




TransModel: An R Package for Linear Transformation Model with Censored Data

Jie Zhou 
University of
South Carolina

Jiajia Zhang
University of
South Carolina

Wenbin Lu
North Carolina
State University

Abstract

Linear transformation models, including the proportional hazards model and proportional odds model, under right censoring were discussed by [Chen, Jin, and Ying \(2002\)](#). The asymptotic variance of the estimator they proposed has a closed form and can be obtained easily by plug-in rules, which improves the computational efficiency. We develop an R package **TransModel** based on Chen's approach. The detailed usage of the package is discussed, and the function is applied to the Veterans' Administration lung cancer data.

Keywords: transformation model, right censored data, R package.

1. Introduction

The proportional hazards (PH) model ([Cox 1992](#)) has been used extensively in many research fields, such as biomedical applications, financial studies and epidemiological studies. However, sometimes the proportional hazards assumption is violated and other models, such as the proportional odds (PO) model ([Dabrowska and Doksum 1988](#)), should be used as alternatives. More generally, the linear transformation model, which includes the PH model and the PO model as special cases, is a broad family of regression models and has attracted considerable attention in recent years due to its flexibility. For example, [Chen *et al.* \(2002\)](#) mentioned that the results from the PO model for the Veterans' Administration lung cancer data are more similar to those reported in the literature compared with the results from the PH model. [Xu, Yang, and Ott \(2005\)](#) claimed the analysis based on transformation models have better prediction capabilities than those from the PH model for their microarray data set.

A general estimation method for the linear transformation model with censored data was proposed in [Cheng, Wei, and Ying \(1995\)](#), and was further developed by [Cai, Wei, and Wilcox \(2000\)](#); [Fine, Ying, and Wei \(1998\)](#); [Cheng, Wei, and Ying \(1997\)](#). A key assumption of their

approach was the independence between the censoring variable and the covariates, which is found to be restrictive in practice. [Chen et al. \(2002\)](#) proposed a unified procedure for the analysis of the linear transformation model, which reduces to the partial likelihood approach in the case of the PH model and its validity does not rely on that independence assumption. Recently, [Hothorn, Möst, and Bühlmann \(2018\)](#) also proposed maximum likelihood estimators in the class of conditional transformation models and developed the **mlt** package ([Hothorn 2020](#)).

Even though the linear transformation model has the attractive property, it has not been widely used due to the lack of easy implemented functions in common statistical software, such as R and SAS. The procedure proposed in [Chen et al. \(2002\)](#) was based on the estimating equations and was easily implemented numerically and computationally efficient. Moreover, the asymptotic variance has a closed form and can be obtained by plug-in rules. Therefore, we develop the R package **TransModel** based on the procedure discussed in [Chen et al. \(2002\)](#), and illustrate its usage in this paper. Package **TransModel** ([Zhou, Zhang, and Lu 2022](#)) is available from the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org/package=TransModel>.

Let T denote the failure time, \mathbf{z} denote a p -dimensional covariate vector and ε represent an error term. The linear transformation model can be expressed as

$$H(T) = -\beta^\top \mathbf{z} + \varepsilon,$$

where $H(\cdot)$ is an unknown monotone transformation function. It links the failure time T with a linear combination of a p -dimensional covariate vector \mathbf{z} and an error term ε . Specific distributions can be assumed for f_ε , which is the density function of ε , to obtain different models. For example, if ε follows a standard extreme value distribution, that is, $f_\varepsilon(s) = \exp(s) \exp(-\exp(s))$, the density function and survival function of T are

$$f_T(t | \mathbf{z}) = \dot{H}(t) \cdot \exp(H(t) + \beta^\top \mathbf{z}) \exp(-\exp(H(t) + \beta^\top \mathbf{z}))$$

and

$$S_T(t | \mathbf{z}) = \exp(-\exp(H(t) + \beta^\top \mathbf{z})),$$

where $\dot{H}(\cdot)$ denotes the first derivative of function $H(\cdot)$. Therefore, the hazard function of T can be written as

$$\lambda_T(t | \mathbf{z}) = \frac{f_T(t | \mathbf{z})}{S_T(t | \mathbf{z})} = \dot{H}(t) \exp(H(t) + \beta^\top \mathbf{z}) = \lambda_0(t) \cdot \exp(\beta^\top \mathbf{z}).$$

Taking $\lambda_0(t) = \dot{H}(t) \exp(H(t))$ as an unspecified baseline hazard function, it reduces to the PH model. Similarly, if ε follows a standard logistic distribution, it becomes a PO model. In our package **TransModel**, we assume the hazard function of ε has the form

$$\lambda_\varepsilon(s) = e^s / (1 + r \times e^s), \quad r \geq 0.$$

Note that the PH model and PO model correspond to $r = 0$ and $r = 1$, respectively.

The rest of the paper is organized as follows. The estimation of both parameters and variances in [Chen et al. \(2002\)](#) paper are summarized in Section 2. Details of the package are discussed in Section 3. The package is illustrated using the Veterans' Administration lung cancer data set in Section 5. Some discussions are outlined in Section 6.

2. Estimation procedure

Let $\mathbf{O} = \{T_j, \delta_j, \mathbf{z}_j; j = 1, \dots, n\}$ denote the observed right censored data set, where $T_j = \min(X_j, C_j)$, $\delta_j = I(X_j \leq C_j)$ and X and C are the nonactive failure time and censoring time, respectively. Let $\Lambda_\varepsilon(\cdot)$ be the cumulative hazard function for ε . Following the usual counting process notation,

$$\begin{aligned} Y(t) &= I(T \geq t) \\ N(t) &= \delta I(T \leq t) \\ M(t) &= N(t) - \int_0^t Y(s) d\Lambda_\varepsilon(\beta_0^\top \mathbf{z} + H_0(s)), \end{aligned}$$

where (β_0, H_0) are the true values of (β, H) . By using the fact that $M(t)$ is a martingale process, the estimating equations are

$$\begin{aligned} U(\beta, H) &= \sum_{j=1}^n \int_0^\infty \mathbf{z}_j [dN_j(t) - Y_j(t) d\Lambda_\varepsilon(\beta^\top \mathbf{z}_j + H(t))] = 0 \\ \sum_{j=1}^n [dN_j(t) - Y_j(t) d\Lambda_\varepsilon(\beta^\top \mathbf{z}_j + H(t))] &= 0, \quad (t \geq 0), \end{aligned} \quad (1)$$

where $H(\cdot)$ is a non-decreasing function satisfying $H(0) = -\infty$.

2.1. Parameter estimation

Follow [Chen et al. \(2002\)](#), the iterative algorithm for computing the β coefficients and H is as follows:

Step 0 : Give initial values for β , say $\hat{\beta}^{(0)} = \mathbf{0}$.

Step 1 : In the i -th iteration, with $\hat{\beta}^{(i)}$, obtain $\hat{H}^{(i)}$ as follows: solving equation

$$\sum_{j=1}^n Y_j(t_1) \Lambda_\varepsilon(\hat{\beta}^{(i)\top} \mathbf{z}_j + H(t_1)) = 1$$

to get $\hat{H}^{(i)}(t_{(1)})$; then calculate

$$\hat{H}^{(i)}(t_{(k)}) = \hat{H}^{(i)}(t_{(k-1)}) + \frac{1}{\sum_{j=1}^n Y_j(t_{(k)}) \lambda_\varepsilon(\hat{\beta}^{(i)\top} \mathbf{z}_j + \hat{H}^{(i)}(t_{(k-1)}))}$$

where $0 < t_{(1)} < t_{(2)} < \dots < t_{(K)} < \infty$ are the K failure times among the n observations.

Step 2 : Update β estimates in $(i+1)$ -th iteration by solving Equation (1) with $H = \hat{H}^{(i)}$ for $k = 2, \dots, K$.

Step 3 : Repeat Step 1 and Step 2 until convergence.

Based on parameter estimates $(\hat{\beta}, \hat{H})$, the survival function for patient with covariate \mathbf{z}_j can be estimated as

$$\hat{S}(t | \mathbf{z}_j) = S_\varepsilon(\hat{H}(t) + \hat{\beta}^\top \mathbf{z}_j).$$

Specifically, in the package **TransModel**, the survival function is defined as

$$\hat{S}(t | \mathbf{z}_j) = \begin{cases} \exp\{-\exp(\hat{H}(t) + \hat{\beta}^\top \mathbf{z}_j)\} & r = 0 \\ \exp\{-\frac{1}{r} \log[1 + r \cdot \exp(\hat{H}(t) + \hat{\beta}^\top \mathbf{z}_j)]\} & r > 0. \end{cases}$$

2.2. Variance estimation of parameters

It is proved in [Chen *et al.* \(2002\)](#) that, under suitable regularity conditions, the derived estimator $\hat{\beta}$ in Section 2.1 is consistent and asymptotic normally distributed. That is,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N\{0, \Sigma_*^{-1} \Sigma^* (\Sigma_*^{-1})^\top\}, \text{ as } n \rightarrow \infty.$$

where Σ_* and Σ^* have closed form solutions

$$\begin{aligned} \hat{\Sigma}^* &= \frac{1}{n} \sum_{j=1}^n \int_0^\tau [\mathbf{z}_j - \bar{\mathbf{z}}(t)] \otimes^2 \lambda\{\hat{\beta}^\top \mathbf{z}_j + \hat{H}(t)\} Y_j(t) d\hat{H}(t) \\ \hat{\Sigma}_* &= \frac{1}{n} \sum_{j=1}^n \int_0^\tau [\mathbf{z}_j - \bar{\mathbf{z}}(t)] \mathbf{z}_j^\top \dot{\lambda}\{\hat{\beta}^\top \mathbf{z}_j + \hat{H}(t)\} Y_j(t) d\hat{H}(t) \end{aligned}$$

respectively. Here we define $\mathbf{b} \otimes^2 = \mathbf{b} \mathbf{b}^\top$ for any vector \mathbf{b} . We also have

$$\begin{aligned} \bar{\mathbf{z}}(t) &= \frac{\sum_{j=1}^n \mathbf{z}_j \lambda\{\hat{\beta}^\top \mathbf{z}_j + \hat{H}(t)\} Y_j(t) \hat{B}(t, T_j)}{\sum_{j=1}^n \lambda\{\hat{\beta}^\top \mathbf{z}_j + \hat{H}(t)\} Y_j(t)} \\ \hat{B}(t, s) &= \exp\left(\int_s^t \frac{\sum_{j=1}^n \dot{\lambda}\{\hat{\beta}^\top \mathbf{z}_j + \hat{H}(x)\} Y_j(x)}{\sum_{j=1}^n \lambda\{\hat{\beta}^\top \mathbf{z}_j + \hat{H}(x)\} Y_j(x)} d\hat{H}(x)\right) \end{aligned}$$

for $t, s \in [0, \tau]$.

2.3. Confidence interval and confidence band of survival function

Since the variance of $H(t)$ does not have a closed form solution, the confidence interval and confidence band for the survival curve are derived through the perturbation techniques. Let $\hat{\varepsilon}_k(\mathbf{z}) = \hat{H}(t_{(k)}) + \hat{\beta}^\top \mathbf{z}$ be the estimate of the error term at time $t_{(k)}$ with covariate \mathbf{z} . Survival probability for subjects with covariate \mathbf{z} at time $t_{(k)}$ is then estimated as

$$\hat{S}(t_{(k)} | \mathbf{z}) = \begin{cases} \exp\{-\exp(\hat{\varepsilon}_k(\mathbf{z}))\} & r = 0 \\ \exp\{-\frac{1}{r} \log[1 + r \cdot \exp(\hat{\varepsilon}_k(\mathbf{z}))]\} & r > 0. \end{cases}$$

The confidence interval of $\hat{S}(t_{(k)} | \mathbf{z})$ can be constructed based on the Wald confidence interval of $\varepsilon_k(\mathbf{z})$, where the variance of $\varepsilon_k(\mathbf{z})$ is derived by perturbation. In the l -th perturbation, $l = 1, 2, \dots, N$, a sequence of random values $\alpha_1^{(l)}, \alpha_2^{(l)}, \dots, \alpha_n^{(l)} \sim \text{Exp}(1)$ is generated, where $\text{Exp}(1)$ is the exponential distribution with scale parameter 1. Then the iterative estimating procedure in Section 2.1 with Step 1 and Step 2 are replaced by the following Step 1* and Step 2*, respectively.

Step 1* : Obtain $\hat{H}_{(l)}^{(i)}(t_{(1)})$ by solving

$$\sum_{j=1}^n Y_j(t_1) \Lambda_\varepsilon \left(\hat{\beta}_{(l)}^{(i)\top} \mathbf{z}_j + H(t_{(1)}) \right) \alpha_j^{(l)} = 1,$$

and for $k \geq 2$

$$\hat{H}_{(l)}^{(i)}(t_{(k)}) = \hat{H}_{(l)}^{(i)}(t_{(k-1)}) + \frac{1}{\sum_{j=1}^n Y_j(t_{(k)}) \lambda_\varepsilon \left(\hat{\beta}_{(l)}^{(i)\top} \mathbf{z}_j + \hat{H}_{(l)}^{(i)}(t_{(k-1)}) \right) \alpha_j^{(l)}}.$$

Step 2* : Update estimates $\hat{\beta}_{(l)}^{(i+1)}$ by solving the following estimation equation,

$$\sum_{j=1}^n \left[dN_j(t) - Y_j(t) d\Lambda_\varepsilon \left(\beta^\top \mathbf{z}_j + \hat{H}_{(l)}^{(i)}(t) \right) \right] \alpha_j^{(l)} = 0, \quad t \geq 0.$$

After convergence, estimates $(\hat{\beta}_{(l)}, \hat{H}_{(l)}(t))$ are obtained and $\hat{\varepsilon}_k^{(l)}(\mathbf{z}) = \hat{H}_{(l)}(t_{(k)}) + \hat{\beta}_{(l)}^\top \mathbf{z}$, $l = 1, 2, \dots, N$. The variance of $(\hat{\varepsilon}_k^{(1)}(\mathbf{z}), \hat{\varepsilon}_k^{(2)}(\mathbf{z}), \dots, \hat{\varepsilon}_k^{(N)}(\mathbf{z}))$ can be used as a consistent estimate of $\text{Var}(\hat{\varepsilon}_k(\mathbf{z}))$. A $(1-\alpha)\%$ point-wise confidence interval for the survival curve can be obtained by transforming the interval $\hat{\varepsilon}_k(\mathbf{z}) \pm \mathbf{Z}_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\varepsilon}_k(\mathbf{z}))}$, $k = 1, 2, \dots, K$, where $\mathbf{Z}_{\alpha/2}$ is the $100(1-\alpha/2)$ -th percentile of the standard normal distribution.

To obtain a critical value for the confidence band, let $\bar{\varepsilon}_k^{(l)}(\mathbf{z})$ be the absolute value of the standardized version of $\hat{\varepsilon}_k^{(l)}(\mathbf{z})$, that is

$$\bar{\varepsilon}_k^{(l)}(\mathbf{z}) = \frac{|\hat{\varepsilon}_k^{(l)}(\mathbf{z}) - \bar{\varepsilon}_k(\mathbf{z})|}{\sqrt{\widehat{\text{Var}}(\hat{\varepsilon}_k(\mathbf{z}))}}, \quad \text{where } \bar{\varepsilon}_k(\mathbf{z}) = \frac{1}{N} \sum_{l=1}^N \hat{\varepsilon}_k^{(l)}(\mathbf{z}).$$

Let $\varepsilon^{(l)}$ denote the maximum value of $(\bar{\varepsilon}_1^{(l)}(\mathbf{z}), \bar{\varepsilon}_2^{(l)}(\mathbf{z}), \dots, \bar{\varepsilon}_K^{(l)}(\mathbf{z}))$, $l = 1, \dots, N$, and Q_α denote the $(1-\alpha)\%$ -th quantile of $(\varepsilon^{(1)}, \dots, \varepsilon^{(N)})$. A $(1-\alpha)\%$ confidence band for the survival curve can be obtained by transforming the interval $\hat{\varepsilon}_k(\mathbf{z}) \pm Q_\alpha \sqrt{\widehat{\text{Var}}(\hat{\varepsilon}_k(\mathbf{z}))}$, $k = 1, 2, \dots, K$.

3. Package description

The contributed R package **TransModel** is used to fit linear transformation models for right censored data using the estimation approach discussed in Section 2. In this section, we list the main functions in this package. The main function in this package is **TransModel**, which can be called with the following syntax:

```
TransModel(formula, data, r, CICB.st = FALSE, subset, dx = 0.001,
  iter.max = 100, num.sim = 200)
```

The required arguments include:

- **formula**: A survival formula based on the **Surv()** function, containing survival time, right censoring indicator and covariates.

- **data**: The data set with all the variables needed in the formula.
- **r**: Parameter in the hazard function of the error term, described in Equation 1.
- **CICB.st**: Whether or not the perturbation for the confidence interval and confidence bands of survival estimates will be done. The default value is **FALSE**.
- **subset**: The conditions used to select a subset of the data.
- **dx**: The tolerance limit of convergence. Default is 0.001.
- **iter.max**: The maximum number of iterations before convergence. Default is 100.
- **num.sim**: The number of perturbations used, only works when **CICB.st = TRUE**. Default is 200.

Returned values are:

- **coefficients**: Estimated β coefficients in the transformation model.
- **vcov**: Estimated covariance matrix for the coefficients.
- **converged**: Convergence status and the number of iterations used for convergence. The value 0 indicates that the algorithm converged.

The **print** command gives the coefficient estimate for each covariate specified in the formula as well as the variance-covariance matrix. Further inference about the coefficients can be obtained by using the **summary** method, where the parameter estimates, standard deviation, test statistics and p values based on the Wald-test are presented in a summary table.

The predicted survival probabilities at given time points for a specific covariates vector can be obtained using the **predict** command for the object from the function **TransModel**. Syntax for the function is:

```
predict(object, newdata, new.time, alpha)
```

The required arguments include:

- **object**: An object returned from function **TransModel**.
- **newdata**: A vector with values for each covariate variable. If not specified, 0 will be used for all variables.
- **new.time**: A vector of ordered time points to be used for survival probability calculation. If null, the time points in the original data set will be used.
- **alpha**: Confidence level for calculating the confidence intervals and confidence bands of the survival estimate. Default value is 0.05.

Possible values returned are:

- **time**: Ordered time points on which survival probabilities are calculated.

- `survival`: Predicted survival probabilities.
- `low.ci`: The lower limit of the confidence interval.
- `up.ci`: The upper limit of the confidence interval.
- `low.cb`: The lower limit of the confidence band.
- `up.cb`: The upper limit of the confidence band.

If new time points are not specified in `new.time`, the ordered event time points in the original data will be used. The predicted survival probabilities at each time point and with the covariate values specified in `newdata` will be returned. Only when `CICB.st = TRUE` is specified in the object, the lower and upper limits for the confidence interval and confidence bands will be returned as well. The `plot` command can be applied to the returned object to get a predicted survival curve, and if the confidence limits or the confidence bands were calculated in the prediction step, they can be shown by specifying `CI = TRUE`, `CB = TRUE` or both.

4. Simulation studies

We design a simulation study to evaluate the performance of the proposed method. We assume the transformation function has the form $H(t) = \log(1+t) + t^{3/2}$. A two-dimensional covariate $\mathbf{z} = (z_1, z_2)^\top$ is considered, where $z_1 \sim N(0, 1)$ and $z_2 \sim U(-1, 1)$. The coefficients are set to be $\beta = (1, -1)$. Different models with $r = 0, 0.5, 1, 2$ are considered. The sample size is chosen as $n = 200$ and 500 . The right censoring proportions of 15% and 40% are considered.

We conduct 1000 replications for each setting, and report the bias and average estimated standard deviation (StErr, or SE), empirical standard deviation (StDev, or SD) and empirical coverage probability (CP) of 95% Wald-type confidence intervals. From the result in Table 4, we can see the package gives unbiased estimates, comparable StErrs and StDevs, and reasonable CPs that are close to the nominal level of 0.95.

5. Veterans' Administration lung cancer data

We use the Veterans' Administration lung cancer data as an example to illustrate the usage of the package **TransModel**. The data has been analyzed in [Prentice \(1973\)](#), [Bennett \(1983\)](#), [Pettitt \(1984\)](#) and [Cheng *et al.* \(1995\)](#), and is available in the current **survival** package ([Therneau 2021](#)). Following [Chen *et al.* \(2002\)](#), we use the subgroup of 97 patients who had no prior therapy usage. The covariates of interest include one categorical variable tumor type (large, adeno, small or squamous) and one continuous variable performance status.

Following [Bennett \(1983\)](#), [Murphy, Rossini, and van der Vaart \(1997\)](#) and [Chen *et al.* \(2002\)](#), we first fit a PO model using the main function `TransModel` in the package with $r = 1$.

```
R> set.seed(100)
R> veteran$celltype <- relevel(veteran$celltype, ref = "squamous")
R> fit <- TransModel(Surv(time, status) ~ karno + factor(celltype),
+   data = veteran, r = 1, CICB.st = TRUE, subset = (prior == 0))
```

<i>n</i>	Cens	Pars	<i>r</i> = 0				<i>r</i> = 0.5			
			Est.	StdErr	SD	CP	Est.	StdErr	SD	CP
200	15%	β_1	1.021	0.102	0.108	0.925	1.016	0.119	0.134	0.914
		β_2	-1.005	0.149	0.155	0.944	-0.996	0.195	0.204	0.944
	40%	β_1	1.023	0.117	0.121	0.946	1.024	0.130	0.145	0.906
		β_2	-1.006	0.176	0.182	0.947	-1.001	0.213	0.227	0.936
500	15%	β_1	1.006	0.063	0.066	0.947	1.004	0.075	0.084	0.928
		β_2	-1.008	0.093	0.094	0.950	-1.005	0.122	0.127	0.933
	40%	β_1	1.008	0.072	0.075	0.944	1.007	0.081	0.092	0.922
		β_2	-1.005	0.110	0.112	0.939	-1.004	0.133	0.139	0.936
			<i>r</i> = 1				<i>r</i> = 2			
200	15%	β_1	1.014	0.151	0.161	0.938	1.016	0.214	0.217	0.945
		β_2	-0.989	0.249	0.254	0.948	-0.973	0.357	0.360	0.951
	40%	β_1	1.028	0.159	0.169	0.925	1.036	0.210	0.216	0.949
		β_2	-1.002	0.258	0.271	0.937	-1.005	0.343	0.356	0.941
500	15%	β_1	1.003	0.095	0.103	0.930	1.002	0.134	0.140	0.941
		β_2	-1.005	0.157	0.160	0.945	-1.006	0.224	0.226	0.949
	40%	β_1	1.007	0.099	0.107	0.927	1.009	0.130	0.136	0.946
		β_2	-1.002	0.161	0.165	0.942	-1.003	0.214	0.214	0.954

Table 1: Simulation results.

The estimated coefficients and covariance matrix estimates can be extracted by using `fit$coefficient` and `fit$vcov`, respectively. Further inference about the coefficients can be obtained by the `summary` method:

```
R> summary(fit)
```

Call:

```
TransModel.default(formula = Surv(time, status) ~ karno + factor(celltype),
  data = veteran, r = 1, CICB.st = TRUE, subset = (prior == 0))
```

	Estimate	StdErr	z.value	p.value	
karno	-0.043716	0.010691	-4.0891	4.331e-05	***
factor(celltype)adeno	1.874020	0.623932	3.0036	0.002668	**
factor(celltype)large	0.409877	0.665074	0.6163	0.537705	
factor(celltype)smallcell	1.621423	0.604667	2.6815	0.007329	**

Other values can also be specified for r , such as 0, 0.5 and 2, the results are listed in Table 2, where $r = 0$ corresponds to the PH model.

In addition to the β coefficients, sometimes comparing survival probabilities among different groups is also an interest. In the lung cancer data, for illustration, we compare survival curves among the four tumor types at median level of the performance status 60. The codes are listed below and the predict survival curves for each tumor type are in Figure 1.

```
R> pred1 <- predict(fit, newdata = c(60, 0, 0, 0))
R> pred2 <- predict(fit, newdata = c(60, 1, 0, 0))
```


Coefficients	$r = 0$			$r = 0.5$			$r = 2$		
	Est.	StdErr	p value	Est.	StdErr	p value	Est.	StdErr	p value
Karno	-0.024	0.006	< 0.001	-0.034	0.008	< 0.001	-0.063	0.016	< 0.001
Adeno vs. Squamous	1.069	0.341	0.002	1.485	0.472	0.002	2.610	0.938	0.005
Large vs. Squamous	0.210	0.347	0.545	0.315	0.502	0.531	0.570	0.994	0.566
Small vs. Squamous	0.769	0.302	0.011	1.228	0.456	0.007	2.316	0.895	0.010

Table 2: Estimated regression coefficients for the Veterans' Administration lung cancer data.

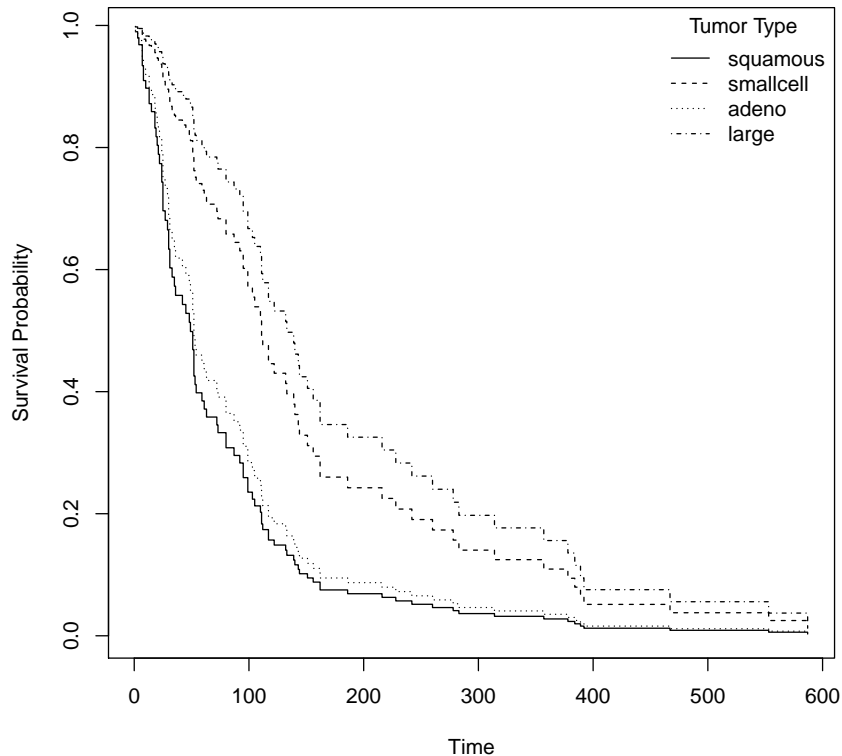


Figure 1: Estimated survival curves for different tumor types.

```
R> pred3 <- predict(fit, newdata = c(60, 0, 1, 0))
R> pred4 <- predict(fit, newdata = c(60, 0, 0, 1))
R> plot(pred1, lty = 1, lwd = 2, cex.axis = 1.5, cex.lab = 1.5)
R> lines(pred2$time, pred2$survival, type = "s", lty = 2, lwd = 2)
R> lines(pred3$time, pred3$survival, type = "s", lty = 3, lwd = 2)
R> lines(pred4$time, pred4$survival, type = "s", lty = 4, lwd = 2)
R> legend("topright", c("squamous", "adeno", "large", "smallcell"),
+       title = "Tumor Type", lty = 1:4, bty = "n")
```

The perturbed variance is used to obtain the 95% confidence interval and confidence bands, which can be presented along with the estimated survival curves. The confidence level can be changed by specifying different values for the argument `alpha` in the `predict` function. For example, the estimated survival curves and 95% confidence interval and confidence bands for patient with performance status at 60 and Squamous type of tumor size can be plotted using the following code and the resulting plot is in Figure 2.

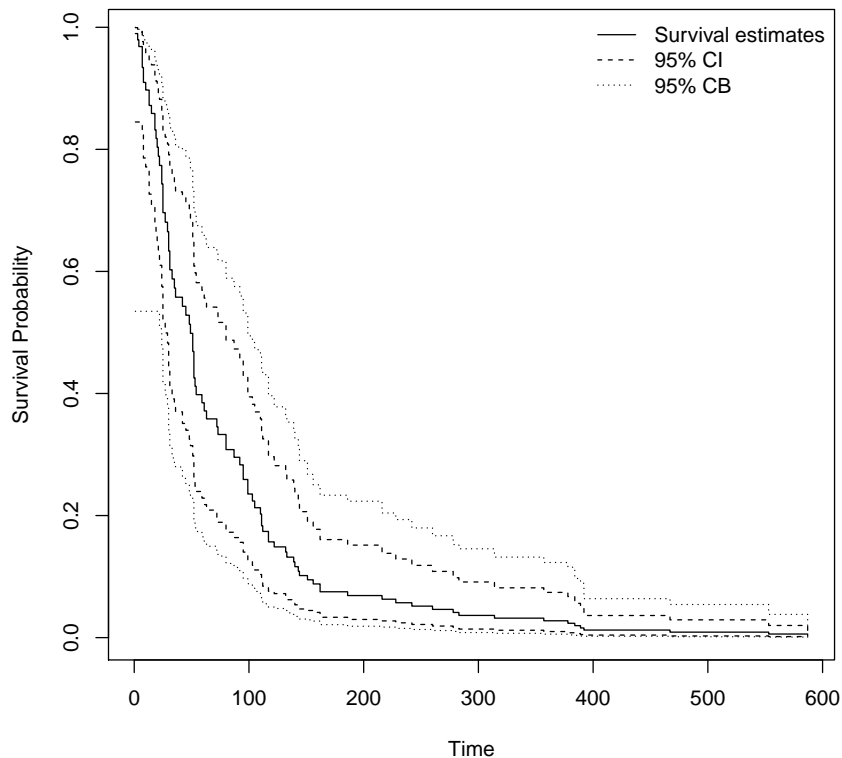


Figure 2: Estimated survival curve for patients with squamous tumor type, with 95% confidence interval and 95% confidence band.

```
R> plot(pred1)
R> lines(pred1$time, pred1$low.ci, lty = 2)
R> lines(pred1$time, pred1$up.ci, lty = 2)
R> lines(pred1$time, pred1$low.cb, lty = 3)
R> lines(pred1$time, pred1$up.cb, lty = 3)
R> legend("topright", c("Survival estimates", "95% CI", "95% CB"),
+       lty = 1:3, lwd = 1, bty = "n")
```

6. Discussion

We develop the package **TransModel** in R to fit linear transformation models with right censored data. The commonly used PH model and PO model are included as special cases with $r = 0$ and 1. The package can provide coefficients, standard deviation, and p value. Furthermore, it can predict survival curves with its confidence interval and confidence band through perturbation.

References

Bennett S (1983). "Analysis of Survival Data by the Proportional Odds Model." *Statistics in Medicine*, **2**(2), 273–277. doi:10.1002/sim.4780020223.

- Cai T, Wei LJ, Wilcox M (2000). “Semiparametric Regression Analysis for Clustered Failure Time Data.” *Biometrika*, pp. 867–878. doi:10.1093/biomet/87.4.867.
- Chen K, Jin Z, Ying Z (2002). “Semiparametric Analysis of Transformation Models with Censored Data.” *Biometrika*, **89**(3), 659–668. doi:10.1093/biomet/89.3.659.
- Cheng SC, Wei LJ, Ying Z (1995). “Analysis of Transformation Models with Censored Data.” *Biometrika*, **82**(4), 835–845. doi:10.1093/biomet/82.4.835.
- Cheng SC, Wei LJ, Ying Z (1997). “Predicting Survival Probabilities with Semiparametric Transformation Models.” *Journal of the American Statistical Association*, **92**(437), 227–235. doi:10.1080/01621459.1997.10473620.
- Cox DR (1992). “Regression Models and Life-Tables.” In *Breakthroughs in Statistics*, pp. 527–541. Springer-Verlag.
- Dabrowska DM, Doksum KA (1988). “Estimation and Testing in a Two-Sample Generalized Odds-Rate Model.” *Journal of the American Statistical Association*, **83**(403), 744–749. doi:10.1080/01621459.1988.10478657.
- Fine JP, Ying Z, Wei LJ (1998). “On the Linear Transformation Model for Censored Data.” *Biometrika*, pp. 980–986. doi:10.1093/biomet/85.4.980.
- Hothorn T (2020). “Most Likely Transformations: The **mlt** Package.” *Journal of Statistical Software*, **92**(1). doi:10.18637/jss.v092.i01.
- Hothorn T, Möst L, Bühlmann P (2018). “Most Likely Transformations.” *Scandinavian Journal of Statistics*, **45**(1), 110–134. doi:10.1111/sjos.12291.
- Murphy SA, Rossini AJ, van der Vaart AW (1997). “Maximum Likelihood Estimation in the Proportional Odds Model.” *Journal of the American Statistical Association*, **92**(439), 968–976. doi:10.1080/01621459.1997.10474051.
- Pettitt AN (1984). “Proportional Odds Models for Survival Data and Estimates Using Ranks.” *Applied Statistics*, pp. 169–175. doi:10.2307/2347443.
- Prentice RL (1973). “Exponential Survivals with Censoring and Explanatory Variables.” *Biometrika*, **60**(2), 279–288. doi:10.1093/biomet/60.2.279.
- Therneau TM (2021). **survival**: *Survival Analysis*. R package version 3.2-13, URL <https://CRAN.R-project.org/package=survival>.
- Xu J, Yang Y, Ott J (2005). “Survival Analysis of Microarray Expression Data by Transformation Models.” *Computational Biology and Chemistry*, **29**(2), 91–94. doi:10.1016/j.compbiolchem.2005.02.001.
- Zhou J, Zhang J, Lu W (2022). **TransModel**: *Fit Linear Transformation Models for Right Censored Data*. R package version 2.3, URL <https://CRAN.R-project.org/package=TransModel>.

Affiliation:

Jiajia Zhang

Department of Epidemiology and Biostatistics

Faculty of Biostatistics

University of South Carolina

915 Greene Street, Columbia, SC 29208, United States of America

Telephone: +1/803/777-4474

E-mail: jzhang@mailbox.sc.edu

URL: http://www.sph.sc.edu/epid_bios/facultystaffdetails.asp?id=575