Gumbel regression models for a monotone increasing continuous biomarker subject to measurement error

Noorie Hyun [a,*], David J. Couper [b], Donglin Zeng [b]

[a] *Biostatistics, Institute of Health and Equity, Medical College of Wisconsin, Milwaukee, Wisconsin, United States*
[b] *Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States*

ARTICLE INFO

Article history:
Received 31 January 2018
Received in revised form 10 September 2018
Accepted 17 March 2019
Available online 21 March 2019

Keywords:
Gumbel distribution
Semiparametric regression models
Generalized linear model
Profile likelihood

ABSTRACT

Right- or left-skewed continuous biomarker measurements are often confronted in practice. The extreme values located in the long tail of skewed distributions indicate abnormal health status, and characterizing the features of a subgroup with abnormally high(or low) values is important for risk management. Furthermore, a biomarker is usually subject to measurement error. In the presence of these two challenging issues, existing methods for analyzing skewed biomarker data are no longer applicable. In this paper, we propose a semiparametric method based on Gumbel distributions while accounting for measurement error in a biomarker and adjusting for a monotone age effect on biomarker levels. We estimate the model parameters using the nonparametric maximum likelihood approach and implement computation via the EM algorithm. We establish the asymptotic properties of the proposed estimators and summarize simulation results to assess the numerical performance of the proposed method. The method is illustrated through an application to data from a diabetes ancillary study to the Atherosclerosis Risk in Communities (ARIC) Study.

1. Introduction

Right- or left-skewed continuous biomarker measurements are often confronted in practice. For analyzing such skewed data, data transformation for approximating to a normal distribution or an application of generalized linear models with a gamma distribution can be commonly used approaches. However, for the case that data is seriously skewed, and extreme outliers are of interest, modeling based on skewed distributions has been proposed and studied in multiple fields, such as climatology, environmental science, biomedicine, and finance. For instance, risk analysis identifying a subgroup at high risk or analysis of extreme events occurring with small probability, such as trends in annual high or low temperatures has used extreme value theory.

Fisher and Tippett (1928) first identified the extreme value limit distribution and Gnedenko (1943) provided the asymptotic proof for the limit distribution of extreme values. The corresponding three types of distributions, Gumbel, Fréchet and Weibull are unified into a single generalized extreme value (GEV)-form (von Mises, 1936; Gumbel, 1958). An alternative approach to extreme events is based on exceedances above thresholds leading to the generalized pareto distribution (Balkema and De Haan, 1974; Pickands, 1975); however, the threshold methods are not covered in this paper. Huerta and Sansó (2007) modeled extreme values using GEV distributions with parameters varying in time. A link function based on GEV distributions was introduced for modeling imbalanced binary and ordinary responses

[*] Corresponding author.
E-mail address: nhyun@mcw.edu (N. Hyun).

https://doi.org/10.1016/j.jspi.2019.03.008

(Wang and Dey, 2010; Roy and Dey, 2014). Ghosh and Mallick (2011) proposed Bayesian hierarchical regression models for repeatedly measured continuous outcomes, which are marginally generated from a GEV distribution. Using the power-transformed quantile regression, Wang and Li (2013) developed a new three-stage estimation procedure, estimating intermediate conditional quantiles and extrapolating the estimates to tails. However, we believe that there has not been a study involving modeling continuous biomarker data based on Gumbel distributions, which are GEV distributions with shape parameter 0, while adjusting for a monotone age effect on biomarker values.

Furthermore, most biomarkers suffer from at least some measurement error: every assay has some inherent variability, so if the assay is run twice on a sample from an individual, the results may not be identical. Additionally, even though there may be a smooth underlying trend in an individual's levels of the biomarker, there is likely to be short-term intra-individual variability, resulting in variation around that underlying trend. Some biases may be more likely in a particular direction, such as with "white-coat hypertension", whereas others are essentially random. In clinical practice, *ad hoc* approaches that are used to take into account biomarker variability include taking two or more measurements over a period of time.

We propose new regression models for investigating associations between exposures and continuous biomarker while addressing abnormally high values and measurement error in a biomarker. We consider a monotone increasing biomarker over time (particularly age) and incorporate a function of time in the model. The proposed method involves a semi-parametric likelihood approach based on a mixture distribution of a normal and Gumbel distribution. We propose an efficient estimator for the model parameters based on nonparametric maximum likelihood estimation. In Section 2, inference procedures using the expectation–maximization algorithm for parameter estimation are presented, and variance estimation is also proposed. Asymptotic results based on the proposed model are established in Section 3. The detailed proofs are provided in the Supplementary material. A simulation study and an application to real data are illustrated in Section 4. Finally, in Section 5 we discuss limitations of our approach and related future research problems.

2. Method

2.1. Model

For subject i, let $\boldsymbol{x}_i$ be time-invariant covariates such as demographic characteristics and potential risk factors at baseline such as health status-related variables, and $y_i^*(t)$ and $y_i(t)$ be the true biomarker value and observed biomarker value at time t, respectively. The observation time, $T = t$ can be fixed or random. Thus, the observed data from n independent and identically distributed subjects are $\{y_i(t_i), t_i, \boldsymbol{x}_i \mid i = 1, \dots, n\}$, which are denoted by $\{\boldsymbol{w}_i \mid i = 1, \dots, n\}$ hereafter.

Normal distributions are commonly used for explaining continuous biomarker values; however, extreme values of a biomarker related to a trait are not well-captured by any symmetric distributions. To study the association between $\boldsymbol{x}_i$ and $y_i(t_i)$, of which the distribution has a long tail, we consider Gumbel distributions:

$$P(Y^*(t) < \xi \mid t, \boldsymbol{x}) = \exp[-\gamma(t)\exp\{-\mu(\xi - \boldsymbol{x}^T\boldsymbol{\beta}\mu^{-1})\}],$$

where ξ denotes true biomarker value $y^*(t)$ in the support $(-\infty, \infty)$; $\mu > 0$ and $\boldsymbol{\beta}$ are unknown parameters, and $\gamma(t)$ is an unknown positive monotone increasing function, that is, true biomarker values are monotone increasing over time. For instance, when $\gamma(t)$ is a function of age, it indicates monotone increasing age effect on biomarker values. When a covariate x value increases by 1 unit with given time t and biomarker value ξ, the cumulative density of a biomarker value decreases (increases) by factor of $\exp(\beta)$ for $\beta > 0$ ($\beta < 0$).

Our second model incorporates the measurement error in the observed biomarker. Specifically we adopt the additive measurement error model (Carroll, 2006; Fuller, 1987; Tsiatis et al., 1995)

$$y_i(t) = y_i^*(t) + \epsilon_i(t), \ \ i = 1, \dots, n. \tag{1}$$

We assume the measurement error $\epsilon_i(t)$ has the normal distribution with mean zero and variance σ^2 for any time t and is independent of $y_i^*(t)$ and $\boldsymbol{x}_i$. The measurement error variance of σ^2 may be estimated in practice by taking repeated measurements. As an example, the coefficient of measurement error variation for blood glucose value is reported to be 3.5% (Hadi et al., 2008; Young et al., 2008). Because information about the measurement error variation can be referred to external data, we consider σ^2 to be known.

Under the above two models, the observed biomarker $Y(t)$ has a mixture distribution of a Gumbel with a normal distribution for given t and $\boldsymbol{x}$. Then the observed likelihood function for $\{y_i(t_i), t_i, \boldsymbol{x}_i \mid i = 1, \dots, n\}$ concerning parameters $(\boldsymbol{\beta}^T, \mu, \gamma(\cdot))$ is

$$\prod_{i=1}^{n} \int_{-\infty}^{\infty} \exp(-\gamma(t_i)e^{\boldsymbol{x}_i^T\boldsymbol{\beta}-\mu\xi})\gamma(t_i)\mu\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} - \mu\xi)\frac{1}{\sigma}\phi\left(\frac{y_i(t_i) - \xi}{\sigma}\right)d\xi, \tag{2}$$

where $\phi(\cdot)$ is the standard normal density function.

2.2. Inference procedure

We maximize (2) to estimate all the parameters, including $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}^T)^T$ and $\gamma(\cdot)$. Specifically, we estimate γ as a step function, which jumps at the observed t_i's. Let $t_{(1)} < \cdots < t_{(K)}$ be ordered observed times of $\{t_i \mid i = 1, \ldots, n\}$ and $\gamma_k = \gamma(t_{(k)})$ and $t_{(0)} = 0$. Then we maximize (2) over $\boldsymbol{\theta}$ and γ_k's, subject to constraints $0 \le \gamma_1 \le \ldots \le \gamma_K$.

To facilitate the maximization, especially over $\{\gamma_k\}$, we employ the EM algorithm by treating the not observed true values ξ as missing data. Then the complete log-likelihood function is

$$l_c(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} I(t_i = t_{(k)}) \Big(-\gamma_k e^{\boldsymbol{x}_i^T \boldsymbol{\beta} - \mu\xi} + \log \gamma_k + \log \mu + \boldsymbol{x}_i^T \boldsymbol{\beta} \\ - \mu\xi - \frac{1}{2} \log \sigma^2 - \frac{(y_i(t_i) - \xi)^2}{2\sigma^2} \Big). \tag{3}$$

In the M-step at the lth iteration of the EM algorithm, we first maximize the conditional expectation of the complete log-likelihood function given observed data over γ_k's. We then update $\boldsymbol{\theta}$ via the Newton–Raphson algorithm. Specifically, we maximize $Q(\boldsymbol{\gamma})$ defined by

$$Q(\boldsymbol{\gamma}) = \sum_{k=1}^{K} \sum_{i=1}^{n} I(t_i = t_{(k)}) E(-\gamma_k e^{\boldsymbol{x}_i^T \boldsymbol{\beta} - \mu\xi} + \log \gamma_k \mid \boldsymbol{w}_i, \boldsymbol{\theta}^{(l)}). \tag{4}$$

Because $Q(\boldsymbol{\gamma})$ is a concave function over a convex cone satisfying $\gamma_1 \le \ldots \le \gamma_K$, this maximization can be carried out using one of the many existing algorithms for convex optimization. To update $\boldsymbol{\theta}$, we apply the following one-step Newton–Raphson algorithm,

$$\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)} + E\left(-\partial^2 l_c/(\partial\theta)^2 \middle| \boldsymbol{w}, \boldsymbol{\theta}^{(l)}\right)^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(l)}} E\left(\partial l_c/\partial\theta \middle| \boldsymbol{w}, \boldsymbol{\theta}^{(l)}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(l)}}. \tag{5}$$

The conditional expectations in (5) are calculated in the E-step of the EM algorithm based on the following expression:

$$E(g(\xi)|\boldsymbol{w}_i, \boldsymbol{\theta}^{(l)}) = \frac{I(t_i = t_{(k)}) \int_{-\infty}^{\infty} g(\xi) \exp[-\gamma_k e^{\boldsymbol{x}_i^T \boldsymbol{\beta} - \mu\xi}] e^{-\mu\xi} \phi\{\frac{y_i(t_i) - \xi}{\sigma}\} d\xi}{\int_{-\infty}^{\infty} \exp[-\gamma_k e^{\boldsymbol{x}_i^T \boldsymbol{\beta} - \mu\xi}] e^{-\mu\xi} \phi\{\frac{y_i(t_i) - \xi}{\sigma}\} d\xi}, \tag{6}$$

where the $g(\xi)$'s to be calculated are ξ, ξ^2, $e^{-\mu\xi}$, $e^{-\mu\xi}\xi$, and $e^{-\mu\xi}\xi^2$. This integration can be approximated by Gauss–Hermite quadrature (Davis, 1984), so it can be approximated by

$$\sum_{k=1}^{N} \Bigg(\sqrt{2}\sigma w_k g \left\{ \sqrt{2}\sigma z_k + y_i(t_i) \right\} \exp\left[-\gamma(t_i) e^{\boldsymbol{x}_i^T \boldsymbol{\beta} - \mu\left\{\sqrt{2}\sigma z_k + y_i(t_i)\right\}} \right] \\ e^{-\mu\left\{\sqrt{2}\sigma z_k + y_i(t_i)\right\}} \Bigg), \tag{7}$$

where N is the number of the quadratures and ω_k and z_k are weights and abscissae for Gauss–Hermite quadrature, respectively. This loop of the expectation step and the maximization step is repeated until $|(\boldsymbol{\theta}^{(l+1)}, \gamma(t)^{(l+1)}) - (\boldsymbol{\theta}^{(l)}, \gamma^{(l)})|$ is smaller than a pre-specified criterion. We denote the final estimators as $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\boldsymbol{\beta}}^T)^T$ and $\hat{\boldsymbol{\gamma}}$.

2.3. Variance estimation

In the asymptotic results given in the supplemental material, we show that the proposed estimator for the true parameter $\boldsymbol{\theta}_0$ is semiparametrically efficient. Moreover, the efficient score function for θ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is

$$l^*_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \gamma_0 \mid \boldsymbol{w}) = \begin{pmatrix} \mu_0^{-1} - E(\kappa\xi \mid \boldsymbol{w}) - E(\kappa \mid \boldsymbol{w}) R_1(t) \\ E(\kappa \mid \boldsymbol{w}) \{\boldsymbol{x} - R_2(t)\} \end{pmatrix}, \tag{8}$$

where $\kappa = 1 - \gamma_0(t) \exp(\boldsymbol{x}\beta_0 - \mu_0\xi)$; γ_0 is the true function of t;

$$R_1(t) = E\big[E(\kappa \mid \boldsymbol{w}) \left\{\mu_0^{-1} - E(\kappa\xi \mid \boldsymbol{w})\right\} \mid t\big] \big/ E\big\{E(\kappa \mid \boldsymbol{w})^2 \mid t\big\},$$
$$R_2(t) = E\left\{\boldsymbol{x} E(\kappa \mid \boldsymbol{w})^2 \mid t\right\} \big/ E\left\{E(\kappa \mid \boldsymbol{w})^2 \mid t\right\}.$$

The derivation for the efficient score function is detailed in Lemma 2 of the Supplementary material. Therefore, the asymptotic variance of $n^{1/2}\hat{\boldsymbol{\theta}}$ is the inverse of the information for $\boldsymbol{\theta}_0$, that is, $I(\boldsymbol{\theta}_0) = E(l_{\boldsymbol{\theta}}^{*\otimes 2})$, where $a^{\otimes 2} = aa^T$ for any vector a.

For the asymptotic variance of $n^{1/2}\hat{\boldsymbol{\theta}}$, we estimate $I(\boldsymbol{\theta}_0)$ by $n^{-1} \sum_{i=1}^{n} \hat{l}_{\boldsymbol{\theta}i}^{*\otimes 2}$, where

$$\hat{l}^*_{\theta i} = \begin{pmatrix} \hat{\mu}^{-1} - \hat{E}(\kappa\xi \mid \boldsymbol{w}_i) - \hat{E}(\kappa \mid \boldsymbol{w}_i) \hat{R}_1(t_i) \\ \hat{E}(\kappa \mid \boldsymbol{w}_i) \big\{\boldsymbol{x}_i - \hat{R}_2(t_i)\big\} \end{pmatrix}, \tag{9}$$

and $\hat{E}(\kappa \mid \boldsymbol{w})$, $\hat{E}(\kappa\xi \mid \boldsymbol{w})$, $\hat{R}_1(t_i)$ and $\hat{R}_2(t_i)$ are some consistent estimators for $E(\kappa \mid \boldsymbol{w})$, $E(\kappa\xi \mid \boldsymbol{w})$, $R_1(t_i)$, and $R_2(t_i)$, respectively. Specifically $\hat{E}(\kappa \mid \boldsymbol{w})$ and $\hat{E}(\kappa\xi \mid \boldsymbol{w})$ are

$$\hat{E}(\kappa \mid \boldsymbol{w}) = 1 - \hat{\gamma}(t)\exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})\hat{E}\{\exp(-\hat{\mu}\xi) \mid \boldsymbol{w}\},$$
$$\hat{E}(\kappa\xi \mid \boldsymbol{w}) = \hat{E}(\xi \mid \boldsymbol{w}) - \hat{\gamma}(t)\exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})\hat{E}\{\exp(-\hat{\mu}\xi)\xi \mid \boldsymbol{w}\},$$

and the other two estimators are some type of kernel estimators with bandwidth h_n:

$$\hat{R}_1(t) = \frac{\sum_{j=1}^n K_{h_n}(t_j - t)\hat{E}(\kappa \mid \boldsymbol{w}_j)\{\hat{\mu}^{-1} - \hat{E}(\kappa\xi \mid \boldsymbol{w}_j)\}}{\sum_{j=1}^n K_{h_n}(t_j - t)\hat{E}(\kappa \mid \boldsymbol{w}_j)^2},$$
$$\hat{R}_2(t) = \frac{\sum_{j=1}^n \boldsymbol{x}_j K_{h_n}(t_j - t)\hat{E}(\kappa \mid \boldsymbol{w}_j)^2}{\sum_{j=1}^n K_{h_n}(t_j - t)\hat{E}(\kappa \mid \boldsymbol{w}_j)^2},$$

where $K_{h_n}(x) = h_n^{-1}\exp(-x^2/h_n)$. In the Supplementary material, we establish the consistency of this variance estimator assuming that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. We choose $(n/2)^{-1/2}$ for h_n.

3. Asymptotic results

In this section, we provide asymptotic results for the proposed estimators under the following conditions. Let $\boldsymbol{\theta}_0$ and γ_0 denote the true regression parameter and monotone increasing function, respectively.

(A1) The finite-dimensional parameter space $\Theta(\subset \mathscr{R}^d)$ is a compact subset of the domain of $\boldsymbol{\theta}$.
(A2) The covariate $\boldsymbol{x}$ has bounded support with probability 1. If $\boldsymbol{x}^T\boldsymbol{\beta} + \alpha = 0$ almost surely (a.s.), then $\boldsymbol{\beta} = 0$ and $\alpha = 0$.
(A3) The support of the observation time, T, is an interval $\mathscr{S}[T] = [l_T, u_T]$, with $0 < l_T \le u_T < \infty$.
(A4) The monotone increasing function $\gamma_0(t)$ has strictly positive derivative on $\mathscr{S}[T]$.

The assumptions that parameter, covariate, and observation time are bounded are standard. Condition (A2) ensures the identifiability of $\boldsymbol{\theta}$ and γ. These conditions hold naturally in most applications.

For convergence of the estimates to the true parameters, we need to define a topology. Let the bounded regression parameter space $\boldsymbol{\Theta}$ be equipped with the Euclidean topology. Regarding infinite dimensional nonparametric space, let $\mathscr{F}$ be the set of all Borel sub-probability measures on $\mathscr{S}[T]$.

Theorem 1 (*Consistency of the MLE*)**.** *Under conditions (A1)–(A3), $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ almost surely, and if $t \in \mathscr{S}[T]$ is a continuity point of γ_0, $\hat{\gamma}(t) \to \gamma_0(t)$ almost surely. Moreover, if γ_0 is continuous, then $\sup_{t\in\mathscr{S}[T]} \mid \hat{\gamma}(t) - \gamma_0(t) \mid \to 0$ almost surely.*

Before discussing the overall convergence rate, we define the distance d on $\mathscr{R}^d \times \mathscr{F}$ as follows:

$$d\{(\theta_1, \gamma_1), (\theta_2, \gamma_2)\} = |\theta_1 - \theta_2| + \|\gamma_1 - \gamma_2\|_{2,P_T},$$

where $|\theta_1 - \theta_2|$ is the Euclidean distance in $\mathscr{R}^d$,

$$\|\gamma_1 - \gamma_2\|_{2,P_T} = \left[\int \{\gamma_1(v) - \gamma_2(t)\}^2 dP_T\right]^{1/2},$$

and P_T is the marginal probability measure of the measurement time variable T.

Our next theorem gives the convergence rates of the estimators in terms of this distance.

Theorem 2 (*Rate of Convergence*)**.** *Under Conditions (A1)–(A3),*

$$d\{(\hat{\boldsymbol{\theta}}, \hat{\gamma}), (\theta_0, \gamma_0)\} = O_p(n^{-1/3}).$$

The overall rate of convergence is dominated by $\hat{\gamma}$. However, it is shown in the next theorem that the convergence rate of $\hat{\boldsymbol{\theta}}$ can be refined to achieve a rate of $n^{1/2}$.

Theorem 3 (*Asymptotic Normality and Efficiency*)**.** *Suppose that θ_0 is an interior point of Θ and that conditions (A1)–(A4) are satisfied. Then*

$$\begin{aligned} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= n^{1/2}(\mathbb{P}_n - P)\left\{I(\boldsymbol{\theta}_0)^{-1} l^*_{\theta_0}(\boldsymbol{w})\right\} + o_p(1) \\ &\to N(0, I(\boldsymbol{\theta}_0)^{-1}) \quad \textit{in distribution,} \end{aligned}$$

where $\mathbb{P}_n$ is the empirical measure of $\boldsymbol{w}_i$, $i = 1, \ldots, n$, that is, $\mathbb{P}_n\{l^*_{\theta_0}(\boldsymbol{w})\} = n^{-1}\sum_{i=1}^n l^*_{\theta_0}(\boldsymbol{w}_i)$, P is the probability measure, that is, $P\{l^*_{\theta_0}(\boldsymbol{w})\} = \int l^*_{\theta_0}(\boldsymbol{w})dP$, $l^*_{\theta_0}(\boldsymbol{w})$ is the efficient score defined in (8), and $I(\theta_0)$ is the information.

Since $\hat{\boldsymbol{\theta}}$ is asymptotically linear with efficient influence function, and the model (the likelihood function) is sufficiently smooth (Hellinger differentiable) with respect to $(\boldsymbol{\theta}, \gamma)$, it is asymptotically efficient in the sense that any regular estimator has asymptotic variance matrix no less than that of $\hat{\boldsymbol{\theta}}$.

Theorem 4 (*Consistency of Information Estimator*). *When the bandwidth h_n satisfies that h_n and $\log n/(nh_n)$ converge to 0 as $n \to \infty$, $\mathbb{P}_n\left\{\hat{l}^{*\otimes 2}_{\boldsymbol{\theta} i}\right\}$ converges to $P\left\{l^{*\otimes 2}_{\theta_0}\right\}$.*

The theorems are justified in Section 2 of the Supplementary material.

4. Numerical examples

4.1. Simulations

Simulation studies were conducted to assess the performance of the estimators proposed in Section 2. We consider two sets of simulations in which the visit times are either discrete or continuous random variables. We compared the proposed estimator (called by Mixture-model) with the estimator not accounting for measurement error, that is, $\sigma^2 = 0$ (called by Gumbel-model). For discrete measurement times, the time point for each subject is distributed as a multinomial distribution with the even probability on the support of $\{0.1, 0.2, 0.4, 0.8\}$, while continuous measurement times are distributed as a uniform distribution over [0, 1]. We round up continuous measurement times to the closest time points among 150 fixed and evenly spaced time points (reported times) between 0 and 1 for reducing computation burden. The error between the true measurement times and rounded reported times ranges −0.003 to 0.003. The model includes two covariates: one is distributed as a Bernoulli distribution with probability 0.5, and the other is generated from a normal distribution with mean 0 and variance 0.1. The true values for (μ, β_1, β_2) are (1, 0.3, 0.3), and $\gamma_0(t)$ is assumed to be monotone increasing over time. We considered three different monotone increasing functions for $\gamma_0(t)$: $2t^{1/5}$, a curve with sharp increase early and plateau later and the other curve with two change points. The latter two curves are generated by cubic I-splines (Ramsay, 2008), and the related simulation settings are described in Section 4 of the Supplementary material.

Consequently, the true biomarker values are generated as follows:

$$y_i^*(t_i) = \mu^{-1}\left[\boldsymbol{x}_i^T\boldsymbol{\beta} - \log\left\{-\log(p_i)/\gamma_0(t_i)\right\}\right], \tag{10}$$

where t_i and p_i are from a uniform distribution over [0, 1]. The observed biomarker values, y_i's are obtained by adding y_i^* and ϵ_i, where ϵ_i is independently generated from a normal distribution with mean 0 and variance $\sigma^2 = 0.25$ and 1, which correspond to the ratio of a measurement error variance to a true biomarker variance, 0.17 and 0.63, respectively. We varied sample sizes from 500 to 2000 and conducted 1000 replicates for each simulation study.

We applied the proposed EM algorithm to simulated data for estimating the parameters. The initial values used for $\boldsymbol{\beta}$ and $\gamma(t)$ in the algorithm were 0's and observed times, respectively. In the M-step, the spectral projected gradient method was used for constrained optimization in (4). The convergence criterion for the EM algorithm was set as 10^{-3}, and the number of nodes, $N = 128$. This results in very stable integrations through all numerical studies. The results do not change when we further increase N. We thus recommend N to be 128.

In the simulations, we noticed that the biomarker effect μ was sensitive to the initial values. Therefore, we first calculated the profile likelihood for μ using the same algorithm except that μ was held at some fixed value; we then carried out a grid search for finding the maximizer for μ. The variance estimation was based on the formula in Section 2.3. We applied the Newton–Raphson algorithm for estimating the parameters of the Gumbel-model and employed the linearization method (Graubard and Fears, 2005) for estimating the asymptotic variance of regression coefficient estimates. Although data-driven methods for choosing a bandwidth in kernel estimating are more practical, it increases computation burden substantially, and the sample size dependent bandwidth works well in the simulation study.

The simulation results are similar between the discrete and continuous time points, so we present the results of simulations for continuous time points for $\gamma_0(t) = 2t^{0.2}$. Table 1 shows that the regression estimators are asymptotically unbiased, and the bias and variance decrease when the sample size increases. The variance estimators are nearly unbiased when the measurement error ratio is less than 0.35, whereas where the measurement error ratio is greater than 0.35, the variance estimates for $\hat{\boldsymbol{\beta}}$ are approximately unbiased, but the variance estimator for $\hat{\mu}$ yields underestimation because σ and μ are confounded in Gauss–Hermite quadrature form (7). In another transformation of model (2) by letting $\mu\xi_i = z_i$, the confounded term $\mu\sigma$ is not separated out (the detail is provided in Section 3 of the Supplementary material). We can prove identifiability of σ theoretically; however, numerically it is difficult to estimate μ directly. High measurement error ratio does not numerically guarantee the asymptotic normality for μ estimates. When the measurement error is ignored, the extent of bias in the estimates is larger by 10-fold than the bias of the proposed estimators. When the true curve $\gamma_0(t)$ is not smooth, $\hat{\mu}$ may be slightly more biased, and the corresponding variance is underestimated; however, estimates for β have little bias regardless (Table 1 and 2 of the Supplementary material).

We evaluated robustness of Mixed-models to non-normality of the measurement error via a numerical study. We let the true distribution for the measurement error be exponential-gamma distributions with mean 0 and variance 0.28 (the

Table 1
Simulation results of continuous random time points and $\gamma_0(t) = 2t^{0.2}$; ESE = empirical standard error; ASE = asymptotic standard error; CP = coverage probability.

N	Parameter	Unit = %							
		Mixture-model				Gumbel-model			
		Bias	ESE	ASE	CP	Bias	ESE	ASE	CP
		$\sigma = 0.5$, measurement error ratio = 0.17							
500	$\mu = 1.0$	0.4	4.6	4.8	96.4	−11.3	3.3	3.1	7.8
	$\beta_1 = 0.3$	0.4	11.2	12.5	97.5	−3.2	10.4	9.8	92.1
	$\beta_2 = 0.3$	−0.0	17.0	19.8	98.2	−3.7	16.1	15.3	93.1
1000	$\mu = 1.0$	−0.1	3.2	3.3	96.3	−11.9	2.4	2.2	0.4
	$\beta_1 = 0.3$	−0.3	7.7	8.5	97.1	−3.8	7.2	7.0	91.1
	$\beta_2 = 0.3$	0.0	12.2	13.4	96.4	−3.4	11.3	11.0	92.6
		$\sigma = 1$, measurement error ratio = 0.63							
1000	$\mu = 1.0$	−4.4	11.0	4.4	70.4	−30.7	1.9	1.9	0.0
	$\beta_1 = 0.3$	−1.4	10.1	10.6	96.2	−9.3	7.8	7.4	74.6
	$\beta_2 = 0.3$	−0.7	15.2	16.8	96.3	−9.0	12.0	11.6	87.0
2000	$\mu = 1.0$	−3.8	10.0	3.1	72.9	−31.0	1.4	1.4	0.0
	$\beta_1 = 0.3$	−1.0	7.7	7.3	93.7	−9.2	5.4	5.3	59.3
	$\beta_2 = 0.3$	−0.9	11.0	11.6	96.0	−9.1	8.2	8.3	79.8

measurement error ratio = 0.28) and with mean 0 and variance 1.16 (the measurement error ratio = 1.25); however, we assumed it was generated from a normal distribution with the same mean and variance as the true distribution. The simulation result shows that estimates of the Mixture-models are robust to non-normality of the measurement error (Table 3 of the Supplementary material).

4.2. Application

We applied our approach to data from an ancillary study to the Atherosclerosis Risk in Communities study. In the ancillary study, the association of type 2 diabetes incidence with six inflammation biomarkers was investigated in four U.S. communities. We used fasting blood glucose (FPG) values as a diabetes biomarker and investigated temporal associations (potential causation) between blood glucose values and factors, such as demographic characteristics, chronic disease status and lab data while accounting for the skewed distribution of blood glucose values, error in blood glucose-measurement and monotone increasing age effect on blood glucose levels. We analyzed a subpopulation of 1560 Caucasian females from Forsyth County, North Carolina, who have complete covariates at the first visit. The ancillary study is longitudinal; however, we use only the FPG values observed at the second visit excluding subjects having diabetes diagnosed by the second visit or being on diabetes medication at the second visit (0.3% in the sub-population) because such subjects may have subsequently taken medication or made dietary changes that could have influenced their blood glucose values.

In the mixture-model, the covariates include body mass index (BMI), current smoking status, hypertension, low density lipoprotein (LDL) and high-density lipoprotein (HDL), which were measured at visit 1; FPG measurement time is age at the visit. The average age at visit 1 of the sub-population was 57.3 years with range 45.6–68.0 years. The average time-to-visit 2 is 2.9 years with standard deviation of 0.20 and range 2.5–5.5 years. The average BMI of the sub-population was 25.0 kg/m^2 with standard deviation 4.54 kg/m^2. The numbers of current smokers and participants with hypertension were 422 (27.1%) and 312 (20.0%), respectively.

The observed biomarker, $y_i(t)$, is defined as the FPG value standardized with sample mean 99.6 mg/dl and standard deviation 11.5 mg/dl so that it has zero mean and variance one. The measurement time is scaled down to (0,1]. The standardized value and the rescaled observation time better facilitate the estimation process than original value or log-transformed value. According to Hadi et al. (2008), the coefficient of variance (CV) for measurement error in laboratory glucose values is 3.5%. We chose $\sigma^2 = 0.3^2$ corresponding to 3.5% CV of measurement error for the standardized FPG value.

For comparison, we applied a Gumbel-model, which ignores the measurement error in FPG values and a linear regression model. The Gumbel-model includes the same covariates used in the Mixture-model. We have considered linear models including different age-adjustments, including isotonic regression using the pooled adjacent violation algorithm, cubic B-splines, categorical dummy age variables divided at quantiles and locations at which the isotonic regression jumps. Among the linear models we considered, the selected model is the linear model including the same covariates used in the Mixture-model and additionally categorical dummy age variables dividend at quantiles. The comparison of different age-adjustments in goodness-of-fit is summarized in Section 5 and Table 4 of the Supplementary material. For a better fit, we log-transformed the observed FPG values and then standardized those using mean 4.59 and variance 0.01.

The result of the analysis is given in Table 2. It shows that the covariates hypertension and higher BMI have significantly increased the probability having a higher FPG value at next visit than the observed one. Given fixed FPG value ξ, compared to normotensive subjects, subjects with hypertension have 1.53 times greater probability of having a higher FPG value at

Table 2
Application to the ARIC Study in Caucasian females from Forsyth County, NC; BG = Blood glucose; Hypert. = Hypertension; Smoking = Current Smoking; Est = regression coefficient estimate; ASE = asymptotic standard error.

Coefficients	Mixture-model			Gumbel-model		
	Est	ASE	p-value	Est.	ASE	p-value
BG(12.2 mg/dL)	1.148	0.021	<0.0001	0.863	0.037	<0.0001
Hypert. = Yes	0.423	0.061	<0.0001	0.567	0.079	<0.0001
Smoking = Yes	0.013	0.052	0.802	0.072	0.091	0.427
BMI(4.5 kg/m^2)	0.106	0.026	<0.0001	0.068	0.036	0.057
LDL(37.7 mg/dL)	−0.015	0.019	0.436	0.003	0.054	0.957
HDL(17.5 mg/dL)	−0.047	0.026	0.073	−0.009	0.043	0.831

Table 3
Linear model application to the ARIC Study in Caucasian females from Forsyth County, NC; ASE = asymptotic standard error.

Coefficients	Estimate	ASE	p-value
Intercept	−0.226	0.059	0.0001
Hypertension = Yes	0.222	0.065	0.0007
Current Smoking = Yes	−0.028	0.057	0.6280
BMI(4.5 kg/m^2)	0.152	0.027	<0.0001
LDL(37.7 mg/dL)	−0.014	0.027	0.6012
HDL(17.5 mg/dL)	−0.090	0.028	0.0013
$50.9 < \text{Age} \leq 55$	0.152	0.078	0.0530
$55 < \text{Age} \leq 59$	0.182	0.079	0.0220
$59 < \text{Age} \leq 63.7$	0.286	0.081	0.0004
$63.7 < \text{Age} \leq 68$	0.316	0.080	0.0001

next visit than ξ. For each 4.5 kg/m^2 increase in BMI, the probability of having a higher FPG value at next visit than ξ increases by factor 1.12.

Comparatively, the analysis of the Gumbel-model yields different results in effect size and significance in BMI (the p-value is on the borderline). The result of the linear regression analysis in Table 3 is not directly comparable with the Mixture- and Gumbel-models. However, it is agreed that the both factors of hypertension and high BMI are significantly associated with increasing FPG values. In addition, age effect on FPG values is seemingly monotone increasing.

To investigate the goodness-of-fit of the applications, we generated predicted glucose values using formula (10) based on the parameter estimation, covariates, cumulative density probability p's randomly generated from a uniform distribution between 0 and 1, and generated measurement error from a normal distribution with mean 0 and variance 0.09. Using the predicted values, we suggest two methods for model diagnosis. First, Quantile–Quantile (QQ) plots are generated to compare the distribution of observed glucose values with the distribution of predicted means (the right column of Fig. 1). Second, we calculated the residuals by subtracting the predicted means from the observed glucose values (the left column of Fig. 1). The residuals and QQ-plots are very comparable between the Mixture- and Gumbel-models although the QQ-plot of the Mixture-model shows slightly better fit to the data than the Gumbel-model. In contrast, the QQ-plot of the linear regression model shows notable distinction between the distributions of observed and predicted values. The three residual plots are randomly scattered around zero. For sensitivity analysis, we applied different measurement error variance values ranging from 0.01 to 0.36, and the analysis results are robust in significance and effect size for binary covariates.

We predicted FPG values at the third visit using the linear, Gumbel- and Mixture-models based on the first and second visits and compared the predicted values with the FPG values observed at the third visit from 1,399 subjects, who have no diabetes diagnosed by the third visit or not being on diabetes medication at the third visit. The correlation between FPG values at the second and third visits is 0.52. Positive and negative residuals of Fig. 2 mean lower and higher prediction than observed values. The Mixture-model has slightly better accuracy than the Gumbel-model for the observations in the middle of range. The linear model more accurately predicts observations near mean 0, whereas the Gumbel-/Mixture model has more accuracy than the linear model for the observations in the high range. The Mixture-/Gumbel-model yields more accurate prediction in the right tail than the lower tail.

5. Discussion

We proposed efficient semiparametric likelihood-based regression models for continuous and skewed biomarker data. Observed biomarker values were analyzed separately as true value and measurement error. An additive model was used to account for biomarker values subject to measurement error by assuming that measurement error follows a Gaussian process with zero-mean and finite variance and is independent of the true biomarker values. We adopted Gumbel

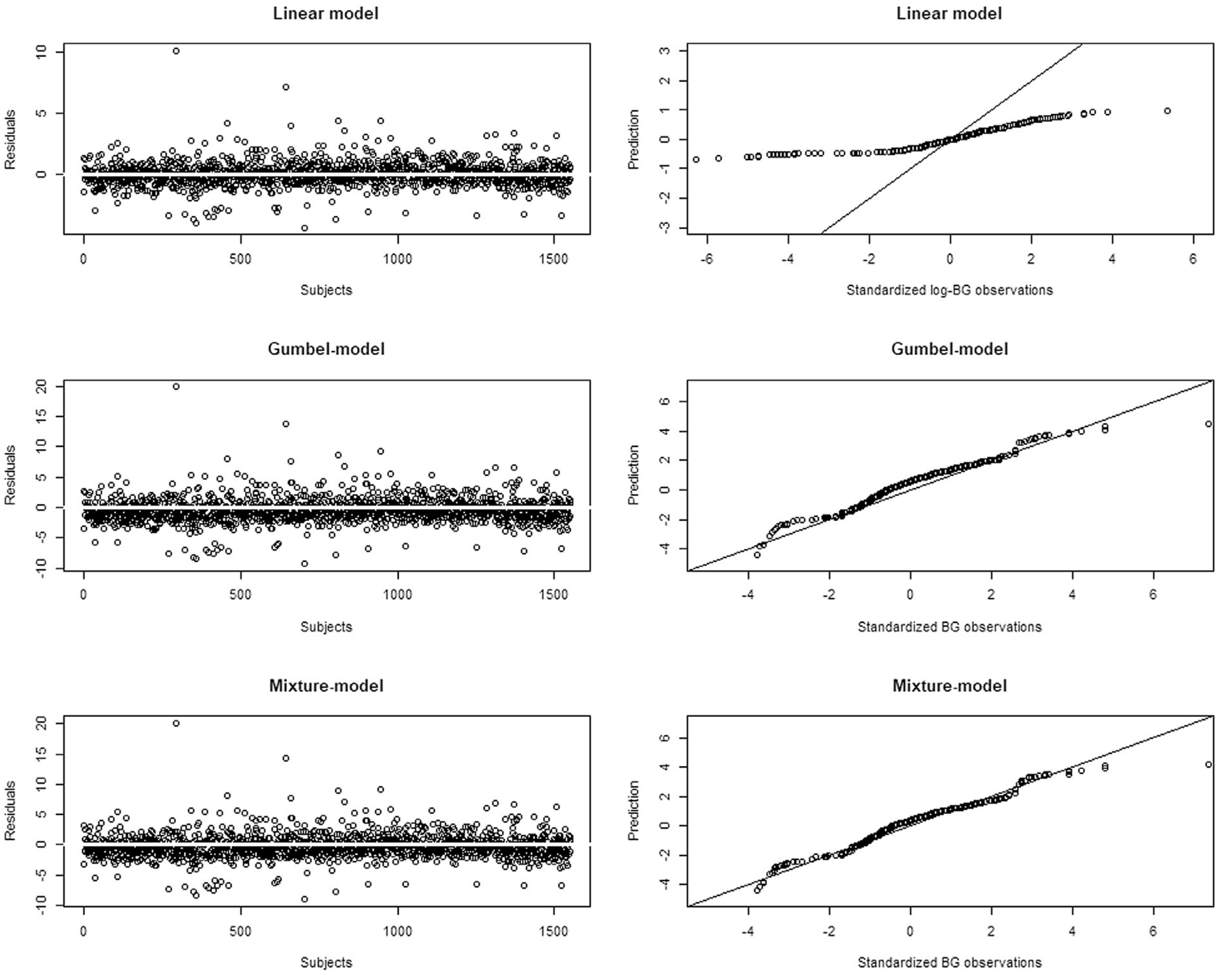

Fig. 1. Goodness-of-fit of the linear, Gumbel-/Mixture-models: left are residuals plots, and right are Quantile–Quantile plots.

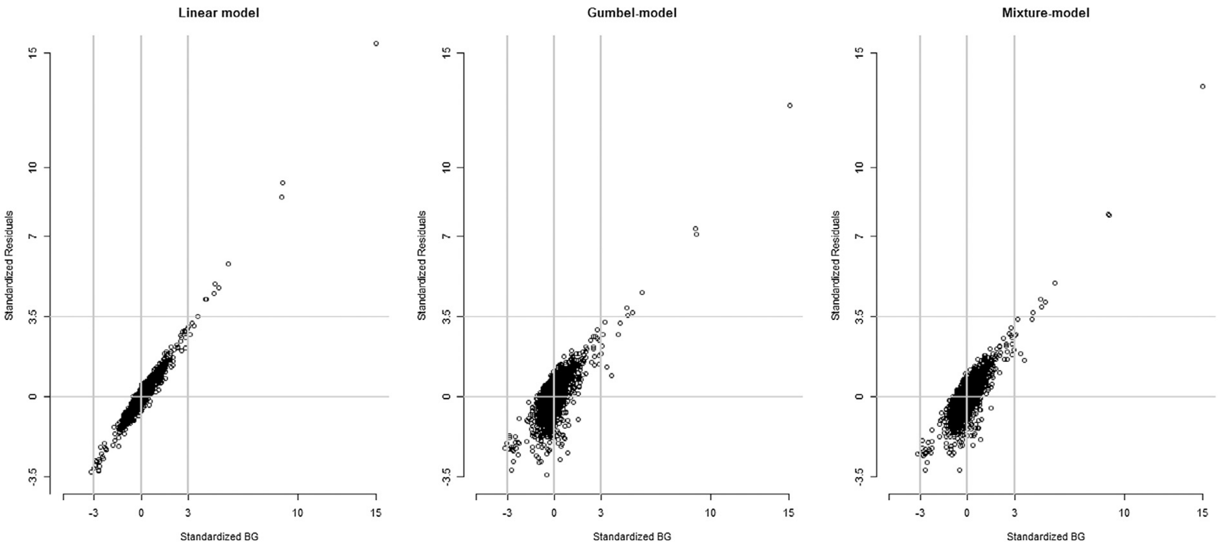

Fig. 2. Prediction for FPG values at the third visit using the linear and Gumbel-/Mixture-models based on the first and second visits.

distributions to construct a stochastic model for the time-varying true biomarker values and restricted the stochastic model to be monotone increasing over time. Then we constructed the marginal observed likelihood for the observed biomarker values using a mixture of a Gumbel with a normal distribution.

The proposed estimator results in asymptotically unbiased estimations and provides asymptotic normality of the regression coefficient estimates when the measurement error ratio is moderate or small. The interpretation of the models is not as straightforward as linear regression models; however the flexibility of the mixture distributions enables the model to better fit to data skewed with a long tail. Gumbel-/Mixture-models yield more accurately prediction in the right end than the lower end. We did not present an application of ordered statistics or extreme events, but the proposed models are also applicable to such extreme values.

The methods proposed in this paper can be generalized to repeated observations using pseudo-likelihood ignoring dependence between biomarker values within the same subject. We built a non-parametric estimation for $\gamma(t)$ as we do not need to specify the number of knots and locations for splines; however, we can relax the monotonicity of the function of time by using splines for modeling non-monotone stochastic trends in biomarker levels. Furthermore, we can consider other regression models, such as linear transformation models or additive models.

Acknowledgments

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute, United States contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jspi.2019.03.008.

Supplementary Material available in the attached file include the proofs of the model identifiability and Theorems 1–4 in Section 3, and simulation results and additional explanations referred in the main context.

References

Balkema, A., De Haan, L., 1974. Residual life time at great age. Ann. Probab. 2, 792–804.

Carroll, R.J., 2006. Measurement Error in Nonlinear Models: A Modern Perspective. Chapman and Hall/CRC.

Davis, P.J., 1984. Methods of Numerical Integration. Academic Press, Orlando.

Fisher, R.A., Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc. Cambr. Philos. Soc. 24, 180–290.

Fuller, W.A., 1987. Measurement Error Models. John Wiley & Sons, Inc.

Ghosh, S., Mallick, B.K., 2011. A hierarchical bayesian spatio-temporal model for extreme precipitation events. Environmetrics 22.

Gnedenko, B., 1943. Sur la distribution limite du terme maximum d'une série aléatoire. Ann. Mat. 44, 423–453.

Graubard, B.I., Fears, T.R., 2005. Standard errors for attributable risk for simple and complex sample designs. Biometrics 61, 847–855.

Gumbel, E., 1958. Statistics of Extremes. Columbia University Press, New York.

Hadi, M., Bacharach, S.L., Whatley, M., Libutti, S.K., Straus, S.E., Rao, V.K., Wesley, R., Carrasquillo, J.A., 2008. Glucose and insulin variations in patients during the time course of a fdg-pet study and implications for the glucose-corrected suv. Nucl. Med. Biol. 35 (4), 441–445.

Huerta, G., Sansó, B., 2007. Time-varying models for extreme values. Environ. Ecol. Stat. 14, 285–299.

von Mises, R., 1936. La distribution de la plus grande de n valeurs. Rev. Math. Union Interbalkaniqu 1, 141–160.

Pickands, J., 1975. Statistical inference using extreme order statistics. Ann. Statist. 3, 119–130.

Ramsay, J.O., 2008. Monotone regression splines in action. Stat. Sci. 3 (4), 425–441.

Roy, V., Dey, D.K., 2014. Propriety of posterior distributions arising in categorical and survival models under generalized extreme value distribution. Statist. Sinica 24 (2), 699–722.

Tsiatis, A.A., DeGruttola, V., Wulfsohn, M.S., 1995. Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and cd4 counts in patients with aids. J. Amer. Statist. Assoc. 90, 27–37.

Wang, X., Dey, D.K., 2010. Generalized extreme value regression for binary response data: an application to b2b electronic payments system adoption. Ann. Appl. Stat. 4 (4), 2000–2023.

Wang, H.J., Li, D., 2013. Estimation of extreme conditional quantiles through power transformation. J. Amer. Statist. Assoc. 108 (503), 1062–1074.

Young, J.K., Ellison, J.M., Marshall, R., 2008. Performance evaluation of a new blood glucose monitor that requires no coding: the onetouch vita system. J. Diabetes Sci. Technol. 2 (5), 814–818.