Journal of Statistical Software

August 2015, Volume 66, Issue 8. *http://www.jstatsoft.org/*

The R Package threg to Implement Threshold Regression Models

Tao Xiao
Shenzhen University

G. A. Whitmore
McGill University

Xin He
University of Maryland

Mei-Ling Ting Lee
University of Maryland

Abstract

This paper introduces the R package **threg**, which implements the estimation procedure of a threshold regression model, which is based on the first-hitting-time of a boundary by the sample path of a Wiener diffusion process. The threshold regression methodology is well suited to applications involving survival and time-to-event data, and serves as an important alternative to the Cox proportional hazards model.

This new package includes four functions: `threg`, and the methods `hr`, `predict` and `plot` for ‘`threg`’ objects returned by `threg`. The `threg` function is the model-fitting function which is used to calculate regression coefficient estimates, asymptotic standard errors and p values. The `hr` method for ‘`threg`’ objects is the hazard-ratio calculation function which provides the estimates of hazard ratios at selected time points for specified scenarios (based on given categories or value settings of covariates). The `predict` method for ‘`threg` objects is used for prediction. And the `plot` method for ‘`threg`’ objects provides plots for curves of estimated hazard functions, survival functions and probability density functions of the first-hitting-time; function curves corresponding to different scenarios can be overlaid in the same plot for comparison to give additional research insights.

Keywords: first-hitting-time, hazard ratios, survival analysis, **threg**, threshold regression, Wiener diffusion process, R.

1. Introduction

Threshold regression is a statistical methodology to analyze time-to-event or survival data, taking covariates into account. Before we discuss threshold regression, we briefly introduce the characteristics of survival data. Survival data record the time until the occurrence of a key event such as death of a patient, occurrence of tumor, discharge from hospital, or cessation of breast feeding. Survival data often contain censored observations for which we are not able to observe the exact event time. For example, in a study of risk factors for lung cancer, the time of diagnosis of lung cancer is to be recorded for each person recruited. At the end of study follow-up, however, some people may not have been diagnosed with lung cancer. Those people might or might not develop lung cancer later, but that is unknown at the time when the data need to be analyzed. The incompletely observed event times encountered in survival data make their analysis different from other classical methods used for completely observed data, such as ordinary linear regression. Survival and time-to-event data arise in many applied fields in addition to medicine and public health, including engineering, sociology, demography, and economics to name a few.

Among the many statistical models used for survival data, the most widely employed is the Cox proportional hazards model (Cox 1972) which superimposes a regression structure for covariates on top of a baseline hazard function that has an arbitrary form. As its name implies, the Cox proportional hazards model assumes that covariates alter the baseline hazard function in a proportional manner. Hence it is crucial to check if the proportional hazards assumption is satisfied before using the Cox model. When the proportional hazards assumption is violated, the model should not be used or at least used with caution.

Threshold regression methodology is a powerful tool to analyze survival data. Threshold regression does not assume proportional hazards and can serve in place of Cox proportional hazards regression, when the proportional hazards assumption does not hold. The defining feature of threshold regression is that the event time is defined as the first time an underlying stochastic process hits a boundary threshold. In a medical context, for example, the event of interest might be death and the time of death is the moment when the patient's latent health status first reaches a boundary at zero.

The R (R Core Team 2015) package **threg** (Xiao 2015), which is presented here, encompasses the most important threshold regression model developed to date and the most widely applied, namely, the Wiener threshold regression model. This model applies to settings where the underlying health process follows a Wiener diffusion process and the failure event is triggered when the process hits a fixed threshold for the first time. The first hitting time in this situation has an inverse Gaussian distribution, which has a very tractable mathematical form. Threshold regression in its general formulation covers a broad collection of models that extend well beyond first hitting times for Wiener diffusion processes. Different families of threshold regression models are created by assuming different types of stochastic processes and different boundaries or thresholds. For example, one can have gamma processes, curvilinear boundaries, Markov chains with absorbing states, and many more. Lee and Whitmore (2006) present a wide-ranging review of the general methodology and types of applications.

The connection between threshold regression and Cox proportional hazards regression is studied in depth in Lee and Whitmore (2010). They show that selected threshold regression models can be constructed that have the proportional hazards property. They also remark, however, that the proportional hazards property is not valid for many real applications even though it is frequently assumed for analytical convenience. The Wiener diffusion model covered by our **threg** package does not require the proportional hazards property. The hazard function of its first hitting time distribution can take a variety of shapes that are commonly encountered in practical work. Lee and Whitmore (2010) discuss a range of advantages that threshold regression has as a modeling approach. These advantages include more realism in describing real world phenomena because it takes explicit account of the underlying health process and has a realistic mechanism for the time-to-event, namely, that of a first hitting time of a threshold or boundary. The richer model structure of threshold regression also tends to offer deeper insights into the data being analyzed. For example, the Wiener threshold regression model can have separate regression functions for a patient's initial health level (baseline health) and for the mean parameter of a patient's health trajectory (the health trend), which together give the model a capacity to capture the multifaceted impact of covariates on a patient's health experience.

Currently there exists the `coxph` function in package **survival** (Therneau and Grambsch 2000; Therneau 2012) in R for implementing the Cox proportional hazards regression, but there is no R package yet to implement the threshold regression, which, as stated above, serves as an important alternative for the Cox proportional hazards regression. This paper presents the R package **threg** (version: 1.0.3) for implementing Wiener threshold regression. This package was created in version 3.0.3 of R, and depends on the following two packages: **survival** (version: 2.36.14) and **Formula** (version: 1.1.0). It is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=threg`.

The paper is organized as follows. In Section 2 we introduce an example dataset from a leukemia study, for which the proportional hazards assumption is not reasonable. We briefly compare the performance of Wiener threshold regression and Cox proportional hazards regression for this dataset. In Section 3, we present a theoretical overview of threshold regression in the Wiener diffusion case. In Section 4 we describe features of the `threg` function and the `hr`, `predict` and `plot` methods for '`threg`' objects returned by function `threg` included in the **threg** package that implements this threshold regression model and also give corresponding examples using a bone marrow transplantation dataset. Conclusions are presented in Section 5.

2. An example with non-proportional hazards

In this section we give a quick comparison between the threshold regression and the Cox proportional hazards regression with `lkr`, a leukemia remission study dataset (Garrett 1997). There are 42 patients in the dataset, who were monitored for whether they relapsed (`relapse`: 1 = yes, 0 = no) and for how long (in weeks) they remained in remission (`weeks`). Two different treatments were given to these patients. For the first treatment, 21 patients received a new experimental drug (drug A), and the other 21 received a standard drug (`treatment1`: 1 = drug A, 0 = standard). For the second treatment, a different drug (drug B) was given to 20 patients, while the remaining 22 patients received a standard drug (`treatment2`: 1 = drug B, 0 = standard). The dataset also records white blood cell count, which is a strong indicator of the presence of leukemia and is categorized into three levels (`wbc3cat`: 1 = normal, 2 = moderate, 3 = high).

The variable `treatment2` violates the proportional hazards assumption. Next we focus on this variable to demonstrate the advantage of the threshold regression model over the Cox model when the proportional hazards assumption is violated.

Firstly Kaplan-Meier non-parametric estimates of the survival curves for the two levels of `treatment2` are given in Figure 1.

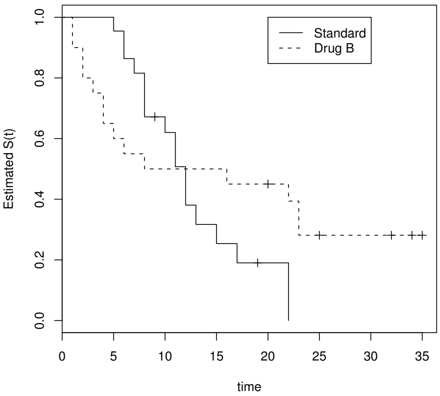

Figure 1: Kaplan-Meier plot by the `treatment2` variable for the leukemia data.

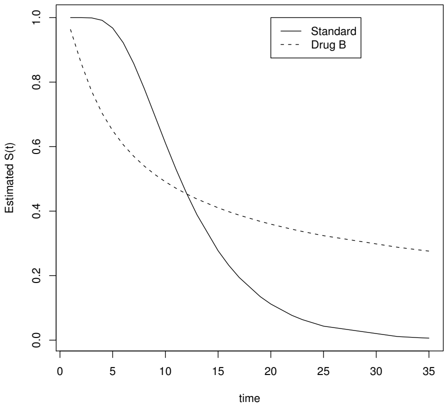

Figure 2: Threshold regression predicted survival functions for the leukemia data.

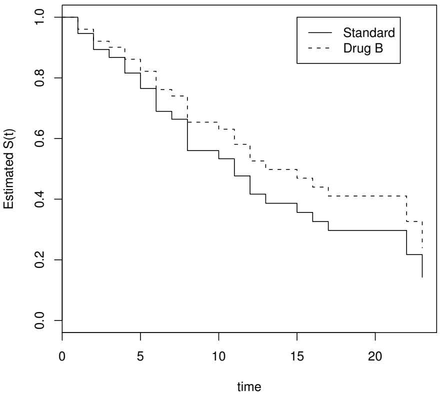

Figure 3: Cox predicted survival functions for the leukemia data.

We can clearly see that the two survival curves in Figure 1 cross each other. Before a time point around week 12, the estimated survival probability (probability of remaining in remission) of the `drug B` group is lower than that of the `standard drug` group. After that point, however, the estimated survival probability of the `drug B` group is higher than the `standard drug` group. The crossing survival curves suggest that the proportional hazards assumption, which is the key assumption for the Cox model, does not hold.

The threshold regression model based on the Wiener process, however, does not assume proportional hazards. We can fit the threshold regression model on the `treatment2` variable in the `lkr` dataset contained in the **threg** package by using the `threg` function:

```
R> library("threg")
R> data("lkr", package = "threg")
R> lkr$f.treatment2 <- factor(lkr$treatment2)
R> fit <- threg(Surv(weeks, relapse) ~ f.treatment2 | f.treatment2,
+    data = lkr)
R> fit
```

```
Call:
threg(formula = Surv(weeks, relapse) ~ f.treatment2 | f.treatment2,
    data = lkr)

                          coef  se(coef)         z       p
lny0: (Intercept)    2.0097844 0.1705141 11.786620 0.0e+00
lny0: f.treatment21 -1.2739233 0.2441634 -5.217504 1.8e-07
  mu: (Intercept)   -0.5886165 0.1340127 -4.392246 1.1e-05
  mu: f.treatment21  0.5888365 0.1535081  3.835866 1.3e-04

Log likelihood =-104.64, AIC =217.28
```

Then we can plot the predicted survival curves for the two levels of the `treatment2` variable by using the `plot` method for '`threg`' objects returned by `threg`:

```
R> plot(fit, var = f.treatment2, graph = "sv", nolegend = 1, nocolor = 1)
R> legend(20, 1, c("Standard", "Drug B"), lty = 1:2)
```

The predicted survival curves generated by the codes above are given in Figure 2. Here we notice that the crossing of the predicted survival curves of the two levels of `treatment2` in the Kaplan-Meier plot (Figure 1) is well captured by the threshold regression model, as illustrated in Figure 2. However, if the Cox model is erroneously used, the predicted survival curves of the two groups shown in Figure 3 fail to capture the crossing pattern.

Using the `hr` method for '`threg`' objects to calculate the hazard ratios at different time points illustrates the advantages of the threshold regression model over the Cox model when the proportional hazards assumption is violated.

The following use of the `hr` method for '`threg`' objects calculates the hazard ratio of the `drug B` group vs. the `standard drug` group at week 5. This hazard ratio is calculated as 2.08.

```
R> hr(fit, var = f.treatment2, timevalue = 5)

     timevalue f.treatment21
[1,]         5      2.075739
```

Similarly we can use the `hr` method for '`threg`' objects again to calculate the corresponding hazard ratio at week 20. This hazard ratio is calculated as 0.12.

It is clear that the hazard ratios calculated at week 5 and week 20 are quite different: at week 5, the hazard ratios of the `drug B` group versus the `standard drug` group is about 2.08; while at week 20, this hazard ratio reverses and sharply decreases to about 0.12. Note that the `standard drug` group is the reference group. Obviously the change of hazard ratio over time is huge and the change has been well captured by using the `threg` function. The graph of the estimated hazard functions of the two treatment groups are also generated by the `threg` function above by using the `graph = "hz"` argument. This graph is given in Figure 4, from which we can see that the hazard function curves of the two treatment groups cross over time, and that is why the estimated hazard ratio of the `drug B` group versus the `standard drug` group changes from a number greater than 1 (2.08) to a number less than 1 (0.12).

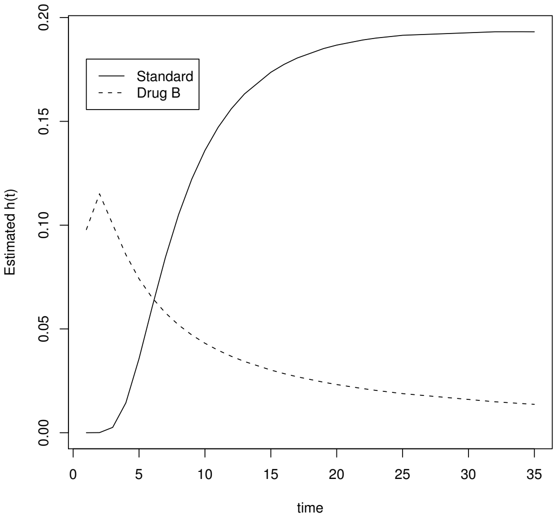

Figure 4: Threshold regression predicted hazard functions for the leukemia data.

On the other hand, we can only estimate a constant hazard ratio from the Cox model (i.e., 0.73), across the whole time span for the `drug B` group versus the `standard drug` group, due to the proportional hazards assumption. Clearly, this outcome is misleading. This constant hazard ratio by the Cox model can be obtained by the `coxph` function of the **survival** package in R as follows:

```
R> library("survival")
R> summary(coxph(Surv(weeks, relapse) ~ treatment2, data = lkr))
```

3. Theory of threshold regression

In this section we will introduce the genesis of the threshold regression methodology briefly. We assume that the latent stochastic process is a Wiener process $Y(t)$ starting at y_0 with drift μ and variance σ^2, and has the following properties:

1. $Y(t)$ has independent increments; for any non-overlapping time intervals (t_1, t_2), (t_3, t_4), $Y(t_2) - Y(t_1)$ and $Y(t_4) - Y(t_3)$ are independent.

2. $Y(t_2) - Y(t_1)$ is normally distributed with mean $\mu(t_2 - t_1)$ and variance $\sigma^2(t_2 - t_1)$ with $t_1 < t_2$.

When we regard this Wiener process as the latent health status process, we can let $Y(0) = y_0 > 0$ be the initial health status, and define T as the first time a sample path of the health status process reaches 0 level, i.e., $T = \inf\{t : Y(t) = 0\}$. T is considered as the event time in survival data analysis by the threshold regression. Figure 5 is an illustration of a sample path of such a latent health status process.

By using either the backward or forward diffusion equations subject to the initial condition and the boundary condition for the absorbing barrier (see Cox and Miller 1965), it can be

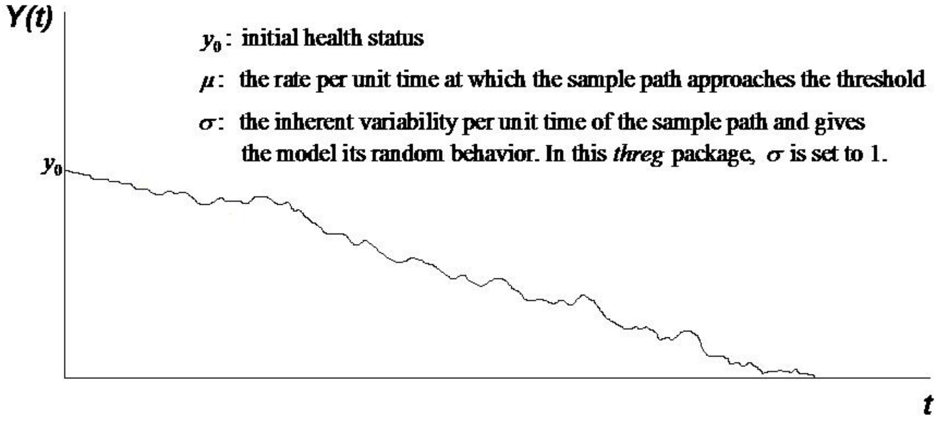
Y(t)
y_0: initial health status
μ: the rate per unit time at which the sample path approaches the threshold
σ: the inherent variability per unit time of the sample path and gives the model its random behavior. In this *threg* package, σ is set to 1.
y_0
t

Figure 5: Wiener process as the latent health status process.

derived that T follows the inverse Gaussian distribution with the following probability density function (p.d.f.).

$$f(t|\mu,\sigma^2,y_0) = \frac{y_0}{\sqrt{2\pi\sigma^2 t^3}} \exp\left[-\frac{(y_0+\mu t)^2}{2\sigma^2 t}\right], \tag{1}$$

where $\sigma^2 > 0, y_0 > 0$ and $-\infty < \mu < \infty$. The p.d.f. is proper if $\mu \leq 0$. The cumulative distribution function (c.d.f.) of the first-hitting-time (FHT) is:

$$F(t|\mu,\sigma^2,y_0) = \Phi\left[-\frac{(y_0+\mu t)}{\sqrt{\sigma^2 t}}\right] + \exp\left(-\frac{2y_0\mu}{\sigma^2}\right)\Phi\left[\frac{\mu t - y_0}{\sqrt{\sigma^2 t}}\right], \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Note that if $\mu > 0$, the Wiener process may never hit the boundary at zero and hence there is a probability that the FHT is ∞, with $P(\text{FHT} = \infty) = 1 - \exp(-2y_0\mu/\sigma^2)$ (Cox and Miller 1965).

By carefully examining Equation 1 and Equation 2, it can be seen that both $f(t|\mu,\sigma^2,y_0)$ and $F(t|\mu,\sigma^2,y_0)$ actually depend on y_0/σ and μ/σ only. Hence we need to fix one of three parameters (μ, y_0, σ) to avoid over-parameterization. Because the degradation process is latent with undefined measurement scale, we choose to set the variance parameter $\sigma^2 = 1$ in the **threg** package without loss of generality. Then we can regress the other two process parameters, y_0 and μ, on the covariate data.

Suppose that the covariate vector is $\boldsymbol{X}^\top = (1, X_1, \ldots, X_k)$, where $X_1, \ldots, X_k$ are covariates and the leading 1 in $\boldsymbol{X}^\top$ allows for a constant term in the regression relationship. We assume that μ and $\ln(y_0)$ are linear in regression coefficients. Then $\ln(y_0)$ and μ can be linked to the covariates with the following regression forms:

$$\ln(y_0) = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_k X_k = \boldsymbol{X}^\top\boldsymbol{\gamma}, \tag{3}$$

$$\mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = \boldsymbol{X}^\top\boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_k)^\top$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^\top$ are regression coefficient vectors. If some covariates are regarded as unimportant in predicting $\ln(y_0)$ or μ, then these covariates can be removed from the regression model by setting the corresponding elements in $\boldsymbol{\gamma}$ or $\boldsymbol{\beta}$ to zero.

For example, if covariate X_1 in $\boldsymbol{X}^\top$ is not important to predict $\ln(y_0)$, we can set γ_1 to zero in Equation 3.

4. The uses of functions provided by the threg package

In this section, we detail how to apply the threshold regression in survival data analysis by using the functions provided by the **threg** package. Examples are given at the end of each subsection, and for all the examples we use the bone marrow transplantation dataset (Klein and Moeschberger 2003). This dataset contains 137 acute leukemia patients treated with bone marrow transplants which are a standard treatment for acute leukemia. Recovery following bone marrow transplantation may depend on many risk factors known at the time of transplantation. In our examples below, we simply choose three risk factors, which are previously reported as important, to illustrate how to use the four functions provided by the **threg** package. These three risk factors are:

1. Patient age (the `recipient_age` variable).

2. Risk categories based on their status at the time of transplantation (the `group` variable). These categories are: acute lymphoblastic leukemia (coded 1 for `group`), low-risk acute myelocytic leukemia (coded 2 for `group`), and high-risk acute myelocytic leukemia (coded 3 for `group`).

3. French-American-British (FAB) classification based on standard morphological criteria (the `fab` variable): coded 1 if the patient has acute lymphoblastic leukemia with an FAB classification of M4 or M5, and 0 otherwise. The former patients were considered to have a possible elevated risk of relapse or treatment-related death.

The study time (time to relapse, death or end of study) variable is `time` (in days) and the censoring indicator variable is `indicator` (coded 1 if dead or relapse, and 0 otherwise). The maximum follow-up for these 137 patients was 7 years.

Next we detail how to use the `threg` function and the `hr`, `plot` and `predict` methods for '`threg`' objects with their arguments.

4.1. Regression coefficient estimation: `threg`

Maximum likelihood estimation is used to estimate the regression coefficients in Equation 3 and Equation 4. A subject i in the sample dataset who has an exact death time observed provides the information on the probability that the event is occurring at this time, and therefore contributes the FHT probability density $f(t^{(i)}|\mu^{(i)}, y_0^{(i)})$ to the sample likelihood function, where $t^{(i)}$ is the observed time of death. A subject j in the sample dataset who lives to the end of the study provides a right censored observation, and all we know about this subject is that the event time is larger than the on study time. Therefore the information contributed by a surviving subject in the sample likelihood function is the survival function evaluated at the corresponding on study time $t^{(j)}$ for this subject: $1-F(t^{(j)}|\mu^{(j)}, y_0^{(j)})$. Among the n subjects in the sample, subjects with observed death times are indexed from 1 to n_1 and subjects with right censored observations are indexed from $n_1 + 1$ to n. Then, the log-likelihood function is

$$\ln L(\beta,\gamma) = \sum_{i=1}^{n_1} \ln f(t^{(i)}|\mu^{(i)}, y_0^{(i)}) + \sum_{j=n_1+1}^{n} \ln \left[1 - F(t^{(j)}|\mu^{(j)}, y_0^{(j)})\right] \tag{5}$$

The `threg` function is used to estimate the regression coefficients of the threshold regression model. Two arguments of the `threg` function are used to specify the inputs of the sample log-likelihood function in Equation 5:

- `formula`: A '`Formula`' object (see Zeileis and Croissant 2010), with the response on the left of a `~` operator, and the independent variables on the right. The response must be a '`Surv`' object as returned by the `Surv` function from package **survival**. On the right of the `~` operator, a `|` operator must be used: on the left of the `|` operator, users specify independent variables that will be used in the linear regression function for $\ln(y_0)$ in the threshold regression model; on the right of the `|` operator, users specify independent variables that will be used in the linear regression function for μ in the threshold regression model. If users just want to use a constant $\ln(y_0)$ or μ, they can put 0 or 1 as a placeholder on the left or right of the `|` operator, instead of listing the independent variables for $\ln(y_0)$ or μ.

- `data`: Specifies input dataset. Such a dataset must be a survival dataset including at least the survival time variable and censoring variable. For the censoring variable, 1 should be used to indicate the subjects with failure observed, and 0 should be used to indicate the subjects that are right censored. The dataset can also include other independent variables that will be used in the threshold regression model.

The optimization function `nlm` from the **stats** package of the base distribution of R is incorporated in the `threg` command to find the minimum of the minus log-likelihood function in Equation 5 and obtain the maximum likelihood estimates of the regression coefficients in vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in the threshold regression model. Note that `nlm` is based on the recurrence relation of the Newton-Raphson method:

$$\theta_{i+1} = \theta_i - \left[\left. \frac{\partial^2 h(\theta)}{\partial\theta\partial\theta^\top} \right|_{\theta=\theta_i} \right]^{-1} \left. \frac{\partial h(\theta)}{\partial\theta} \right|_{\theta=\theta_i}, \tag{6}$$

where the matrix of the second derivatives is called the *Hessian* matrix and the vector of the first derivatives is called the *gradient*. The convergence speed of `nlm` is fairly fast for log-likelihood function of the threshold regression model in Equation 5. Note that some warning messages passed from the `nlm` function may appear along with the outputs of the `threg` function when the *Hessian* matrix is not invertible at the values of some starting points of the Newton-Raphson procedure, but these warning messages can be ignored.

Below we use the `threg` function introduced above to fit a threshold regression model on the bone marrow transplantation dataset. We use `recipient_age` and `fab` as the predictors for $\ln(y_0)$ and `group` and `fab` as the predictors for μ. Note that `group` and `fab` are transformed to factor variables, `f.group` and `f.fab`, since they are categorical variables. From the outputs we can see that all of these three predictors are significantly important.

```
R> library("threg")
R> data("bmt", package = "threg")
R> bmt$f.group <- factor(bmt$group)
R> bmt$f.fab <- factor(bmt$fab)
R> fit <- threg(Surv(time, indicator) ~ recipient_age + f.fab | f.group +
+    f.fab, data = bmt)
R> fit

Call:
threg(formula = Surv(time, indicator) ~ recipient_age + f.fab |
    f.group + f.fab, data = bmt)

                           coef    se(coef)          z       p
lny0: (Intercept)    3.09629902 0.276452140 11.2001268 0.0e+00
lny0: recipient_age -0.03196957 0.007711182 -4.1458717 3.4e-05
lny0: f.fab1        -0.42065450 0.177761812 -2.3663941 1.8e-02
  mu: (Intercept)    0.01493740 0.009118815  1.6380856 1.0e-01
  mu: f.group2       0.02679506 0.011619710  2.3060005 2.1e-02
  mu: f.group3       0.01281571 0.013266828  0.9659962 3.3e-01
  mu: f.fab1        -0.02396574 0.010601310 -2.2606397 2.4e-02

Log likelihood =-674.64, AIC =1363.28
```

4.2. Hazard ratio calculation: `hr` method

Suppose that we have used the `threg` function with predictor variables $\{X_1, \ldots, X_{k-1}, G\}$ to predict $\ln(y_0)$ and μ in the *threshold regression* model, where G is a categorical variable, then we can use `hr` to estimate the hazard ratio of level $G = g$ over level $G = 0$ for a scenario in which the values of the other predictors and the time are given: $X_1 = X_1, \ldots, X_{k-1} = X_{k-1}$, and $t = t_0$. Such a calculation is sorted out as follows.

Set $(\boldsymbol{z}^g)^\top = (1, X_1, \ldots, X_{k-1}, g)$, then by using the `threg` function, y_0 and μ can be estimated for the given $(\boldsymbol{z}^g)^\top$ as: $\hat{y}_0^g = \exp\{(\boldsymbol{x}^g)^\top \hat{\boldsymbol{\gamma}}\}$ and $\hat{\mu}^g = (\boldsymbol{x}^g)^\top \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ are the vectors of the estimated values for the regression coefficients (see Section 4.1). Now we can calculate the estimates of the density and survival functions at time t_0 as:

$$f(t_0|\hat{\mu}^g, \hat{y}_0^g), \tag{7}$$

$$S(t_0|\hat{\mu}^g, \hat{y}_0^g) = 1 - F(t_0|\hat{\mu}^g, \hat{y}_0^g). \tag{8}$$

And further, the estimate of the hazard function at time t_0 is calculated as:

$$h(t_0|\hat{\mu}^g, \hat{y}_0^g) = \frac{f(t_0|\hat{\mu}^g, \hat{y}_0^g)}{S(t_0|\hat{\mu}^g, \hat{y}_0^g)} = \frac{f(t_0|\hat{\mu}^g, \hat{y}_0^g)}{1 - F(t_0|\hat{\mu}^g, \hat{y}_0^g)}, \tag{9}$$

where $f(\cdot)$ and $F(\cdot)$ are given in Equation 1 and Equation 2, with σ^2 set to 1, and with y_0 and μ replaced by their estimates in level g. Similarly, if we change the non-reference level g to the reference level 0, we can obtain $f(t_0|\hat{\mu}^0, \hat{y}_0^0)$, $S(t_0|\hat{\mu}^0, \hat{y}_0^0)$ and $h(t_0|\hat{\mu}^0, \hat{y}_0^0)$. The hazard ratio of level $G = g$ over level $G = 0$ at $X_1 = X_1, \ldots, X_{k-1} = X_{k-1}$ and time $t = t_0$ is therefore:

$$\text{Hazard Ratio} = \frac{h(t_0|\hat{\mu}^g, \hat{y}_0^g)}{h(t_0|\hat{\mu}^0, \hat{y}_0^0)}. \tag{10}$$

Using the formulas above, the hr method for 'threg' objects estimates the hazard ratios for a categorical variable with three arguments: var, scenario and timevalue. The uses of these three arguments are given as follows:

- object: An object of class 'threg', returned by the threg function.
- var: Specifies the name of the categorical variable G for which the hazard ratios are to be calculated. Note that the categorical variable G specified for the var argument needs to be a factor variable in R. If G is not already a factor variable yet, you need to use the factor function in R to transform G into a factor variable first and then include this factor variable in threg when fitting the threshold regression model. Then you can specify this factor variable in the var argument of the hr function to calculate the hazard ratios for G.
- timevalue: Specifies the desired time value at which the hazard ratios are to be calculated. A vector is allowed for this argument.
- scenario: Specifies the values of all predictors except G. A setting of these values is referred as a *scenario*. The calculated hazard ratios are with reference to the specified scenario. We do not need to specify a level value for the categorical G in the scenario argument since all non-reference levels g of G are enumerated in calculating hazard ratios relative to the reference level. The reference level is chosen as the lowest level of the factor variable specified in the var argument, and all other levels of this factor variable are non-reference levels. Note that in the scenario argument, the order of presentation of the predictors does not matter, and the terms in this argument are separated by + signs. If G is the only predictor in the model, then the scenario argument is not needed.

Below we use the hr method for 'threg' objects to calculate the hazard ratios for the factorized group variable, f.group, for the specified scenario that "the patient age is 18 years old and the FAB classification is 0" at time "500 days". Note that the factor variable f.group has already been included in the threshold regression model earlier by threg which returned the object fit.

```
R> hr(fit, var = f.group, timevalue = 500,
+     scenario = recipient_age(18) + f.fab1(0))

     timevalue  f.group2  f.group3
[1,]       500 0.3691302 0.6379671
```

4.3. Curves of estimated survival, hazard, density functions: plot method

The plot method for 'threg' objects can be used to display the graphs of the estimated survival, hazard or density functions at different levels of a factor predictor variable which has been included in the threshold regression by threg.

From Equations 7–9, the estimates of the density, survival and hazard functions in level g at a given time t_0 can be calculated. If we replace t_0 with t which varies over the range of all the observed time points in the data, Equations 7–9 become functions of t in level g and those functions curves can be plotted. When we overlay the curves of different levels of G in one plot, we can compare the corresponding function estimates at different levels of G. These kinds of plots can give additional research insights. The `plot` method for '`threg`' objects generates these plots with the following arguments:

- `x`: An object of class '`threg`', returned by the `threg` function.
- `var`: Specifies the name of the categorical variable G, for each level of which the plots would be generated at given scenario specified by the `scenario` argument. The use of the `var` argument is the same as that in the `hr` method.
- `scenario`: Specifies the values of all predictors except G. The use of the `scenario` argument is the same as that in the `hr` method.
- `graph`: Specifies the type of curves to be generated. The `"hz"` option is to plot hazard function curves, the `"sv"` option is to plot survival function curves, and the `"ds"` option is to plot density function curves.
- `nolegend`: The `nolegend` argument needs to be set to 1 if users do not want the **threg** package to generate legends for the plot. Note that even if `nolegend` is set to 1, users can still add legends by themselves after the graph is generated, by using the `legend` function in R.
- `nocolor`: The `nocolor` argument needs to be set to 1 if users want to depict all curves in black.

Below are examples of using the `plot` method of '`threg`' objects to overlay curves of survival, hazard and probability density functions corresponding to different levels of `f.group`. The specified scenario is still that "the patient age is 18 years old and the FAB classification is 0".

Survival functions (Figure 6):

```
R> plot(fit, var = f.group, scenario = recipient_age(18) + f.fab1(0),
+    graph = "sv", nocolor = 1)
```

Hazard functions (Figure 7):

```
R> plot(fit, var = f.group, scenario = recipient_age(18) + f.fab1(0),
+    graph = "hz", nocolor = 1)
```

Probability density functions (Figure 8):

```
R> plot(fit, var = f.group, scenario = recipient_age(18) + f.fab1(0),
+    graph = "ds", nocolor = 1)
```

4.4. Predictions: `predict` method

The `predict` method for '`threg`' objects can be used to predict the initial health status value y_0, the drift value of the health process μ, the probability density function of the survival

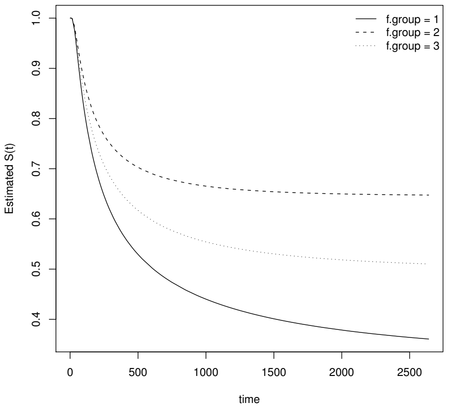

Figure 6: Estimated survival functions by the threshold regression for the bone marrow transplantation data.

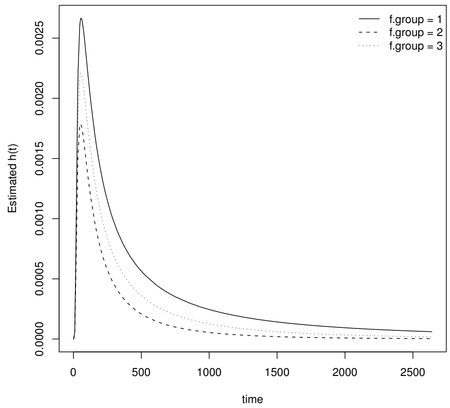

Figure 7: Estimated hazard functions by the threshold regression for the bone marrow transplantation data.

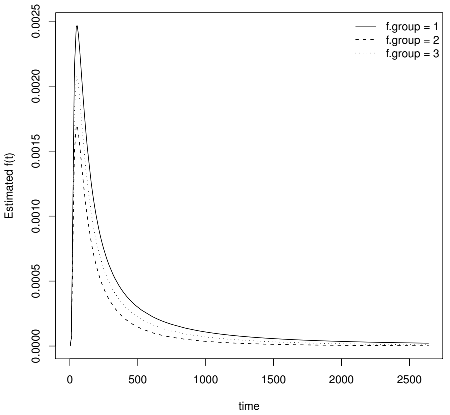

Figure 8: Estimated probability density functions by the threshold regression for the bone marrow transplantation data.

time $f(t \mid \mu, y_0)$, the survival function $S(t \mid \mu, y_0)$ and the hazard function $h(t \mid \mu, y_0)$ for a specified scenario and time value. The scenario specified here is similar to those in the `hr` and `plot` methods. The only difference is that we need to provide the scenario values for the dummy variables expanded from the factor variable G when using the `predict` method while we do not need to provide these values when using `hr` and `plot`, since the program will automatically enumerate all levels of G for `hr` and `plot`. The `predict` method has three arguments:

- `object`: An object of class '`threg`', returned by the `threg` function.
- `timevalue`: Specifies the desired time value at which the predicted values are to be calculated. A vector is allowed for this argument. If this argument is omitted, then the predicted values for the study time of all subjects would be calculated.
- `scenario`: Specifies the values of all predictors, including the dummy variables expanded

for the factor variable G. If this argument is omitted, then the predicted values at a specified time value for all subjects would be calculated, and in this case the covariate values for each subject are used as their corresponding scenario values.

Below we use the `predict` method for ‘`threg`’ objects to calculate the predicted values for the specified scenario that “the patient age is 18 years old, the FAB classification is 0 and the risk category is 3” at time “2000 days”.

```
R> predict(fit, timevalue = 2000,
+    scenario = recipient_age(18) + f.fab1(0) + f.group2(0) + f.group3(1))

     timevalue       y0         mu            f         S            h
[1,]      2000 12.43912 0.02775311 1.749514e-05 0.5184183 3.374714e-05
```

5. Conclusion

This paper introduces the use of the **threg** package to implement threshold regression in R. As a newly developed methodology in the area of survival data analysis, threshold regression brings useful research insights by modeling the underlying health process. Four functions are provided by the **threg** package: the `threg` function can be used to estimate regression coefficients, the `hr` method for ‘`threg`’ objects can be used to calculate hazard ratios, the `predict` method for ‘`threg`’ objects can be used to calculated predicted values and the `plot` method for ‘`threg`’ objects can be used to draw predicted plots for the threshold regression model. Application of threshold regression on a wide varieties of survival data can be carried out easily with the help of the **threg** package. For those readers who are interested in implementing threshold regression in Stata (StataCorp. 2013), we refer them to Xiao, Whitmore, He, and Lee (2012).

Acknowledgments

This project is supported in part by CDC/NIOSH Grant RO1 OH008649.

References

Cox DR (1972). “Regression Models and Life Tables.” *Journal of the Royal Statistical Society B*, **34**(2), 187–230.

Cox DR, Miller HD (1965). *The Theory of Stochastic Processes*. Chapman and Hall, London.

Garrett JM (1997). “**sbe14**: Odds Ratios and Confidence Intervals for Logistic Regression Models with Effect Modification.” *Stata Technical Bulletin*, **36**, 15–22. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 104–114. College Station, TX: Stata Press.

Klein JP, Moeschberger ML (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd edition. Springer-Verlag.

Lee MLT, Whitmore GA (2006). "Threshold Regression for Survival Analysis: Modeling Event Times by a Stochastic Process." *Statistical Science*, **21**(4), 501–513.

Lee MLT, Whitmore GA (2010). "Proportional Hazards and Threshold Regression: Their Theoretical and Practical Connections." *Lifetime Data Analysis*, **16**(2), 196–214.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

StataCorp (2013). *Stata Data Analysis Statistical Software: Release 13*. StataCorp LP, College Station. URL http://www.stata.com/.

Therneau T (2012). ***survival**: A Package for Survival Analysis in S*. R package version 2.36.14, URL http://CRAN.R-project.org/package=survival.

Therneau TM, Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Xiao T (2015). ***threg**: Threshold Regression*. R package version 1.0.3, URL http://CRAN.R-project.org/package=threg.

Xiao T, Whitmore GA, He X, Lee MLT (2012). "Threshold Regression for Time-To-Event Analysis: The **stthreg** Package." *Stata Journal*, **12**(2), 257–283.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. URL http://www.jstatsoft.org/v34/i01/.

Affiliation:

Tao Xiao
Shenzhen University
Shenzhen, People's Republic of China
E-mail: taoxiao1@gmail.com

G. A. Whitmore
McGill University
Montreal, Quebec, Canada
E-mail: whitmore@mcgill.ca

Xin He
University of Maryland
College Park, MD, United States of America
E-mail: xinhe@umd.edu

Mei-Ling Ting Lee
University of Maryland
College Park, MD, United States of America
E-mail: mltlee@umd.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 66, Issue 8 *Submitted:* 2012-04-19
August 2015 *Accepted:* 2014-10-13