Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2022, Vol. 31(12) 2352–2367

DOI: 10.1177/09622802221122406
journals.sagepub.com/home/smm
SAGE

Flexible semiparametric mode regression for time-to-event data

Alexander Seipp[1], Verena Uslar[2], Dirk Weyhe[2], Antje Timmer[1], and Fabian Otto-Sobotka[1]

Abstract

The distribution of time-to-event outcomes is usually right-skewed. While for symmetric and moderately skewed data the mean and median are appropriate location measures, the mode is preferable for heavily skewed data as it better represents the center of the distribution. Mode regression has been introduced for uncensored data to model the relationship between covariates and the mode of the outcome. Starting from nonparametric kernel density based mode regression, we examine the use of inverse probability of censoring weights to extend mode regression to handle right-censored data. We add a semiparametric predictor to add further flexibility to the model and we construct a pseudo Akaike's information criterion to select the bandwidth and smoothing parameters. We use simulations to evaluate the performance of our proposed approach. We demonstrate the benefit of adding mode regression to one's toolbox for analyzing survival data on a pancreatic cancer data set from a prospectively maintained cancer registry.

Keywords

Iteratively weighted least squares, P-splines, inverse probability weights, inverse probability of censoring, pancreatic cancer

1 Introduction

Mode regression (or modal regression) can be used to predict the mode of a continuous outcome conditional on covariates. The mode is of particular interest for highly skewed and multimodal data, where the mean and even the median can be unlikely values in a region of low density. They can be far away from the highest density region in the sense of Hyndman,[1] which we intuitively think of as the center of the distribution. In contrast, the mode can be seen as a more typical observation, an interpretation that might be rather appealing for applied scientists.

In this article, we extend mode regression to right-censored data. We imagine mode regression being a useful tool in analyzing time-to-event data with heavy skewness or multimodal data. In our application, we used the method to analyze survival times of pancreatic cancer patients. Pancreatic cancer is a highly fatal disease. It is often detected late and many patients die within the first year after diagnosis. In 2016, the one-year survival rate in Germany was 33% relative to survival in the general population.[2] While the underlying mechanisms are still unclear, some patients can have long survival times. The five-year survival rate in Germany was 9%. Overall survival was therefore expected to be right-skewed and results from mode regression were of interest.

Existing approaches to mode regression are often based on kernel density estimation. Other approaches include direct derivative estimation and obtaining the mode via quantile regression among others.[3,4] A summary of kernel density based mode regression methods can be found in a review by Chen.[5] Mode regression can be categorized into global and local mode regression. Local mode regression aims to model the relationship between covariates and multiple local maxima

[1]Division of Epidemiology and Biometry, Faculty of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg, Germany
[2]University Hospital for Visceral Surgery, Pius-Hospital Oldenburg, Germany

Corresponding author:
Fabian Otto-Sobotka, Division of Epidemiology and Biometry, Faculty of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg, Ammerländer Heerstr. 114-118, 26129 Oldenburg (Oldb), Germany.
Email: fabian.otto-sobotka@uni-oldenburg.de

of the conditional density. This can be done by estimating the joint density of outcome and regressors with multivariate kernel density estimation and maximization of the density for fixed covariate values.[6] Continuous covariates are modeled without the need to specify an explicit functional form. This flexibility, however, also leads to reduced interpretability, since covariate effects cannot be summarized in a few numerical estimates, but rather have to be inspected visually. Global mode regression, on the other hand, is concerned with the relationship between covariates and the global maxima of the conditional density. Common approaches are based on minimizing expected loss.[7] Our method is based on Kemp and Santos Silva,[8] who proposed minimizing empirical loss by maximizing a one-dimensional kernel density estimator. Covariate effects were included with a linear predictor, which simplified interpretability in comparison to smooth modeling of covariate effects. Still, flexible modeling as an option is desirable in many situations, which is why we aim to include flexible modeling in our approach.

Mode regression falls into the class of distributional regression methods that have been developed in the last decades to analyze different measures than just the conditional mean in the regression setting. While methods like quantile regression,[9] expectile regression[10] and Generalized Additive Models for Location, Scale and Shape (GAMLSS)[11] can be used to analyze tail behavior and are particularly useful in heteroscedastic settings, mode regression is an alternative method to quantify effects on the center of the distribution. Mode regression is a useful alternative if the data or context imply that the expectation and the median are not estimands of interest.

The approach taken in this paper starts from linear mode regression proposed by Kemp and Santos Silva[8] and adds useful extensions. The first extension is the inclusion of semiparametric predictors, allowing for the flexible modeling of spatial and temporal information as well as nonlinear effects of continuous covariates. We use penalized splines (P-splines), as made popular by Eilers and Marx,[12] for inclusion of nonlinear effects in mode regression. By using the normal kernel, computation of the estimator given in Kemp and Santos Silva reduces to a weighted least squares problem, which allows direct inclusion of least squares methods like P-splines.

The second important component of our model is the inclusion of right-censored data with inverse probability of censoring (IPC) weights, which has been examined, for example, for local mode regression[13] and expectile regression.[14] The weights lead to an exclusion of every censored event time, while every uncensored event time is used with an increased weight in estimation.

Kernel density estimation requires specification of the so-called bandwidth parameter, that controls smoothness of the estimated density. A bandwidth that is too small leads to an estimate with high variability, large bandwidths lead to smooth estimates with high bias. The importance of bandwidth selection for univariate kernel density estimation is often emphasized, and considerable research has been conducted to find an optimal selection method.[15] Specification of a bandwidth is also needed for kernel density based mode regression. Minimizing the asymptotic MSE has been proposed to choose the bandwidth for linear mode regression.[16] We examine the use of a pseudo Akaike's information criterion (AIC) approach for selection beyond linear mode regression and uncensored data. We use the pseudo AIC to select both the bandwidth and the smoothness parameters of the penalized effects. Further, it can also be used as a model selection criterion.

In this article, we introduce and evaluate an extended mode regression model for right-censored data and use it to analyze pancreatic cancer data. In the section on "Method development," we first summarize global mode regression based on kernel density estimation. We then extend the model to include semiparametric predictors and right-censored data. In the section concerning the "Analysis of overall survival of pancreatic cancer patients," we present an analysis of survival times of pancreatic cancer patients, showing the potential benefits of using mode regression for survival analysis. In the section on "Numerical studies," we show results from a simulation study, evaluating performance of our approach in comparison with GAMLSS.

2 Method development

2.1 Mode estimation and mode regression

The (global) mode of a continuous random variable T with density $f(t)$ is defined as the value that maximizes the density,

$$\text{Mode(T)} = \arg\max_t f(t) = \arg\max_t \lim_{s\to 0} \frac{P(t \le T < t+s)}{s}$$

The mode of continuous variables can be interpreted similarly to the discrete case. For discrete variables, the mode is the value that occurs with the highest probability. For a continuous variable, since $f(t) \approx P(t \le T < t+s)/s$ for some small s, we can imagine binning the data into intervals of length s and choosing the interval with the highest probability. The chosen interval contains the mode if s is small enough.

There are various ways to estimate the mode, see for example Chacon[17] for an overview. Parzen[18] already considered mode estimation in his first publication on kernel density estimation. It turns out that simply maximizing the kernel density leads to a consistent estimator of the mode. For independent observations $t_1, \ldots, t_n$ it is given by

$$\hat{\beta}_0 = \arg\max_{\beta_0} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t_i - \beta_0}{h}\right),$$

where K is a Kernel function and h is a smoothness parameter called bandwidth. The choice of the bandwidth parameter is generally recognized as an important decision. Undersmoothing leads to higher variance, low bias and tends to produce too many local modes. Oversmoothing, on the other hand, can obfuscate the underlying structure.

Kemp and Santos Silva[8] as well as Yao and Li[16] extended estimation of the mode to the regression setting. The framework is similar to usual linear least squares regression. Let $T_1, \ldots, T_n$ depend on covariates $x_1, \ldots, x_p$ linearly with parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. The relationship is given by

$$T_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

where the ε_i are independent, but not necessarily identically distributed. ε_i are assumed to have a unique mode of 0, which assures that the mode of T_i is the linear predictor $\mathbf{x}_i^\top \boldsymbol{\beta}$. The risk function

$$R(\boldsymbol{\beta}, \mathbf{x}) = \mathrm{E}\left[1 - \mathrm{K}(0)^{-1}\mathrm{K}\left(\frac{\mathrm{T} - \mathbf{x}_\mathrm{i}^\top \boldsymbol{\beta}}{\mathrm{h}}\right)\right],$$

is assumed, whose sample analog is minimized by the parameter estimates $\hat{\boldsymbol{\beta}}_h$, defined as

$$\hat{\beta}_h = \arg\max_{\beta} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{T - \mathbf{x}_i^\top \boldsymbol{\beta}}{h}\right). \tag{1}$$

Optimization in (1) can be done by using an expectation-maximization algorithm, as proposed in Yao and Li.[16] The algorithm reduces to minimizing iteratively weighted least squares for the Gaussian kernel, while no closed form expressions follow from using other kernels. For the remainder of this article, K is chosen as the Gaussian kernel to reduce computational complexity. The estimates are then given by

$$\begin{aligned} \hat{\beta}_h^{[j+1]} &= \arg\max_{\beta} \sum_{i=1}^{n} w_i\left(\hat{\beta}_h^{[j]}, h\right)\left(t_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2, \\ w_i\left(\hat{\beta}_h^{[j]}, h\right) &= \frac{\phi_h(t_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_h^{[j]})}{\sum_{i=1}^{n} \phi_h(t_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_h^{[j]})}, \end{aligned} \tag{2}$$

where $\hat{\boldsymbol{\beta}}_h^{[j]}$ is the estimate in iteration j and $\phi_h(t)$ is the density of a normal distribution with zero mean and a standard deviation of h. The weights w are larger for observations close to the mode and smaller or even close to zero for observations more distant. It has been shown that convergence of the algorithm is guaranteed, although the solution is dependent on the starting values.[16] Therefore, multiple starting values should be used. Further, differentiability of the normal kernel was used to prove consistency of the estimator.

We propose two extensions to the regression approach formulated in equation (1). First, we allow right-censored data by adding IPC weights. Second, since iteratively weighted least squares are used, we can incorporate semiparametric predictors like P-splines, Gaussian Markov random fields and other flexibly modeled effects.

2.2 Time-to-event data

In the case of right-censored time-to-event data, estimation of the survival function is typically done with the Kaplan-Meier estimator.[19] The estimator has many desirable properties like strong consistency and can be viewed as a nonparametric maximum-likelihood estimator.[20,21] On the other hand, the Kaplan-Meier estimator is always a step function, even for continuous event times, which is an undesirable property. To remedy this, smoothing of the Kaplan-Meier estimator has been proposed, using Bezier curves or kernel smoothing among others.[22,23] Kernel smoothing uses the step sizes of the Kaplan-Meier estimator v_i as weights. Defining censoring times $c_1, \ldots, c_n$, follow-up times $y_i = \min(t_i, c_i)$ and event indicator $\delta_i = 1(t_i \leq c_i)$, the density of the event time can be estimated with

$$\hat{f}_h(t) = \frac{1}{h}\sum_{i=1}^{n} v_i K\left(\frac{y_i - t}{h}\right),$$

$$v_i = \frac{1}{n}\frac{\delta_i}{\hat{P}(C_i > y_i)},$$

where $\hat{P}(C_i > y_i)$ is the Kaplan-Meier estimator of the censoring time. The combination of the Kaplan-Meier estimator and kernel density estimation can be viewed as smoothing the Kaplan-Meier estimator, but also as extending kernel density estimation to right-censored data. We use this idea to extend mode regression to right-censored data. By adding Kaplan-Meier weights, unknown event times t_i can be replaced by follow-up times y_i. Estimates are then defined as

$$\hat{\beta}_h = \arg\max_{\beta} \frac{1}{h}\sum_{i=1}^{n} v_i K\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{h}\right), \tag{3}$$

which can again be maximized with iteratively weighted least squares.

The weights v_i can be seen as either the step sizes of the Kaplan-Meier estimator of the event time or, equivalently, as inverse probability of censoring (IPC) weights.[24] The weights exclude censored observations, while the weights are larger than one for uncensored observations. An uncensored observation can be thought of as representative of several unobserved event times, indicated by the size of the weight. IPC weights are a special case of inverse probability weights, which is a fairly general and well-known approach used in the analysis of missing data. In time-to-event analysis IPC weights have been used for extending mean regression,[25] median regression,[24] expectile regression[14] and also local mode regression[13] to model right-censored data.

A potential problem that is encountered in using IPC is that some weights are lost if the longest observation(s) are censored. For mean regression, Zhou[25] added the lost weights to the longest observation to remedy this. In our experience, this problem is not as relevant for mode regression and a correction is not necessary, since the longest observations are usually far enough away from the mode of the distribution to have little to no effect on the estimates.

2.3 Semiparametric regression

For the models in the previous sections, effects that could be modeled by simple and known functions were assumed. We prefer a more flexible approach by including a semiparametric predictor. A structured additive model as proposed in Fahrmeir et al.[26] allows us to capture effects more broadly. It is generally defined as

$$T_i = \mathbf{x}_i^\top \boldsymbol{\beta} + f_1(z_1) + \cdots + f_r(z_r) + \varepsilon_i,$$

where the f_j are unknown functions and z_j are covariates which can be of various types and can have, for example, nonlinear, spatial, or random effects. An overview of the different possible effects and estimation methods can be found in Fahrmeir et al.[26] In this article we present P-splines as an example to model nonlinear effects. In this approach, a set of equally spaced, local polynomial functions $B(z)$ form a basic spline base. The functions are combined linearly with parameter vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)$, such that the true unknown function f_j can be approximated by

$$\hat{f}_j(z_j) = \sum_{k=1}^{d} \gamma_k B_k(z_j).$$

A large number of polynomial functions d is chosen, while a penalty dependent on a smoothing parameter λ is used to avoid overfitting. As in Eilers and Marx,[12] the penalty here consists of the sum of squared second order differences of neighboring coefficients, that is,

$$\lambda \boldsymbol{\gamma}^\top \mathbf{K} \boldsymbol{\gamma} = \lambda \sum_{j} (\gamma_j - 2\gamma_{j-1} + \gamma_{j-2})^2,$$

with penalty matrix $\mathbf{K}$. Extending our model, we define the parameter vector $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r)$ and the design matrix $\mathbf{Z} = (\mathbf{X}, \mathbf{B}_1, \ldots, \mathbf{B}_r)$. $\mathbf{X}$ is the usual design matrix of all parametric effects and the intercept, while each $\mathbf{B}$ represents further additive effects and includes the corresponding set of polynomial functions. The outcome can then be described with

$$\mathbf{T} = \mathbf{Z}\boldsymbol{\vartheta} + \boldsymbol{\varepsilon}$$

The coefficients can be estimated with an iteratively weighted least squares algorithm. Let $\mathbf{P}(\boldsymbol{\lambda})$ denote a block matrix with blocks of penalty terms, $\mathbf{P}(\boldsymbol{\lambda}) = \operatorname{diag}(\mathbf{0}, \lambda_1\mathbf{K}, \ldots, \lambda_r\mathbf{K})$. Assuming that h and $\boldsymbol{\lambda}$ were already selected, estimates are updated as follows:

$$\hat{\boldsymbol{\vartheta}}^{[m+1]}_{h,\lambda} = \arg\min_{\boldsymbol{\vartheta}} \sum_{i=1}^{n} \omega_i\left(\hat{\boldsymbol{\vartheta}}^{[m]}_{h,\lambda}, h\right)\left(y_i - \mathbf{z}_i^\top \boldsymbol{\vartheta}\right)^2 + \boldsymbol{\vartheta}^\top \mathbf{P}(\boldsymbol{\lambda})\boldsymbol{\vartheta}$$

$$\omega_i\left(\hat{\boldsymbol{\vartheta}}^{[m]}_{h,\lambda}, h\right) = \frac{1}{2h^2}\frac{v_i\phi_h(y_i - \mathbf{z}_i^\top \hat{\boldsymbol{\vartheta}}^{[m]}_{h,\lambda})}{\sum_{i=1}^{n} v_i\phi_h(y_i - \mathbf{z}_i^\top \hat{\boldsymbol{\vartheta}}^{[m]}_{h,\lambda})}$$

The algorithm reduces to equation (2) for linear covariates. The algorithm maximizes the following objective function:

$$J_{h,\lambda}(\boldsymbol{\vartheta}) = \log\left(\frac{1}{h}\sum_{i=1}^{n} v_i K\left(\frac{y_i - \mathbf{z}_i^\top \boldsymbol{\vartheta}}{h}\right)\right) - \boldsymbol{\vartheta}^\top \mathbf{P}(\boldsymbol{\lambda})\boldsymbol{\vartheta}$$

Convergence for fixed h and $\boldsymbol{\lambda}$ is guaranteed, as shown in the Appendix.

2.4 Hyperparameter selection

While the bandwidth h and the smoothing parameter λ can be chosen manually, a more objective and convenient way is data-driven selection of hyperparameters. Using the normal kernel, Yao and Li[16] derived asymptotics in the linear case and proposed to choose h by minimization of the mean squared error with a plug-in approach. Common approaches to select λ used in GAMs include restricted maximum likelihood or using model fit criteria like generalized cross-validation or AIC. In our extended model, we propose to choose h and λ simultaneously by minimizing a pseudo AIC, calculated with a pseudo-likelihood approach. The pseudo-likelihood approach has been developed by Duin[27] as one of the first bandwidth selection approaches for univariate kernel density estimation. The pseudo-likelihood is defined as the product of 'leave-one-out' kernel density estimates of every data point. For mode regression, we use the pseudo-likelihood of the residuals. We define

$$L(h, \lambda, \hat{\boldsymbol{\varepsilon}}) = \prod_{i=1}^{n} \hat{f}_{-i}(\hat{\varepsilon}_i, h)^{\delta_i}(1 - \hat{F}_{-i}(\hat{\varepsilon}_i, h))^{1-\delta_i},$$

$$\hat{f}_{-i}(u, h) = \frac{1}{h}\sum_{j\neq i} v_j K\left(\frac{\hat{\varepsilon}_j^{[-i]} - u}{h}\right),$$

where $\hat{F}_{-i}(k, h) = \int_{-\infty}^{k} \hat{f}_{-i}(u, h)\,\mathrm{d}u$, $\hat{\varepsilon}_i = y_i - \mathbf{z}_i^\top \hat{\boldsymbol{\vartheta}}_{h,\lambda}$ and $\hat{\varepsilon}_j^{[-i]}$ is the jth cross-validated residual of a fit, that leaves out the ith observation. The pseudo AIC can then be defined as

$$AIC(h, \lambda, \hat{\boldsymbol{\varepsilon}}) = -2\log(L(h, \lambda, \hat{\boldsymbol{\varepsilon}})) + 2p \tag{4}$$

where p represents the model size, taken to be the effective degrees of freedom from the weighted GAM. Optimization of the pseudo AIC requires refitting of the model several times, since hyperparameters and residuals depend on each other. However, the pseudo-likelihood for fixed residuals can be calculated efficiently from the fit with all observations. The influence matrix $\mathbf{H}$ of the weighted GAM can be used for calculation of the cross-validated residuals. They can be calculated with

$$\hat{\varepsilon}_j^{[-i]} = \hat{\varepsilon}_j + H_{ji}\frac{\hat{\varepsilon}_i}{1 - H_{ii}},$$

where H_{ji} is the value in the jth row and ith column of $\mathbf{H}$. The proof follows directly from results given in Wood, Section 6.2.2.[28]

2.5 Implementation of point estimates

We implemented mode regression with the software R.[29] A summary of our algorithm is given in Algorithm 1. As a first step, starting values for hyperparameters and the mode are chosen. An arbitrary starting value of one is chosen for λ. A GAM is fit and the resulting predictions for the conditional mean $\mathbf{z}_i^\top \hat{\boldsymbol{\vartheta}}_{[LS]}$ are used as starting values for the mode. We use the R package "mgcv"[28] for fitting GAMs. A starting value for the bandwidth is calculated by assuming that for the normal kernel only values with a distance less than three standard deviations (3 h) have a meaningful contribution. Assuming relevance of 50% of the data points results in the starting value $\hat{h}$ = median($|\hat{\boldsymbol{\varepsilon}}|/3$).

Next, hyperparameters are selected. We use nonlinear optimization algorithms from the R package ”nloptr”.[30] After every update of the hyperparameters, the iteratively weighted GAM is refitted and the pseudo-likelihood is recalculated. After selection of hyperparameters, we calculate multiple starting values for the mode, since estimates can be sensitive to the initial values. We use 10 different sets of starting values in our simulations. Since for unimodal distributions, the difference between mean and mode is at most $\sqrt{3}$ standard deviations,[31] we propose the starting values

$$\mathbf{z}_i^\top \hat{\boldsymbol{\vartheta}}(c) = \mathbf{z}_i^\top \hat{\boldsymbol{\vartheta}}_{[LS]} + c\hat{\sigma},$$

where $\hat{\sigma}^2 = \sum_{i=1}^{n} v_i(y_i - \mathbf{z}_i^\top \hat{\boldsymbol{\vartheta}}_{[LS]})^2/(n-p)$ and c is chosen from a regular grid between $-\sqrt{3}$ and $\sqrt{3}$. Based on the 10 different starting values, regression models are fit and the model with the highest value for the objective function is selected. Then, hyperparameters are recalculated and the final model is fit.

We tested the computation time of the algorithm on our test system (CPU: AMD Ryzen 5 3600). Without P-splines, median computation time was 2.0 s (Q1: 1.7, Q3: 2.5) for 200 observations, 7.4 s (6.8, 8.2) for 1000 observations and 91.0 s (85.8, 96.4) for 5000 observations. With P-splines, median computation time was 6.8 s (5.2, 9.8) for 200 observations, 21.6 s (17.9, 28.4) for 1000 observations and 201.1 s (174.3, 355.6) for 5000 observations. More details can be found in the Appendix.

The computationally most demanding step was the optimization of hyperparameters. We found the number of optimization steps to be right-skewed and in some cases the number of required steps was quite large. In the interest of computational efficiency, we stopped maximization after 200 iterations. Details on the number of required steps can also be found in the Appendix.

Algorithm 1. Pseudo code of the implementation

```
1  set $\hat{\lambda} = 1$, calculate $\boldsymbol{v}$ from a Kaplan-Meier estimate
2  set starting values $\boldsymbol{z}_i^\top \hat{\boldsymbol{\vartheta}}_{[LS]}$ from mean regression
3  set $\hat{h}$ = median($|\hat{\varepsilon}|/3$)
4  begin
5      nonlinear optimization of the pseudo AIC:
6      repeat
7          repeat
8              calculate weights $\mathbf{w}$
9              fit weighted GAM with fixed $\hat{\lambda}$
10         until convergence of $\hat{\boldsymbol{\vartheta}}$
11         update $\hat{h}$ and $\hat{\lambda}$
12     until convergence of $\hat{h}$ and $\hat{\lambda}$
13 end
14 begin
15     generate multiple sets of starting values from $\boldsymbol{z}_i^\top \hat{\boldsymbol{\vartheta}}_{[LS]}$
16     for each set of starting values do
17         repeat
18             calculate weights $\mathbf{w}$
19             fit weighted GAM with fixed $\hat{\lambda}$
20         until convergence of $\hat{\boldsymbol{\vartheta}}$
21     end
22     select estimate with the highest value of the objective function $J_{h,\lambda}(\boldsymbol{\vartheta})$
23 end
24 nonlinear optimization of $\hat{h}$ and $\hat{\lambda}$ as in 4-13
```

2.6 Confidence intervals

For the calculation of confidence intervals, we propose a nonparametric residual bootstrap.[32] It consists of the following steps in the case of uncensored data:

(1) Fit the mode regression model and calculate predictions $\mathbf{z}_1^\top \hat{\boldsymbol{\vartheta}}, \ldots, \mathbf{z}_n^\top \hat{\boldsymbol{\vartheta}}$ and residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ for every data point.

(2) Sample n times from the residuals with replacement, such that $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$ are generated. Adding the sample and the predictions generates a bootstrap sample $(\tilde{y}_i, \mathbf{z}_i)$ with

$$\tilde{y}_i = \mathbf{z}_i^\top \hat{\boldsymbol{\vartheta}} + \tilde{\varepsilon}_i.$$

(3) Bootstrap estimates $\tilde{\vartheta}$ are then calculated, using the bootstrap sample $(\tilde{y}_i, \mathbf{z}_i)$. The bandwidth is recalculated for the estimation.

(4) Steps (2) and (3) are repeated k times and bootstrap estimates $\bar{\boldsymbol{\vartheta}}_{1,\ldots,}\bar{\boldsymbol{\vartheta}}_k$ are calculated.

(5) Bounds of the $1 - \alpha$ pointwise confidence interval are obtained with the $\alpha/2$ and $(1 - \alpha/2)$ quantiles of the bootstrap estimates.

We propose the following modification in the censored case: Since IPC weights indicate that an observation represents multiple unobserved event times, we use the IPC weights for sampling $\tilde{\varepsilon}_j$ from the residuals. Each residual ε_i is sampled with probability $p_i = v_i / \sum_{k=1}^n v_k$. After generating the bootstrap sample, IPC weights $\tilde{v}_i$ are calculated using the Kaplan-Meier estimator of C used in the original fit. We define

$$\tilde{v}_i = \frac{1}{n} \frac{\delta_i}{\hat{P}(C_i > \tilde{y}_i)}.$$

The weights $\tilde{v}_i$ are then used for estimation instead of v_i.

3 Analysis of overall survival of pancreatic cancer patients

We performed a retrospective exploratory analysis of overall survival and associated variables for pancreatic cancer patients. The goal of the analysis was a comparison of estimated regression coefficients between mode regression and parametric accelerated failure time (AFT) models.

3.1 Data source

The analyzed data set consisted of pancreatic cancer patients treated at the certified transregional pancreatic cancer center of the University Clinic for Visceral Surgery of the Pius Hospital Oldenburg. The Pius Hospital is an acute care hospital in the north-west of Germany and the University Clinic for Visceral Surgery performs a total of about 3000 inpatient and about 500 outpatient surgeries annually. The university clinic maintains a registry with prospective data collection, where the analyzed data originated from. The pancreatic cancer database lists 542 patients treated between 2010 and 2020, 158 of whom were included. We included patients with pancreatic adenocarcinoma as a histologic diagnosis who underwent surgical resection. Adjuvant therapy, that is, as supportive treatment after surgery, was given in accordance with standard guidelines.[33]

The data that support the findings of this analysis are available on request from the corresponding author. The data are not publicly available due to privacy/ethical restrictions. The Pius-Hospital is an institution of the Catholic Church. In accordance with the Law on Data Protection of the Catholic Church in Germany, no ethics approval is needed for the retrospective analysis of anonymized data. Patients gave their informed consent for their data to be entered into the prospectively maintained cancer registry database of the Pius-Hospital Oldenburg, and for their data to be used in retrospective analyzes.

3.2 Variables

The endpoint was overall survival measured since diagnosis. Chemotherapy was dichotomized (yes/no). Contraindications (yes/no) prevented some patients from receiving chemotherapy. pT, pN, pM, and R status were determined from pathological analysis of tissue removed during surgery. The UICC classification[34] of pancreatic cancer was used to classify the tumors in stages I-IV. Lymph node ratio (LNR) is defined as the number of affected lymph nodes divided by the total number of examined lymph nodes. R status refers to the presence of residual tumor after surgery. Microscopic or macroscopic residual tumor at the edge of the resection is classified as R1 and R2, respectively. R0 (no residual tumor) can be further grouped by the distance between the resection margin and the tumor tissue. The circumferential resection margin (CRM) is negative if the distance is more than 1mm and positive else. Finally, patients who already suffered from other cancers at some point in their lives are described as having pre-existing cancer.

Table 1. Descriptive statistics of the study population. Numbers are absolute (relative) frequencies, unless stated otherwise.

	Chemotherapy		
	No (*n*=52)	Yes (*n*=106)	Total (*n*=158)
Sex			
Female	25 (48.1)	47 (44.3)	72 (45.5)
Male	27 (51.9)	59 (55.7)	86 (54.4)
Mean age (SD)	73.4 (7.8)	64.8 (9.9)	67.7 (10.1)
UICC stage			
I	8 (15.4)	8 (7.5)	16 (10.1)
II	35 (67.3)	78 (73.6)	113 (71.5)
III	5 (9.6)	12 (11.3)	17 (10.8)
IV	4 (7.7)	8 (7.5)	12 (7.6)
Primary tumor			
pT1	2 (3.8)	7 (6.6)	9 (5.7)
pT2	16 (30.8)	13 (12.3)	29 (18.4)
pT3	32 (61.5)	82 (77.4)	114 (72.2)
pT4	2 (3.8)	4 (3.8)	6 (3.8)
Regional lymph nodes			
pN0	16 (30.8)	23 (21.7)	39 (24.7)
pN1	30 (57.7)	73 (68.9)	103 (65.2)
pN2	6 (11.5)	10 (9.4)	16 (10.1)
Remote metastases			
pM0	48 (92.3)	98 (92.5)	146 (92.4)
pM1	4 (7.7)	8 (7.5)	12 (7.6)
Median lymph node ratio (Q1, Q3)	0.06 (0.00, 0.23)	0.11 (0.04, 0.27)	0.11 (0.03, 0.25)
Resection margin			
R0	35 (67.3)	74 (69.8)	109 (69.0)
R1	11 (21.2)	21 (19.8)	32 (20.3)
R2	6 (11.5)	11 (10.4)	17 (10.8)
Circumferential resection margin			
Positive	34 (65.4)	81 (76.4)	115 (72.8)
Negative	18 (34.6)	25 (23.6)	43 (27.2)
Pre-existing cancer	11 (21.2)	25 (23.6)	36 (22.8)

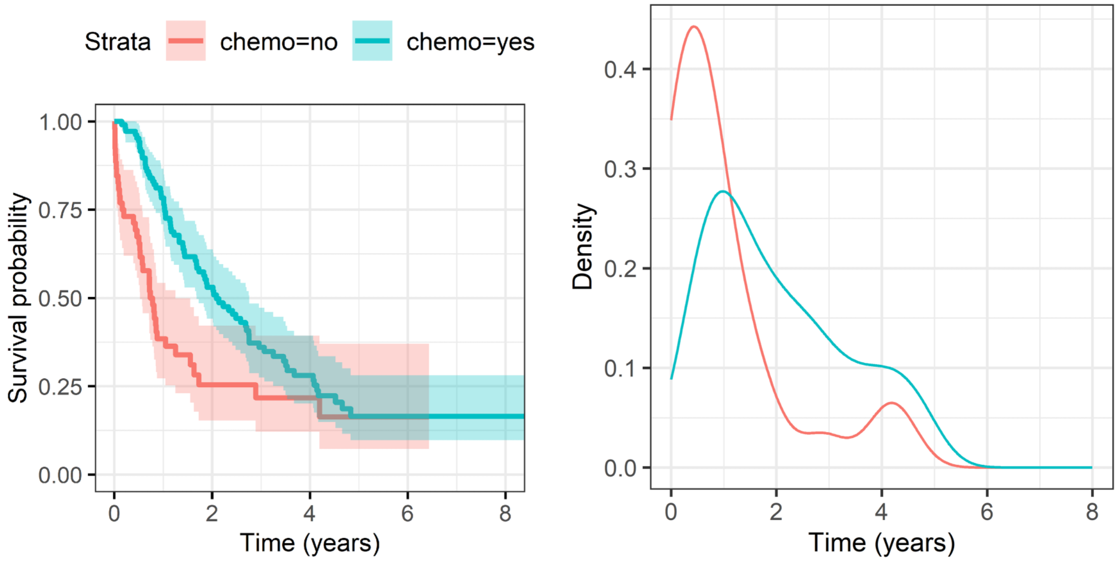

Figure 1. Kaplan-Meier estimates stratified by chemotherapy with pointwise confidence interval (left) and smoothed density estimates (right) of survival times.

3.3 Descriptive statistics

Baseline characteristics are shown in Table 1. Around two-thirds of patients received chemotherapy, in nine cases (6%) combined with radiotherapy. Most patients (113, 71.5%) had stage II pancreatic cancer. Survival times were censored in 43 cases (27%). Using reverse Kaplan-Meier estimation[35] (potential follow-up is censored by death), median potential follow-up was estimated to be 4.4 years (Q1: 2.8, Q3: 5.7). The Kaplan-Meier estimates of the survival times stratified by chemotherapy and the kernel smoothed densities based on the Kaplan-Meier estimates are shown in Figure 1. The bivariate analysis showed a positive association between chemotherapy and overall survival. Median survival was estimated at around 2.2 years (1.1, 4.1) for patients with chemotherapy treatment and 0.9 years (0.3, 2.5) without. The positive association with chemotherapy was particularly evident during the first year. Also, many patients died within the first year: estimated one-year survival was 77% with chemotherapy and 45% without. However, some patients survived a long time, the longest survival time was censored at 7.9 years. The density functions on the right side of Figure 1 show moderate and heavy skewness (1.1 with chemotherapy, 3.5 without), as well as signs of a second mode after 4 years of survival. Thus, there might be meaningful differences between results from mean, median and mode regression, suggesting practical relevance of mode regression for this group of patients.

3.4 Model selection

We fit a mode regression model and, for comparison, an AFT model assuming a parametric distribution as well as identically distributed residuals.

We began variable selection by including chemotherapy and LNR in our model, as they have been found to be the best predictors for long-term survival for pancreatic ductal adenocarcinoma.[36] It should be noted that for our observational data, the association between survival times and chemotherapy should be interpreted with care since treatment with chemotherapy might be subject to channeling bias. Treatment recommendations are usually given through an interdisciplinary tumor board. Various factors like expected survival chances, success of surgery, contraindications, general health and age might be considered. Boxplots for age in relation to chemotherapy and R status are depicted in Figure 2. 75% of patients without chemotherapy were 70 or older, while it were only 37% of patients with chemotherapy. Because of the potential confounding effect of age, we included it in our model as well and examined an interaction term after variable selection. In 16 cases, chemotherapy was not used because of contraindications. The S3 guidelines of the Association of the Scientific Medical Societies[33] define these as, for example, limited autonomy (ECOG score of 3 or higher), uncontrolled infections, cirrhosis of the liver (Child-Pugh score B or C) or severe coronary heart disease. Patients with contraindications were included in the analysis.

Inclusion of all other variables was determined by a best subset selection approach using the pseudo AIC defined in equation (4) as criterion. The model with the lowest pseudo AIC included an R status (residual tumor after surgery) in addition to chemotherapy, age and LNR. After variable selection, we added an interaction term between chemotherapy and age, but elected not to include the term in the final model due to fairly high standard errors for both the mode regression and AFT model. At this point, we compared the linear model with models using P-splines for age and LNR. The model with linear associations had the lowest pseudo AIC. However, we also show results from a model using P-splines to estimate the association with age to illustrate our method in more detail.

3.5 Results from multivariate analysis

Parameter estimates from the fitted models can be found in Table 2, along with estimates from a parametric AFT model, using the same variables. The Weibull distribution was assumed since it showed the best parametric fit (determined by graphical inspection of residuals). The estimated nonlinear association between age and survival is shown in Figure 3.

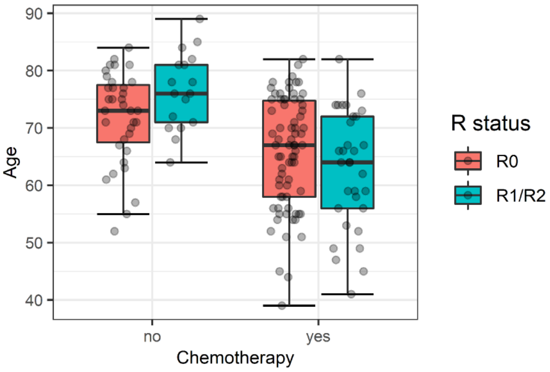

Figure 2. Boxplots of patients age seperated by chemotherapy and R status.

Table 2. Linear parameter estimates (95% confidence intervals) for mode regression and the Weibull model.

	Weibull model	Mode regression (linear)	Mode regression (nonlinear age)
Intercept	1.41 (−0.10 to 2.93)	−1.01 (−2.71 to 0.68)	−0.15 (−1.04 to 0.74)
Chemotherapy	0.79 (0.36 to 1.21)	1.14 (0.42 to 1.86)	1.21 (0.27 to 2.15)
LNR	−1.65 (−2.55 to −0.75)	−0.02 (−1.59 to 1.54)	0.11 (−1.68 to 1.90)
R1	−0.78 (−1.22 to −0.34)	−0.67 (−1.12 to −0.23)	−0.67 (−1.12 to −0.22)
R2	−1.24 (−1.76 to −0.72)	−0.89 (−1.39 to −0.39)	−0.66 (−1.19 to −0.13)
Age	−0.01 (−0.03 to 0.01)	0.01 (−0.01 to 0.04)	

LNR: lymph node ratio; R: resection margin.

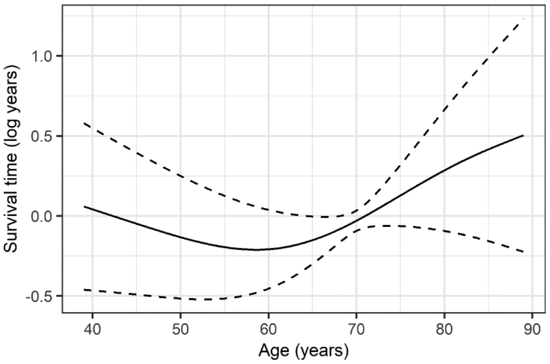

Figure 3. Estimate (solid line) and 95% confidence interval (dashed lines) for the relationship between age and mode of the log survival time.

The estimated association between chemotherapy and the log survival time was positive and smaller for the Weibull model (0.79, 95% CI: 0.36–1.21) in comparison to linear mode regression (1.14, 0.99–2.37). Point estimates of the linear mode regression model were fairly close to the estimates for the nonlinear model for all variables except age. Microscopic and macroscopic residues (R Status) were found to be associated with shorter survival time for all models, but more so for the Weibull model. The nonlinear association between age and log survival time was slightly negative up to 59 years and positive after. In comparison, no association was estimated between age and log survival time for linear mode regression. LNR had a negative association with survival for the AFT model, but no association was found between survival and LNR for mode regression.

As shown in this analysis, the estimates from mode regression and AFT models can differ substantially. However, differing estimates are not too surprising, first, because the survival time might be multimodal and violate the Weibull assumption and, second, the estimated parameters have different interpretations. For mode regression, the estimates relate to changes of the modal log survival time. Receiving chemotherapy was estimated to increase the modal log survival time by 1.14 log years. For the AFT model, chemotherapy was estimated to increase the log survival time by 0.79 log years. The AFT model estimate can be interpreted as a location shift for the whole distribution. The differing estimates for chemotherapy are also in line with the bivariate Kaplan-Meier estimates, showing a stronger association for patients dying within the first year and a weaker association for patients with longer survival times. While a constant location shift is assumed for the AFT model, mode regression is able to detect associations that are specific for a patient dying within the first year.

4 Numerical studies

4.1 Simulation setup

We compared our proposed mode regression method with AFT and GAMLSS models as well as our bandwidth selection with the plug-in bandwidth approach in a simulation study. We aimed to study the properties of our estimator in realistic finite sample scenarios. We started with a base scenario derived from the data and results of the analysis of pancreatic cancer patients and generated different scenarios through modifications of the base case, like the amount of censoring, signal-to-noise ratios and choice of event-time distributions.

The base scenario was developed using AFT models of our pancreatic cancer data as starting points. For the censoring time of overall survival, we found the log-logistic AFT model to be the best parametric fit (determined by graphical inspection of residuals). The censoring time was estimated with

$$\log(C) = 1.39 + 0.39e,$$

where $e \sim$ Logistic(0, 1). Simulation of the event time was based on the Weibull AFT model given in Table 2:

- log(T) was Gumbel distributed with a scale of 0.97.
- The signal-to-noise ratio was 30%.

We chose one binary and one metric nonlinear covariate for the base model. Coefficients were chosen in agreement with the observed signal-to-noise ratio. An intercept was chosen such that the share of censoring was 27%, as observed in the pancreatic cancer data. Finally, we also added heteroscedasticity, such that the scale was increased or decreased by up to one-third, depending on the values of the covariates. The resulting event time model was

$$\log(T) = 0.46 + 1.04x_1 + f_{age}(x_2) + \sigma(x_1, x_2)\varepsilon,$$

$$\sigma(x_1, x_2) = 0.97 - 0.32x_1 + 0.32x_2^2,$$

with $x_1 \sim$ Bin(1, 0.5), $x_2 \sim U(0, 1)$ and $\varepsilon \sim$ Gumbel(0, 1). $f_{age}(x_2)$ was a fifth degree polynomial modeled after the estimated association between age and the mode of the log survival time (Figure 3).

In addition to the base scenario, we simulated several scenarios with variations from the base case. An overview is given in Table 3. Share of censoring was varied (50% and uncensored data), as well as the signal-to-noise ratio (15% and 60%). Covariate influence on C was introduced, violating an assumption of our model. Linear trends were used instead of nonlinear trends. We also replaced the Gumbel distribution with a logistic distribution while retaining the variance. Each scenario was simulated with 200 and 500 observations and 200 repetitions.

We compared our proposed mode regression approach with mode regression using plug-in bandwidth selection described in Yao and Li[16] (only for uncensored data), with two GAMLSS and one parametric AFT model. In short, GAMLSS is a parametric model which allows modeling parameters other than just the location. In this case, the GAMLSS models assumed covariate influence on the variance, while for the AFT model a constant variance was assumed. For the first GAMLSS model, we assumed a Weibull distribution for the event time (only correct for scenarios 1–8), the second incorrectly assumed a log-normal distribution. A Weibull distribution was also used for the AFT model. All models assumed the proper linear relationships in scenarios 8 and 9 and used P-splines in all appropriate cases (nonlinear trends in scenarios 1–7, estimation of the scale).

To evaluate point estimates, predictions were made for observations on a regular grid with 100 points between zero and one for x_2 and for $x_1 = 0$ and $x_1 = 1$. Bias and mean squared error (MSE) were calculated. Bias and MSE for each repetition were calculated as the average difference and average squared difference between predicted modes $\hat{m}(x_1, x_2)$ and true modes $m(x_1, x_2)$, that is

$$\text{Bias} = \frac{1}{200}\sum_{i=1}^{200} \hat{m}(x_{1i}, x_{2i}) - m(x_{1i}, x_{2i}),$$

$$\text{MSE} = \frac{1}{200}\sum_{i=1}^{200} \left[\hat{m}(x_{1i}, x_{2i}) - m(x_{1i}, x_{2i})\right]^2.$$

Table 3. Summary of the simulation setup and variations.

Number	Scenario	Variation
1	Base scenario	–
2	0% censoring	$\log(C) = \infty$
3	50% censoring	Log event time intercept of 1.31
4	15% signal-to-noise ratio	$\sigma(x_1, x_2) = \sqrt{2}(0.97 - 0.32x_1 + 0.32x_2^2)$
5	60% signal-to-noise ratio	$\sigma(x_1, x_2) = \sqrt{0.5}(0.97 - 0.32x_1 + 0.32x_2^2)$
6	Covariate influence on *C*	$\log(C) = 1.39 + x_1 - x_2 + 0.39e$
7	Linear trend	$\log(T) = 1.25 + 1.04x_1 - 1.8x_2 + \sigma(x_1, x_2)\varepsilon$
8	Linear trend, 0% censoring	$\log(T) = 1.25 + 1.04x_1 - 1.8x_2 + \sigma(x_1, x_2)\varepsilon$, $\log(C) = \infty$
9	Different event time distribution	$\varepsilon \sim$ Logistic$(0, 1/\sqrt{2})$

Table 4. Bias, mean squared error (MSE) and mean empirical coverage (Cov.) in different simulation scenarios.

		Mode regression			Wb. GAMLSS		LN. GAMLSS		Wb. AFT	
Scenario	Sample size	Bias	MSE	Cov.	Bias	MSE	Bias	MSE	Bias	MSE
1	200	−0.17	0.17	90.0	−0.01	0.03	−0.45	0.26	0.01	0.04
	500	−0.14	0.10	93.2	−0.01	0.02	−0.45	0.23	0.01	0.02
2	200	−0.14	0.11	83.6	−0.01	0.02	−0.53	0.32	0.00	0.03
	500	−0.09	0.06	82.0	0.00	0.01	−0.53	0.30	0.01	0.02
3	200	−0.36	0.41	71.3	−0.01	0.07	−0.35	0.23	0.04	0.08
	500	−0.26	0.24	77.6	−0.02	0.03	−0.35	0.17	0.04	0.04
4	200	−0.28	0.34	89.8	−0.02	0.07	−0.63	0.51	0.01	0.07
	500	−0.21	0.19	92.7	−0.01	0.03	−0.61	0.43	0.02	0.03
5	200	−0.14	0.10	89.0	−0.01	0.02	−0.33	0.14	0.00	0.02
	500	−0.09	0.05	90.0	−0.01	0.01	−0.33	0.13	0.01	0.01
6	200	−0.33	0.32	70.7	−0.02	0.04	−0.44	0.25	−0.04	0.04
	500	−0.30	0.25	67.6	−0.01	0.02	−0.44	0.22	−0.03	0.02
7	200	−0.18	0.11	92.5	0.00	0.02	−0.45	0.25	0.05	0.03
	500	−0.13	0.06	95.2	0.00	0.01	−0.45	0.23	0.05	0.01
8	200	−0.12	0.08	86.3	0.00	0.02	−0.52	0.30	0.01	0.02
	500	−0.08	0.05	88.2	0.00	0.01	−0.53	0.30	0.00	0.01
9	200	0.00	0.20	92.7	0.40	0.23	0.00	0.05	0.45	0.26
	500	0.00	0.11	97.7	0.41	0.20	0.00	0.02	0.46	0.24

Wb.: Weibull; LN.: log-normal.

Bias and MSE were then averaged over the 200 repetitions. 95% confidence intervals for the predictions were constructed through residual bootstrap. For each point on the grid, the bootstrap percentile interval was used. Empirical coverage was measured by the proportion of cases in which the true mode was within the bounds of the interval.

We used the software R 4.0.2[29] for all simulations. The package GAMLSS[11] was used for the parametric models. Our implementation of mode regression has been published in a dedicated R package,[37] our simulation code is available on request. Random seeds were used to generate pseudo random numbers and saved to allow reproducibility.

4.2 Results

Our simulation results are shown in Table 4. In general, the results showed good performance of mode regression. Bias and MSE were for the most part within the bounds of the parametric GAMLSS models with correct and partially incorrect assumptions. Further, bias and MSE were reduced with increasing sample size, indicating possible consistency. The empirical coverage of the estimates did not reach the desired confidence level but was close to the level in many scenarios. Further details are given below.

Mode regression had a negative bias in every scenario except one and was always reduced with increasing sample size. In the base case (scenario 1), the bias was −0.17 for mode regression, −0.45 for log-normal GAMLSS and, as expected, close to zero for the Weibull GAMLSS and Weibull AFT model. Increasing the sample size to 500 reduced the bias to −0.14 for mode regression, while the bias for the log-normal GAMLSS was unaffected. The bias for the parametric models was lower than for mode regression if the assumed distribution was correct and larger for incorrect distributions, except in scenario 9 (log-logistic distribution), where the log-normal assumption lead to low bias as well. The Weibull GAMLSS and Weibull AFT models showed close results in all scenarios.

The MSE showed similar trends as the bias. In most scenarios, mode regression had a larger MSE than the parametric model with the correct distribution and lower MSE than the model with the incorrect distribution. However, the MSE of mode regression was closer to the incorrect GAMLSS model than the bias. In the base case, the MSE was 35% and 57% lower than the MSE of the log-normal GAMLSS for n = 200 and n = 500. In comparison, the bias was 62% and 69% lower. The poorest results for mode regression were in scenario 3 (50% censoring) and scenario 6 (covariate influence on censoring). Bias was close to the log-normal GAMLSS and MSE was larger than the MSE of the parametric models. Bias and MSE were also reduced with increasing sample size in both scenarios. The MSE was close to the log-normal GAMLSS for 500 observations.

As a comparison, we also performed mode regression with the plug-in bandwidth approach in scenario 8 (uncensored data, linear trends). Bandwidth selection with the pseudo-likelihood and the plug-in approach performed equally well, the

MSE was 0.09 for the plug-in approach at 200 observations, at 500 observations, it was 0.05. In comparison, the MSE of the GAMLSS model with correct assumptions was between 0.02 and 0.01 for 200 and 500 observations. The bias of the pseudo-likelihood and plug-in approach were also similar to each other.

The residual bootstrap resulted in empirical coverages below 95%. However, coverages were above 85% in most scenarios. Scenario 2 (uncensored data, nonlinear trends), scenario 3 (50% censoring) and scenario 6 (covariate influence on censoring) had the lowest coverage. For scenarios 3 and 6, this was in agreement with the reported elevated bias and MSE.

5 Discussion

We extended mode regression with IPC weights and semiparametric predictors, resulting in a flexible method for time-to-event analysis. Mode regression allows us to estimate the conditional mode, which is a more appropriate location measure for heavily skewed or multimodal data than the mean or median. The conditional mode might also be a more intuitive measure than the hazard ratio, which has been criticized for being less appealing for patients and doctors than direct interpretation of the survival time.[38] While direct interpretation of the survival time is quite intuitive, similar to acceleration factors in AFT models, we have avoided direct interpretation of the raw survival time in our application and focused on the log-survival time instead. That is because transformations of the estimates to the raw survival time require additional information or assumptions. A partial remedy might be the use of the softplus response function, a combination of the exponential response function and the identity, as proposed by Wiemann and Kneib.[39] The transformation assumes a multiplicative association for shorter survival times and a smooth transition to a linear association after some specified time. While the multiplicative part of the survival time would still be hard to interpret, changes after the transition to the identity could be interpreted linearly and a transformation would be unnecessary.

The main caveats of other conventional methods for survival analysis are, however, their basic assumptions. This could be the expected proportionality of hazards or the choice of a parametric distribution for the response. Violations of these assumptions can lead to biased results, for example, the average hazard ratio is underestimated for converging hazard rates when the proportional-hazards model is used.[40] By using mode regression, we are able to circumvent these assumptions and are therefore in less danger of obtaining biased results.

The choice to use mode regression should depend mostly on the research question. For pancreatic cancer, mode regression is suitable for the special research interest in patients who die within the first year. Results from our analysis of pancreatic cancer suggested a stronger association of chemotherapy with the mode of overall survival than the association estimated in a parametric AFT model.

In our simulation study, bias and MSE were shown to be mostly bound between parametric models with correct and incorrect distributional assumptions. We have seen a strong difference between the two in such a way that a correct selection of the distribution of the response is crucial for the usability of the corresponding regression coefficients. While mode regression never outperforms a parametric model with correct assumptions, its results are far closer to a correctly assumed model than to a parametric model with false assumptions. The simulations further showed that bias and MSE were reduced with increasing sample size. Although residuals were not identically distributed in our simulations, we used a residual bootstrap for calculation of confidence intervals. Empirical coverages were below 95%, but above 85% in most scenarios. The scenarios with covariate influence on the censoring time and a larger share of censoring had elevated bias and MSE and the empirical coverage was much lower than in other scenarios. For heavy censoring, sample sizes larger than 500 might be necessary. For uncensored data, results from mode regression with the pseudo-likelihood were similar to the plug-in approach, which has been shown to be consistent.[16]

A further area that would benefit from additional research is model selection. While we used the pseudo AIC to perform model selection in our application to pancreatic cancer survival and the chosen variables agreed with our general intuitions about the variables, performance of model selection criteria should be studied in more detail.

Acknowledgements

We thank Thomas Kneib for helpful discussions and exchange of ideas. We also would like to thank two anonymous referees for their valuable comments that led to considerable improvements of this article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

We gratefully acknowledge financial support from the German Research Foundation (DFG) grant SO1313/1-1 project "Distributional Regression for Time-to-Event Data."

ORCID iD

Alexander Seipp https://orcid.org/0000-0002-1496-4818

References

1. Hyndman RJ. Computing and graphing highest density regions. *Am Stat* 1995; **50**: 120–126.
2. Robert Koch-Institut and Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. Krebs in Deutschland 2015/2016 [Cancer in Germany 2015/2016], 2019.DOI: 10.25646/5977.2.
3. Sasaki H, Ono Y and Sugiyama M. Modal regression via direct log-density derivative estimation. In Hirose A, Ozawa S, Doya K et al. (eds.) *Neural Information Processing*. Springer International Publishing. ISBN 978-3-319-46672-9, pp. 108–116.
4. Ota H, Kato K and Hara S. Quantile regression approach to conditional mode estimation. *Electron J Stat* 2019; **13**: 3120–3160.
5. Chen YC. Modal regression using kernel density estimation: a review. *WIREs Comput Stat* 2018; **10**: e1431.
6. Chen YC, Genovese CR, Tibshirani RJ et al. Nonparametric modal regression. *Ann Stat* 2016; **44**: 489–514.
7. Lee MJ. Mode regression. *J Econom* 1989; **42**: 337–349.
8. Kemp GC and Santos Silva J. Regression towards the mode. *J Econom* 2012; **170**: 92–101.
9. Koenker K and Bassett G. Regression quantiles. *Econometrica* 1978; **46**: 33–50.
10. Newey WK and Powell JL. Asymmetric least squares estimation and testing. *Econometrica* 1987; **55**: 819–847.
11. Rigby RA and Stasinopoulos DM. Generalized additive models for location, scale and shape. *Appl Statist* 2005; **54**: 507–554.
12. Eilers PHC and Marx BD. Flexible smoothing with B-splines and penalties. *Stat Sci* 1996; **11**: 89–102.
13. Khardani S, Lemdani M and Ould Sad E. Uniform rate of strong consistency for a smooth kernel estimator of the conditional mode for censored time series. *J Stat Plan Inference* 2011; **141**: 3426–3436.
14. Seipp A, Uslar V, Weyhe D et al. Weighted expectile regression for right-censored data. *Stat Med* 2021; **40**: 5501–5520.
15. Heidenreich NB, Schindler A and Sperlich S. Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Adv Stat Anal* 2013; **97**: 403–433.
16. Yao W and Li L. A new regression model: modal linear regression. *Scand J Stat* 2014; **41**: 656–671.
17. Chacón JE. The modal age of statistics. *Int Stat Rev* 2020; **88**: 122–141.
18. Parzen E. On estimation of a probability density function and mode. *Ann Math Statist* 1962; **33**: 1065–1076.
19. Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958; **53**: 457–481.
20. Peterson AV. Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *J Am Stat Assoc* 1977; **72**: 854–858.
21. Johansen S. The product limit estimator as maximum likelihood estimator. *Scand J Stat* 1978; **5**: 195–199. http://www.jstor.org/stable/4615715.
22. Kim C, Park BU, Kim W et al. Bezier curve smoothing of the Kaplan-Meier estimator. *Ann Inst Stat Math* 2003; **55**: 359–367.
23. Földes A, Rejto L and Winter B. Strong consistency properties of nonparametric estimators for randomly censored data, ii: Estimation of density and failure rate. *Period Math Hung* 1981; **12**: 15–29.
24. Huang J, Ma S and Xie H. Least absolute deviations estimation for the accelerated failure time model. *Stat Sin* 2007; **17**: 1533–1548. http://www.jstor.org/stable/24307687.
25. Zhou M. M-estimation in censored linear models. *Biometrika* 1992; **79**: 837–841.
26. Fahrmeir L, Kneib T and Lang S. Penalized structured additive regression for space-time data: a Bayesian perspective. *Stat Sin* 2004; **14**: 731–761. http://www.jstor.org/stable/24307414.
27. Duin RPW. On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans Comput* 1976; **C-25**: 1175–1179.
28. Wood S. *Generalized additive models: an introduction with R*. edition2 ed. New York, NY: Chapman and Hall/CRC, 2017.
29. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. https://www.R-project.org/.
30. Johnson SG. *The NLopt nonlinear-optimization package*, 2020. R package version 1.2.2.2, http://github.com/stevengj/nlopt.
31. Johnson NL and Rogers CA. The moment problem for unimodal distributions. *Ann Math Statist* 1951; **22**: 433–439.
32. Efron B. *The jackknife, the bootstrap and other resampling plans*. Philadelphia, PA: SIAM, 1982.
33. Seufferlein T, Porzner M, Becker T et al. S3-guideline exocrine pancreatic cancer. *Z Gastroenterol* 2013; **51**: 1395–1440.
34. Brierley JD, Gospodarowicz MK and Wittekind C eds. *TNM classification of malignant tumours*. 8 ed. Oxford, UK: John Wiley & Sons, 2016.
35. Schemper M and Smith TL. A note on quantifying follow-up in studies of failure time. *Control Clin Trials* 1996; **17**: 343–346.
36. Weyhe D, Obonyo D, Uslar VN et al. Predictive factors for long-term survival after surgery for pancreatic ductal adenocarcinoma: making a case for standardized reporting of the resection margin using certified cancer center data. *PLoS ONE* 2021; **16**: e0248633.
37. Seipp A and Otto-Sobotka F. *dirttee: Distributional regression for time to event data*, 2022. R package version 1.0, https://CRAN.R-project.org/package=dirttee.

38. Chen YQ, Jewell NP, Lei X et al. Semiparametric estimation of proportional mean residual life model in presence of censoring. *Biometrics* 2005; **61**: 170–178.

39. Wiemann P and Kneib T. Using the softplus function to construct alternative link functions in generalized linear models and beyond, 2021. Working paper.

40. Schemper M. Cox analysis of survival data with non-proportional hazard functions. *Statistician* 1992; **41**: 455–465.

Appendix

A.1 Convergence of nonlinear mode regression

We show convergence of the estimator for the nonlinear and right-censored case if bandwidth h and smoothing parameter λ are fixed. This is a modified version of the proof given in Yao and Li.[16] To simplify notation we omit dependence of $\hat{\boldsymbol{\vartheta}}$ on h and λ.

Through transformations we get the following:

$$\begin{aligned}
&\log\left(\sum_{i=1}^{n} v_i\phi_h\left(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k+1]}\right)\right) - \log\left(\sum_{i=1}^{n} v_i\phi_h\left(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k]}\right)\right)\\
&= \log\left(\sum_{i=1}^{n} \frac{v_i\phi_h\left(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k+1]}\right)}{\sum_{i=1}^{n} v_i\phi_h\left(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k]}\right)}\right)\\
&= \log\left(\sum_{i=1}^{n} \omega_i\left(\hat{\boldsymbol{\vartheta}}^{[k]}, h\right)2h^2\frac{\phi_h\left(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k+1]}\right)}{\phi_h\left(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k]}\right)}\right)
\end{aligned}$$

Applying Jensen's inequality, we obtain

$$\begin{aligned}
&J_{h,\lambda}\left(\hat{\boldsymbol{\vartheta}}^{[k+1]}\right) - J_{h,\lambda}\left(\hat{\boldsymbol{\vartheta}}^{[k]}\right)\\
&= \log\left(\sum_{i=1}^{n} v_i\phi_h(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k+1]})\right) - \log\left(\sum_{i=1}^{n} v_i\phi_h(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k]})\right) - \hat{\boldsymbol{\vartheta}}^{[k+1]^\top}\mathbf{P}(\boldsymbol{\lambda})\hat{\boldsymbol{\vartheta}}^{[k+1]} + \hat{\boldsymbol{\vartheta}}^{[k]^\top}\mathbf{P}(\boldsymbol{\lambda})\hat{\boldsymbol{\vartheta}}^{[k]}\\
&\geq \sum_{i=1}^{n} \omega_i\left(\hat{\boldsymbol{\vartheta}}^{[k]}, h\right)2h^2\log\left(\frac{\phi_h(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k+1]})}{\phi_h(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k]})}\right) - \hat{\boldsymbol{\vartheta}}^{[k+1]^\top}\mathbf{P}(\boldsymbol{\lambda})\hat{\boldsymbol{\vartheta}}^{[k+1]} + \hat{\boldsymbol{\vartheta}}^{[k]^\top}\mathbf{P}(\boldsymbol{\lambda})\hat{\boldsymbol{\vartheta}}^{[k]}\\
&= -\left(\sum_{i=1}^{n} \omega_i\left(\hat{\boldsymbol{\vartheta}}^{[k]}, h\right)(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k+1]})^2 + \hat{\boldsymbol{\vartheta}}^{[k+1]^\top}\mathbf{P}(\boldsymbol{\lambda})\hat{\boldsymbol{\vartheta}}^{[k+1]}\right)\\
&\quad + \sum_{i=1}^{n} \omega_i\left(\hat{\boldsymbol{\vartheta}}^{[k]}, h\right)(y_i - \mathbf{z}_i^\top\hat{\boldsymbol{\vartheta}}^{[k]})^2 + \hat{\boldsymbol{\vartheta}}^{[k]^\top}\mathbf{P}(\boldsymbol{\lambda})\hat{\boldsymbol{\vartheta}}^{[k]} \geq 0,
\end{aligned}$$

which shows convergence of the algorithm.

A.2 Computation time and hyperparameter optimization

We tested the computation time of our algorithm on our test system for five different sample sizes. The computation times are shown in Figure 4(a). Because of long computation times in some cases, we chose a stopping criterion for the number of steps for hyperparameter optimization. We chose a maximum number of 200 steps as stopping criterion. For the objective value, we chose a difference of 0.001 in objective values of consecutive optimization steps as stopping criterion. As an example, we focus on scenario 1 with 200 observations. In scenario 1, there were three cases that did not converge within 200 steps. They are shown in Figure 4(b). While case 1 benefits from additional optimization beyond 200 steps, cases 2 and 3 converge shortly after 200 steps.

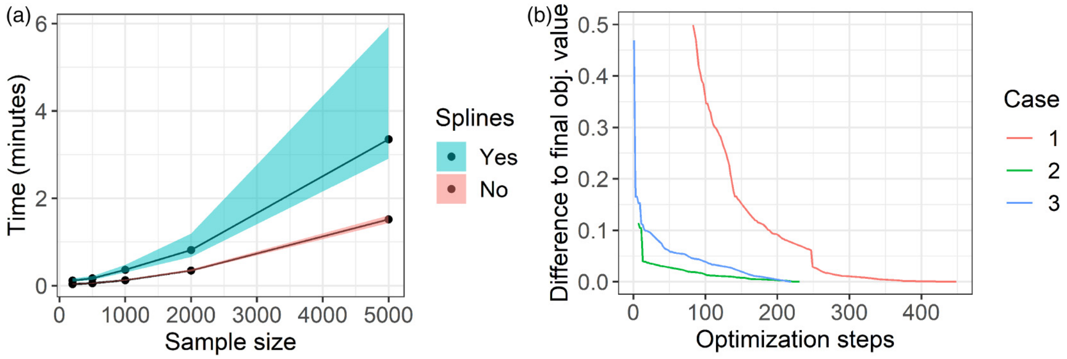

Figure 4. Further simulation results on the convergence of the algorithm. (a) Computation time of the algorithm in dependence of sample size, with either linear effects or with P-splines. Points and lines represent the median, the shaded areas are intervals from Q1 to Q3. (b) Progression of objective values in three cases where the hyperparameter optimization did not converge within 200 optimization steps.