expressed as:

$$S(t \mid \mathbf{Z}) = \exp\left[-\exp\left(\frac{\log(t) - \mu - \boldsymbol{\alpha}'\mathbf{Z}}{\sigma}\right)\right] \tag{3}$$

$$h(t \mid \mathbf{Z}) = \exp\left[-\frac{\mu + \boldsymbol{\alpha}'\mathbf{Z}}{\sigma}\right] \frac{1}{\sigma} t^{\frac{1}{\sigma}-1} \tag{4}$$

The coefficients for the explanatory variables ($\boldsymbol{\beta}$) in the hazard function ($h(t \mid \mathbf{Z})$) are equal to $-\frac{\boldsymbol{\alpha}}{\sigma}$. Moreover, there is a one-one correspondence between the parameters $\lambda, \gamma, \boldsymbol{\beta}$ in equations (1)-(2) and the parameters $\mu, \sigma, \boldsymbol{\alpha}$ in equations (3)-(4), where $\lambda = \exp(-\frac{\mu}{\sigma}), \gamma = \sigma^{-1}$ and $\beta_j = -\frac{\alpha_j}{\sigma}$ (Collett, 2003).

The log-linear form of the Weibull model can be generalized to allow arbitrary baseline hazard functions within subgroups defined by a stratum indicator $S = 1, \cdots, s$. Thus, the stratified Weibull regression model for an individual in the j^{th} stratum is expressed as:

$$log\left(T \mid \mathbf{Z}, S = j\right) \quad = \quad \mu_j + \alpha_1 Z_1 + \cdots \alpha_p Z_p + \sigma_j \epsilon$$

where μ_j and σ_j denote stratum specific intercept and scale parameters. This model is implemented in the R package **survival** (Therneau, 2012). In this model, the regression coefficients $\boldsymbol{\alpha}$ on the AFT scale are assumed to be stratum independent.

However, the hazard ratio comparing two individuals with covariate vectors and stratum indicators denoted by ($\mathbf{Z}$, $S = j$) and ($\mathbf{Z}^*, S = k$) is stratum specific and is given by:

$$\frac{h\left(t \mid S = j, \mathbf{Z}\right)}{h\left(t \mid S = k, \mathbf{Z}^*\right)} = t^{1/\sigma_j - 1/\sigma_k} \frac{\sigma_k}{\sigma_j} \exp\left(\frac{\mu_k}{\sigma_k} - \frac{\mu_j}{\sigma_j}\right) \exp\left(\boldsymbol{\alpha}'\left(\mathbf{Z}^*/\sigma_k - \mathbf{Z}/\sigma_j\right)\right)$$

For $j \neq k$, the hazard ratio varies with time t. However, when $j = k$, the hazard ratio comparing two individuals within the same stratum $S = j$ is invariant with respect to time t but is stratum-dependent and reduces to:

$$\frac{h\left(t \mid S=j, \boldsymbol{Z}\right)}{h\left(t \mid S=j, \boldsymbol{Z}^{*}\right)}=\exp\left(\frac{\boldsymbol{\alpha}^{'}}{\sigma_{j}}\left(\boldsymbol{Z}^{*}-\boldsymbol{Z}\right)\right) \tag{5}$$

Stratified Weibull regression model implemented in R package straweib

In this section, we describe the stratified Weibull regression model that is implemented in the new R package, **straweib** (Gu and Balasubramanian, 2013).

To relax the proportional hazards assumption in the Weibull regression model, we propose the following model for an individual in the stratum $S = j$:

$$h(t \mid \boldsymbol{Z}, S=j)=\lambda_{j} \exp\left(\boldsymbol{\beta}^{\prime} \boldsymbol{Z}\right) \gamma_{j} t^{\gamma_{j}-1} \tag{6}$$

Equivalently, the model can be stated in terms of the survival function as:

$$S(t \mid \boldsymbol{Z}, S=j)=\exp\left(-\lambda_{j} \exp\left(\boldsymbol{\beta}^{\prime} \boldsymbol{Z}\right) t^{\gamma_{j}}\right)$$

Here, we assume that the scale and shape parameters (λ, γ) are stratum specific - however, the regression coefficients $\boldsymbol{\beta}$ are assumed to be constant across strata (S). The hazard ratio comparing two individuals with covariate vectors and stratum indicators denoted by ($\boldsymbol{Z}$, $S = j$) and ($\boldsymbol{Z}^{*}$, $S = k$) is given by:

$$\frac{h(t \mid S=j, \boldsymbol{Z})}{h(t \mid S=k, \boldsymbol{Z}^{*})}=t^{\gamma_{j}-\gamma_{k}} \exp\left(\boldsymbol{\beta}^{'}\left(\boldsymbol{Z}-\boldsymbol{Z}^{*}\right)\right) \frac{\lambda_{j} \gamma_{j}}{\lambda_{k} \gamma_{k}}$$

For $j \neq k$, the hazard ratio varies with time t and thus relaxes the PH assumption. However, for $j = k$, the hazard ratio comparing two individuals within the same stratum $S = j$ reduces to:

$$\frac{h(t \mid S=j, \boldsymbol{Z})}{h(t \mid S=j, \boldsymbol{Z}^{*})}=\exp\left(\boldsymbol{\beta}^{'}\left(\boldsymbol{Z}-\boldsymbol{Z}^{*}\right)\right) \tag{7}$$

This hazard ratio is invariant with respect to time t and stratum S, as in the stratified Cox model (Collett, 2003).

Estimation

Let $u_j = \log(\lambda_j)$ and $v_j = \log(\gamma_j)$. Let n_j denote the number of subjects in stratum $S = j$. For the k^{th} subject in stratum j, let $\boldsymbol{Z}_{jk}$ denote the p dimensional vector of covariates and let a_{jk} and b_{jk} denote the left and right endpoints of the censoring interval. That is, a_{jk} denotes the time of the last negative test and b_{jk} denotes the time of the first positive test for the event of interest. Then the log-likelihood function can be expressed as:

$$\begin{aligned} l(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{\beta})= & \sum_{j=1}^{s} \sum_{k=1}^{n_{j}} \log\{\exp[-\exp[u_{j}+\boldsymbol{\beta}^{\prime} \boldsymbol{Z}_{jk}+\exp(v_{j}) \log(a_{jk})]] \\ & -\exp[-\exp[u_{j}+\boldsymbol{\beta}^{\prime} \boldsymbol{Z}_{jk}+\exp(v_{j}) \log(b_{jk})]]\} \end{aligned}$$

The unknown parameters to be estimated are $\boldsymbol{v}$, $\boldsymbol{u}$, and $\boldsymbol{\beta}$. The log-likelihood function can be optimized using the `optim` function in R. The shape and scale parameters can be estimated from the estimates of $\boldsymbol{v}$ and $\boldsymbol{u}$. The covariance matrix of the estimates of these unknown parameters can be obtained by inverting the negative Hessian matrix that is output from the optimization routine (Cox and Hinkley, 1979).

Test of the PH assumption

One can test whether or not the baseline hazard functions of each strata are proportional to each other, by testing the equality of shape parameters across strata $S = 1, \cdots, s$. That is,

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_s$$

or equivalently,

$$H_0 : v_1 = v_2 = \cdots = v_s.$$

The null hypothesis H_0 can be tested using a likelihood ratio test, by comparing a reduced model that assumes that $\gamma_1 = \gamma_2 == \cdots = \gamma_s$ to the full model in (6) assuming stratum specific shape parameters. We note that the reduced model is equivalent to the Weibull PH model that includes the stratum indicator S as an explanatory variable. Thus the reduced model has $s - 1$ fewer parameters than the

stratified model, or the full model. Let l_F and l_R denote the log-likelihoods of the full and reduced models evaluated at their MLE. Then the test statistic $T = -2(l_R - l_F)$ follows a χ^2_{s-1} distribution under H_0. In addition to the likelihood ratio test, one can also use a Wald test to test the null hypothesis H_0. The R package **straweib** illustrated in Section 3 outputs both the Wald and Likelihood Ratio test statistics.

Estimating hazard ratios

The log hazard ratio comparing two individuals with covariate vectors and stratum indicators denoted by ($\mathbf{Z}$, $S = j$) and ($\mathbf{Z^*}$, $S = j^*$) at time t can be expressed as:

$$r_{tjj^*} = \log\left(R_{tjj^*}\right) = u_j + v_j + \log(t)\exp\left(v_j\right) - u_{j^*} - v_{j^*} - \log(t)\exp\left(v_{j^*}\right) + \boldsymbol{\beta}'\left(\mathbf{Z} - \mathbf{Z}^*\right)$$

Let $\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimates for $\boldsymbol{v}$, $\boldsymbol{u}$ and $\boldsymbol{\beta}$, then r_{tjj^*} can be estimated by

$$\hat{r}_{tjj^*} = \hat{u}_j + \hat{v}_j + \log(t)\exp\left(\hat{v}_j\right) - \hat{u}_{j^*} - \hat{v}_{j^*} - \log(t)\exp\left(\hat{v}_{j^*}\right) + \hat{\boldsymbol{\beta}}'\left(\mathbf{Z} - \mathbf{Z}^*\right)$$

Let $\boldsymbol{w} = (\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{\beta}) = (v_1, v_2, \cdots, v_s, u_1, u_2, \cdots, u_s, \beta_1, \cdots, \beta_p)$. Let $\widehat{\Sigma}$ denote the estimate of the covariance matrix of $\hat{\boldsymbol{w}}$. Let $\boldsymbol{J}_{tjj^*}$ denote the Jacobian vector, $\boldsymbol{J}_{tjj^*} = \frac{\partial r_{tjj^*}}{\partial \boldsymbol{w}}|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}$. Thus, the estimate of the variance of $\hat{r}_{tjj^*}$ is obtained by:

$$\widehat{Var}(\hat{r}_{tjj^*}) = \boldsymbol{J}^T_{tjj^*}\widehat{\Sigma}\boldsymbol{J}_{tjj^*}$$

We obtain a 95% confidence interval for r_{tjj^*} as $\left(\hat{r}_{tjj^*} - 1.96\sqrt{\widehat{Var}(\hat{r}_{tjj^*})}, \hat{r}_{tjj^*} + 1.96\sqrt{\widehat{Var}(\hat{r}_{tjj^*})}\right)$. We exponentiate $\hat{r}_{tjj^*}$ and its corresponding 95% confidence interval to obtain the estimate and the 95% confidence interval for the hazard ratio, R_{tjj^*}. We illustrate the use of the **straweib** R package for obtaining hazard ratios and corresponding confidence intervals in Section 4.

Comparison of models implemented in packages survival and straweib

In this section, we compare the stratified Weibull regression model implemented in the **survival** package to that implemented in our package, **straweib**.

In the absence of stratification, both models are identical and reduce to the Weibull PH model. However, in the presence of a stratification factor, the models implemented by **survival** and **straweib** correspond to different models, resulting in different likelihood functions and inference. As we discussed in Section 2, the hazard ratio between two subjects with different covariate values within same stratum depends on their stratum in the model implemented in the R package **survival** (Equation (5)), whereas the hazard ratio comparing two individuals within the same stratum is invariant to stratum in the model implemented in the R package **straweib** (Equation (7)). In particular, the Weibull model implemented in the **straweib** shares similarities with the semi-parametric, stratified Cox model for right censored data.

To illustrate the difference between the models implemented in the R packages **survival** and **straweib**, we conducted a simulation study in which 1000 datasets were simulated under the model assumed in the **straweib** package (Equation (6)). For each simulated dataset, since both models have the same number of unknown parameters, we compare the values of the log-likelihood evaluated at the MLEs. Datasets were simulated based on the assumptions that there are 3 strata, each with a 100 subjects; the shape parameters (γ) in the three strata were set to 1.5, 2, and 1, respectively; the baseline scale parameters in the three strata (λ) were set to 0.01, 0.015, and 0.02, respectively. We assumed that there are two independent explanatory variables available for each subject, randomly drawn from $N(0, 1)$ random variables. The coefficients corresponding to each of the two covariates were set to 0.5 and 1, respectively. To simulate interval censored outcomes, we first simulated the true event time for each subject by sampling from a Weibull distribution with the appropriate parameters. We assumed that each subject has 20 equally spaced diagnostic tests, at which the true event status is observed. Each test has a probability of 70% being missing. To obtain the maximum likelihood estimates under each model, we used the `survreg` function in the R package **survival** and the `icweib` function in the **straweib** package.

Figure 1 compares the maximized value of the log-likelihoods under both models, when the data are generated using a simulation mechanism that corresponds to the model implemented in the R package **straweib**. The maximized value of the log-likelihood from the R package **survival** is lower than that from the R package **straweib** for 93.1% of simulated datasets. This is expected as in this simulation study the data generating mechanism is identical to the model implemented in the R package **straweib**. In applications where the proportional hazards assumption is questionable, we recommend fitting both models and comparing the resulting maximized values of the log likelihood.

Whether one model is better than another depends on the data.

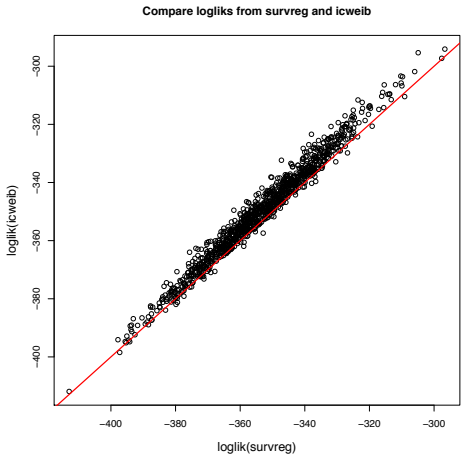

Figure 1: Comparing the maximized values of the log-likelihood obtained from the models implemented in the R package **survival** (X axis) to that from the R package **straweib** (Y axis), when the data is simulated under the model implemented in the R package **straweib**

Example

We illustrate the R package **straweib** with data from a study on the timing of emergence of permanent teeth in Flemish children in Belgium (Leroy et al., 2003). The data analyzed were from the Signal-Tandmobiel project (Vanobbergen et al., 2000), a longitudinal oral health study in a sample of 4430 children conducted between 1996 and 2001. Dental examinations were conducted annually for a period of 6 years and tooth emergence was recorded based on visual inspection. As in Gomez et al. (2009), we will illustrate our R package by analyzing the timing of emergence of the permanent upper left first premolars. As dental exams were conducted annually, for each child, the timing of tooth emergence is known up to the interval from the last negative to the first positive dental examination.

```
data(tooth24)
head(tooth24)
```

```
  id left right sex dmf
1  1  2.7   3.5   1   1
2  2  2.4   3.4   0   1
3  3  4.5   5.5   1   0
4  4  5.9   Inf   1   0
5  5  4.1   5.0   1   1
6  6  3.7   4.5   0   1
```

The dataset is formatted to include 1 row per child. The variable denoted **id** corresponds to the ID of the child, **left** and **right** correspond to the left and right endpoints of the censoring interval in years, **sex** denotes the gender of the child (0 = boy, and 1 = girl), and **dmf** denotes the status of primary predecessor of the tooth (0 = sound, and 1 = decayed or missing due to caries or filled). Right censored observations are denoted by setting the variable **right** to "Inf".

In our analysis below, we use the function `icweib` in the package **straweib**, to fit a stratified Weibull regression model, where the variable **dmf** is the stratum indicator (S) and the variable **sex** is an explanatory variable (Z).

```
fit <- icweib(L = left, R = right, data = tooth24, strata = dmf, covariates = ~sex)
fit
```

```
Total observations used: 4386. Model Convergence: TRUE

Coefficients:
```

```
    coefficient     SE    z p.value
sex       0.331 0.0387 8.55       0

Weibull parameters - gamma(shape), lambda(scale):
 straname strata gamma   lambda
      dmf      0  5.99 1.63e-05
      dmf      1  4.85 1.76e-04

Test of proportional hazards for strata (H0: all strata's shape parameters are equal):
             test TestStat df  p.value
             Wald     44.2  1 2.96e-11
 Likelihood Ratio     44.2  1 3.00e-11

Loglik(model)=  -5501.781   Loglik(reduced)=  -5523.87
Loglik(null)=  -5538.309  Chisq= 73.05611   df= 1  p.value= 0
```

The likelihood ratio test of the PH assumption results in a p value of 3.00e-11, indicating that the PH model is not appropriate for this dataset. Or in other words, the data suggest that the hazard functions corresponding to the strata defined by $dmf = 0$ and $dmf = 1$ are not proportional. From the stratified Weibull regression model, the estimated regression coefficient for **sex** is 0.331, corresponding to a hazard ratio of 1.39 (95% CI: 1.29 - 1.50). In the output above, the maximized value of the log likelihood of the null model corresponds to the model stratified by covariate **dmf** but excluding the explanatory variable **sex**.

The p value from the Wald test of the null hypothesis of no effect of gender results in a p value of approximately 0 ($p < 10^{-16}$), which indicates that the timing of emergence of teeth is significantly different between girls and boys.

To test the global null hypothesis that both covariates **sex** and **dmf** are not associated with the outcome (time to teeth emergence), we obtain the log-likelihood for global null model, as shown below.

```
fit0 <- icweib(L = left, R = right, data = tooth24)
fit0

Total observations used: 4386. Model Convergence: TRUE

Weibull parameters - gamma(shape), lambda(scale):
 straname strata gamma   lambda
   strata    ALL   5.3 7.78e-05

Loglik(model)=  -5596.986
Loglik(null)=  -5596.986
```

The likelihood ratio test testing the global null hypothesis results in a test statistic $T = -2(l_R - l_F) = -2(-5596.986 + 5501.781) = 190.41$, which follows a χ^2_3 distribution under H_0, resulting in a p value of approximately 0 ($p < 10^{-16}$).

We illustrate the `HRatio` function in the **straweib** package to estimate the hazard ratio and corresponding 95% confidence intervals for comparing boys without tooth decay ($dmf = 0$) to boys with evidence of tooth decay ($dmf = 1$), where the hazard ratio is evaluated at various time points from 1 through 7 years.

```
HRatio(fit, times = 1:7, NumStra = 0, NumZ = 0, DemStra = 1, DemZ = 0)

  time NumStra DemStra beta*(Z1-Z2)        HR      low95    high95
1    1       0       1            0 0.1143698 0.06596383 0.1982972
2    2       0       1            0 0.2520248 0.18308361 0.3469262
3    3       0       1            0 0.4000946 0.33112219 0.4834339
4    4       0       1            0 0.5553610 0.49863912 0.6185351
5    5       0       1            0 0.7162080 0.66319999 0.7734529
6    6       0       1            0 0.8816470 0.79879884 0.9730878
7    7       0       1            0 1.0510048 0.91593721 1.2059899
```

The output indicates that the hazard ratio for boys comparing the stratum $dmf = 0$ to stratum $dmf = 1$ is small initially (e.g. 0.11 at 1 year) but tends to 1 in later years (e.g. 0.88 at 6 years and 1.05 at 7

years). Prior to 6 years, the hazard ratio is significantly less than 1, indicating that the timing of teeth emergence is delayed in children with tooth decay ($dmf = 1$) when compared to children without tooth decay ($dmf = 0$).

We illustrate estimation of the survival function in Figure 2 by plotting the survival functions and corresponding 95% point wise confidence intervals for girls ($Z = 1$), with and without tooth decay .

```
plot(fit, Z = 1, tRange = c(1, 7), xlab = "Time (years)", ylab = "Survival Function",
     main = "Estimated survival function for girls")
```

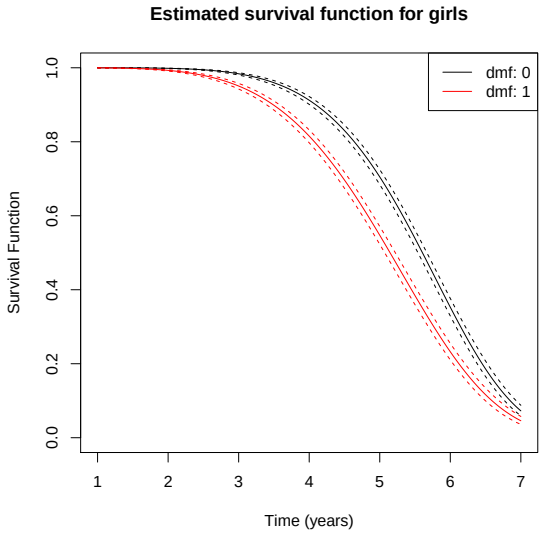

Figure 2: Estimated survival functions for girls, comparing the subgroup with sound primary predecessor of the tooth (dmf = 0) to the subgroup with unsound primary predecessor of the tooth (dmf = 1).

We compare our results from the **straweib** package to that obtained from the **survival** package.

```
library(survival)
tooth24.survreg <- tooth24
tooth24.survreg$right <- with(tooth24, ifelse(is.finite(right), right, NA))
fit1 <- survreg(Surv(left, right, type="interval2") ~ sex + strata(dmf) + factor(dmf),
                data = tooth24.survreg)
fit1

Call:
survreg(formula = Surv(left, right, type = "interval2") ~ sex +
    strata(dmf) + factor(dmf), data = tooth24.survreg)

Coefficients:
 (Intercept)          sex factor(dmf)1
  1.84389938  -0.06254599  -0.06491729

Scale:
dmf=Sound1 dmf=Sound2
 0.1659477  0.2072465

Loglik(model)= -5499.3   Loglik(intercept only)= -5576.2
        Chisq= 153.8 on 2 degrees of freedom, p= 0
n= 4386
```

The maximized value of the log-likelihood from the R package **survival** is -5499.3 (shown below), as compared to the maximized value of the log-likelihood of -5501.8 from the R package **straweib**.

To clarify the specific assumptions made by the models implemented in the **survival** and **straweib** packages, we carried out subgroup analyses in which we fit a Weibull PH model separately to each of the strata $dmf = 0$ and $dmf = 1$. The results from the Weibull PH model fit to the subgroup of children in the $dmf = 0$ stratum is shown below:

```
fit20 <- icweib(L= left, R=right, data=tooth24[tooth24$dmf==0, ], covariates = ~sex)
fit20 ### Partial results shown below
Coefficients:
    coefficient     SE    z  p.value
sex       0.448 0.0543 8.25 2.22e-16
```

The results from the Weibull PH model fit to the subgroup $dmf = 1$ is shown below:

```
fit21 <- icweib(L= left, R=right, data=tooth24[tooth24$dmf==1, ], covariates = ~sex)
fit21 ### Partial results shown below
Coefficients:
    coefficient     SE    z  p.value
sex       0.208 0.0554 3.76 0.000169
```

The model using the PH scale (implemented by **straweib** package) replaces the stratum specific hazard ratios for sex of $e^{0.448} = 1.57$ for the subgroup $dmf = 0$ and $e^{0.208} = 1.23$ for the subgroup $dmf = 1$ with a common value, $e^{0.331} = 1.39$.

Since the Weibull distribution has both the PH and accelerated failure time (AFT) property (Collett, 2003), the identical set of subgroup analyses can be fit using the **survival** package. Results from the fit using the **survival** package for the subgroup $dmf = 0$ are shown below:

```
fit20.survreg <- survreg(Surv(left, right, type="interval2") ~ sex,
                data = tooth24.survreg[tooth24.survreg$dmf==0, ])
fit20.survreg ### Partial results shown below
Coefficients:
(Intercept)         sex
 1.85029150 -0.07453785
```

Similar results using the **survival** package for the subgroup $dmf = 1$ are shown below:

```
fit21.survreg <- survreg(Surv(left, right, type="interval2") ~ sex,
                data = tooth24.survreg[tooth24.survreg$dmf==1, ])
fit21.survreg ### Partial results shown below
Coefficients:
(Intercept)         sex
 1.76931556 -0.04303767
```

In particular, the model assuming a common sex coefficient in the AFT scale (implemented by **survival** package) replaces the value of sex coefficient -0.075 for the subgroup with $dmf = 0$ and sex coefficient of -0.043 for the subgroup $dmf = 1$ with a shared common value, -0.063.

To assess the goodness of fit of the stratified Weibull model implemented by **straweib**, we created a multiple probability plot, as described in chapter 19 of Meeker and Escobar (1998). This diagnostic plot was created by splitting the dataset into 4 subgroups based on the values of **sex** and **dmf**. Within each group, we estimated the cumulative incidence at each visit time using a non-parametric procedure for interval censored data (Turnbull, 1976). The non-parametric estimates of cumulative incidence within each subgroup were compared to that obtained from the stratified Weibull model implemented by **straweib** package. We use the R package **interval** (Fay and Shaw, 2010) to obtain Turnbull's NPMLE estimates and the R package **straweib** for the estimates from the stratified Weibull model (code available upon request). Figure 3 shows the diagnostic plot.

Table 1 presents the estimates of hazard ratio for **sex**, within each of the strata defined by $dmf = 0$ and $dmf = 1$, comparing three different analyses - (1) Using the **survival** package to stratify on the variable **dmf** and including **sex** as an explanatory variable; (2) Using the **straweib** package to stratify on the variable **dmf** and including **sex** as an explanatory variable; (3) Fitting a Weibull PH model with **sex** as an explanatory variable, separately within each of the two subgroups defined by $dmf = 0$ and $dmf = 1$.

```
HR.straweib <- exp(fit$coef[1, 1])
HR.survreg <- exp(-fit1$coefficients['sex']/fit1$scale)
HR.subgroup <- exp(c(fit20$coef[1, 1], fit21$coef[1, 1]))
```

Concluding remarks

We have developed and illustrated an R package **straweib** for the analysis of interval-censored outcomes, based on a stratified Weibull regression model. The proposed model shares similarities with the semi-parametric stratified Cox model. We illustrated the R package **straweib** using data from

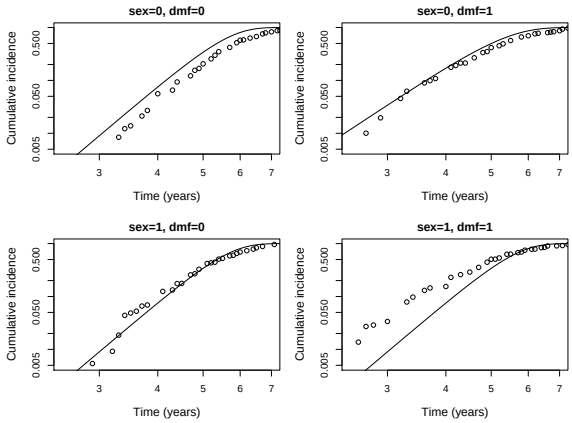

Figure 3: Comparing non-parametric (points) and Weibull model (lines) based estimates of cumulative incidence within each group based on covariates **sex** and **dmf**.

Table 1: Hazard ratio estimates for gender, comparing the models implemented in the R packages **survival**, **straweib** and subgroup analyses

stratum	Package **survival**	Package **straweib**	Stratum specific subgroup analyses
dmf = 0	1.46	1.39	1.56
dmf = 1	1.35	1.39	1.23

a prospective study on the timing of emergence of permanent teeth in Flemish children in Belgium (Leroy et al., 2003).

Although the models and R package are illustrated for the analysis of interval-censored time-to-event outcomes, the methods proposed here are equally applicable for the analysis of right-censored outcomes. The syntax for the analysis of right-censored observations is explained in the manual accompanying the **straweib** package available on CRAN (Gu and Balasubramanian, 2013).

Acknowledgements

This research was supported by NICHD grant R21 HD072792.

Bibliography

D. Collett. *Modelling Survival Data in Medical Research, Second Edition*. Texts in statistical science. Taylor & Francis, 2003. ISBN 9781584883258. [p32, 33, 38]

D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall, 1979. [p33]

D. T. Dunn, R. J. Simonds, M. Bulterys, L. A. Kalish, J. Moye, A. de Maria, C. Kind, C. Rudin, E. Denamur, A. Krivine, C. Loveday, and M. L. Newell. Interventions to prevent vertical transmission of HIV-1: effect on viral detection rate in early infant samples. *AIDS*, 14(10):1421–1428, 2000. [p31]

M. P. Fay and P. A. Shaw. Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. *Journal of Statistical Software*, 36(2):1–34, 2010. URL http://www.jstatsoft.org/v36/i02/. [p38]

D. M. Finkelstein. A proportional hazards model for interval-censored failure time data. *Biometrics*, 42 (4):845–854, 1986. [p31]

E. Goetghebeur and L. Ryan. Semiparametric regression analysis of interval-censored data. *Biometrics*, 56(4):1139–1144, 2000. [p31]

G. Gomez, M. L. Calle, R. Oller, and K. Langohr. Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling*, 9(4):259–297, 2009. [p31, 35]

X. Gu and R. Balasubramanian. *straweib: Stratified Weibull Regression Model*, 2013. URL http://CRAN.R-project.org/package=straweib. R package version 1.0. [p31, 33, 39]

R. Leroy, K. Bogaerts, E. Lesaffre, and D. Declerck. The emergence of permanent teeth in flemish children. *Community Dentistry and Oral Epidemiology*, 31(1):30–39, 2003. [p31, 35, 39]

J. C. Lindsey and L. M. Ryan. Tutorial in biostatistics - methods for interval-censored data. *Statistics in Medicine*, 17(2):219–238, 1998. [p31]

W. Meeker and L. Escobar. *Statistical Methods for Reliability Data*. Wiley Series in Probability and Statistics. Wiley, 1998. ISBN 9780471673279. [p38]

T. Therneau. *A Package for Survival Analysis in S*, 2012. R package version 2.36-14. [p31, 32]

B. W. Turnbull. Empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B-Methodological*, 38(3):290–295, 1976. [p31, 38]

J. Vanobbergen, L. Martens, E. Lesaffre, and D. Declerck. The Signal-Tandmobiel project a longitudinal intervention health promotion study in flanders (belgium): baseline and first year results. *Eur J Paediatr Dent*, 2:87–96, 2000. [p35]

Xiangdong Gu
Division of Biostatistics and Epidemiology
University of Massachusetts, Amherst, MA, USA
xdgu@schoolph.umass.edu

David Shapiro
Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
shapiro@sdac.harvard.edu

Michael D. Hughes
Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
mhughes@hsph.harvard.edu

Raji Balasubramanian
Division of Biostatistics and Epidemiology
University of Massachusetts, Amherst, MA, USA
rbalasub@schoolph.umass.edu