Chapter 11
Subgroup Analysis with Partial Linear Regression Model

Yizhao Zhou, Ao Yuan, and Ming T. Tan

Abstract In clinical trials it is common that the treatment has different effects on different subjects. This motivates the precision medicine, the goal is to identify the treatment favorable or unfavorable subgroups, if they exist, and classify the subjects into one of the subgroups based on their covariate values. In practice, some covariate(s) is known to affect the response non-linearly, in this case the existing linear model is not adequate. To address this issue, we use a partial linear model, in which the effect of some specific covariates is a non-linear monotone function, along with a linear part for the rest of the covariates. This approach not only makes the model more flexible than the parametric linear model, and more interpretable and efficient than the full nonparametric model. The Wald statistics is used to test the existence of subgroups, and the Neyman-Pearson rule is used to classify the subjects. Simulation studies are conducted to evaluate the performance of the method, and then the method is used to analyze a real clinical trial data.

11.1 Introduction

In clinical studies, often treatment effect is not uniform over all the patients, some subgroup of patients may benefit significantly from the treatment and others may not so. Thus one of goals of precision medicine is to find out if such subgroups exist or not, and if existence is justified, identify the subgroups of patients according to their covariate values. For example, in IBCSG (2002), patients with ER-negative tumors were likely to benefit from chemotherapy, while those with ER-positive tumors did not.

Y. Zhou (✉) · A. Yuan · M. T. Tan
Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC, USA
e-mail: yz459@georgetown.edu

Subgroup analysis is recently a very active research area see, e.g., Sabine (2005), Song and Chi (2007), Ruberg et al. (2010), Foster et al. (2011), Lipkovich et al. (2011), Friede et al. (2012), Shen and He (2015), Fan et al. (2017), and Ma and Huang (2017). Rothmann et al. (2012) discussed issues for subgroups testing and analysis. Fokkema (2018) used generalized linear mixed-effect model tree (GLMM tree) algorithm detecting treatment-subgroup interactions in clustered datasets. Yuan et al. (2018, 2020) proposed semiparametric methods for this problem.

Existing methods for this problem often use linear model. In practice, sometimes it is known that some covariate has non-linear effect on the response, incorporating such information can improve the quality of the analysis. Here we consider such case and apply a more featured partial linear model to identify the existence of subgroups and to classify the subjects into different subgroups if the existence of subgroup is confirmed. This model assumes a monotone non-linear effect of some covariate, and linear effects from the rest covariates. First, a partial model with individual subgroup membership as latent variable and with a covariate whose effect are known as non-linear are formulated and the model regression parameters is estimated with expectation-maximization algorithm (E-M algorithm), and isotonic regression method is used for the maximum likelihood of the nonparametric non-linear part. Then null hypothesis of non-existence of subgroups are tested with Wald Statistics. If the existence of subgroup is confirmed, we use the Neyman-Pearson rule to classify each subject so that the misclassification error for the treatment favored group is under control while the misclassification error for the other subgroup is minimized.

The rest of the chapter is organized as follows. In Sect. 11.2 we describe the model and parameter estimation, Sect. 11.3 elaborates the testing and classification method, and Sect. 11.4 illustrates the simulation study and real data analysis.

11.2   The Method

The observed data is denoted as $D_n = \{(y_i, \boldsymbol{x}_i, z_i), i = 1, \ldots, n\}$, where $y_i \in R$ is the response variable of i-th subject, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})' \in R^d$ and z_i is another covariate, which is known to have a non-linear monotone effect on the response. Each subject i receives the same treatment, and we assume that bigger value of the response corresponds to better treatment effects. We want to test if there are treatment favorable and non-favorable subgroups in the patients. If subgroup does exist, we need to classify each subject into corresponding subgroup based on his/her covariate profile. In this paper, we assume that there are only two potential subgroups: treatment-favorable and treatment-nonfavorable subgroups. We need first to specify the model, estimate the model parameters, and then perform the hypothesis test and classification of subjects.

11.2.1 The Semiparametric Model Specification

We specify the semiparametric partial linear model as

$$y_i = \boldsymbol{\beta}' \boldsymbol{x}_i + g(z_i) + \delta_i \eta + \epsilon_i, \qquad \epsilon \sim N(0, 1), \qquad g \in \mathcal{G},$$

where δ_i is a latent indicator for whether subject i belongs to the treatment favorable subgroup ($\delta_i = 1$) or not ($\delta_i = 0$). $\boldsymbol{\beta}$ is a d-vector of unknown parameters, η is the effect of treatment favorable subgroup, and the constraint $\eta \geq 0$ is used for the identifiability with the intercept vector term in $\boldsymbol{\beta}$. It is assumed that the covariate z_i has a non-linear effect $g(\cdot)$ to the response y_i, we only know that $g(\cdot) \in \mathcal{G}$, the collection of all monotone increasing functions on R.

Denote the i.i.d. copy of the $(y_i, \boldsymbol{x}_i, z_i, \delta_i, \varepsilon_i)$'s as $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\delta}, \boldsymbol{\varepsilon})$. Let $\lambda = P(\delta = 1)$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \eta, \lambda)'$ be the vector of all the Euclidean parameters. Conditioning on $(\boldsymbol{x}, z)$, the density of y is the mixture

$$h(y | \boldsymbol{x}, z, \boldsymbol{\theta}) = \lambda \phi \left(y - \boldsymbol{\beta}' \boldsymbol{x} - g(z) - \eta \right) + (1 - \lambda) \phi \left(y - \boldsymbol{\beta}' \boldsymbol{x} - g(z) \right).$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. The log-likelihood of the observed data is

$$\ell(\boldsymbol{\theta}, g | D_n) = \sum_{i=1}^{n} \log \left(\lambda \phi (y_i - \boldsymbol{\beta}' \boldsymbol{x}_i - g(z_i) - \eta) + (1 - \lambda) \phi (y_i - \boldsymbol{\beta}' \boldsymbol{x}_i - g(z_i)) \right),$$

$$\boldsymbol{\theta} \in \boldsymbol{\Theta}, \ g \in; \mathcal{G}. \tag{11.1}$$

Direct computation of the maximum likelihood estimate (MLE) from a mixture model (11.1) is not convenient, especially in the presence of the nonparametric component $g(\cdot)$, and it is known that E-M algorithm (Dempster et al. 1977) is typically easy to use. For this, we treat the latent variable δ_i's as missing data, with $\delta_i = 1$ if the i-th subject belongs to the treatment-favorable subgroup, otherwise $\delta_i = 0$. The likelihood based on the 'complete data' $D_n^c = \{(y_i, \boldsymbol{x}_i, z_i, \delta_i) : i = 1, \ldots, n)\}$ is

$$L(\boldsymbol{\theta}, g | D_n^c) = \prod_{i=1}^{n} \left(\lambda \phi (y_i - \boldsymbol{\beta}' \boldsymbol{x}_i - g(z_i) - \eta) \right)^{\delta_i} \left((1 - \lambda) \phi (y_i - \boldsymbol{\beta}' \boldsymbol{x}_i - g(z_i)) \right)^{1 - \delta_i},$$

the corresponding log-likelihood is

$$\ell(\boldsymbol{\theta}, g | D_n^c) = \sum_{i=1}^{n} \Big(\delta_i \log \phi (y_i - \boldsymbol{\beta}' \boldsymbol{x}_i - g(z_i) - \eta)$$

$$+ (1 - \delta_i) \log \phi (y_i - \boldsymbol{\beta}' \boldsymbol{x}_i - g(z_i)) + \delta_i \log \lambda + (1 - \delta_i) \log(1 - \lambda) \Big). \tag{11.2}$$

The semiparametric MLE $(\hat{\boldsymbol{\theta}}_n, \hat{f}_n)$ of the true parameter $(\boldsymbol{\theta}_0, f_0)$ is given by

$$(\hat{\boldsymbol{\theta}}_n, \hat{g}_n) = \arg \max_{(\theta, g) \in (\Theta, \mathcal{G})} \ell(\boldsymbol{\theta}, g | D_n^c). \tag{11.3}$$

11.2.2 Estimation of Model Parameters

As the δ_i's are missing, $(\hat{\boldsymbol{\theta}}_n, \hat{g}_n)$ in (11.3) cannot be computed directly, the EM algorithm is used instead. For this a starting value $\boldsymbol{\theta}^{(0)}$ of $\boldsymbol{\theta}$ is needed, then find $g^{(1)}(\cdot) \in \mathcal{G}$ as the maxima of $\ell(\boldsymbol{\theta}^{(0)}, g | D_n^c)$, then fix $g^{(1)}$, find $\boldsymbol{\theta}^{(1)} \in \Theta$ as the maxima of $\ell_n(\boldsymbol{\theta}, g^{(1)})$, and so on . . . . until convergence of the sequence $\{(\boldsymbol{\theta}^{(r)}, g^{(r)})\}$, which is increasing the likelihood at each iteration, and will converge to at least some local maxima of $\ell_n(\boldsymbol{\theta}, g)$. In fact, the increasing likelihood property is obvious, as for all integer r,

$$\ell(\boldsymbol{\theta}^{(r+1)}, g^{(r+1)} | D_n^c) \geq \ell(\boldsymbol{\theta}^{(r)}, g^{(r+1)} | D_n^c) \geq \ell(\boldsymbol{\theta}^{(r)}, g^{(r)} | D_n^c).$$

A formal justification of the convergence of the above iterative algorithm is a case of the block coordinate descent methods in Bertsekas (2016).

Our algorithm is a semiparametric version of EM algorithm, see also Tan et al. (2009, chap. 2) for bio-medical applications of this algorithm. The semiparametric and nonparametric EM algorithm was used in a large number of literatures, such as in Muñoz (1980), Campbell (1981), Hanley and Parnes (1983), Groeneboom and Wellner (1992, Section 3.1), and see the argument there for the convergence of such algorithm (p. 67–68). Chen et al. (2002) applied the EM algorithm to a semiparametric random effects model, Bordes et al. (2007) applied the EM algorithm to a semiparametric mixture model, using simulation studies to justify the convergence of the algorithm. Balan and Putter (2019) developed an R-package of EM algorithm for semiparametric shared frailty models.

Now we give the detail of the algorithm. At each iteration r, do the following:

Step 0. For fixed $(g^{(0)}, \boldsymbol{\theta}^{(0)})$, compute $\{\delta_i^{(0)}\}$ with E-step of E-M algorithm.
Step 1. For fixed $(g^{(r)}, \boldsymbol{\theta}^{(r)})$, compute

$$
\begin{aligned}
H_n(\boldsymbol{\theta}, g | \boldsymbol{\theta}^{(r)}, g^{(r)}) &= E_{\boldsymbol{\delta}}[\ell(\boldsymbol{\theta}, g | D_n^c) | D_n, \boldsymbol{\theta}^{(r)}, g^{(r)}] \\
&= \sum_{i=1}^n \Big(\delta_i^{(r)} \log \phi(y_i - \boldsymbol{\beta}' \boldsymbol{x}_i - g(z_i) - \eta) \\
&\quad + \delta_i^{(r)} \log \lambda + (1 - \delta_i^{(r)}) \log \phi(y_i - \boldsymbol{\beta}' \boldsymbol{x}_i - g(z_i)) \\
&\quad + (1 - \delta_i^{(r)}) \log(1 - \lambda)) \Big),
\end{aligned}
\tag{11.4}
$$

where the expectation is taken with respect to the missing δ, and as if the true data is generated from parameters $(\theta^{(r)}, g^{(r)})$. In particular, the r-th step estimates of the δ_i's (for $i = 1, \ldots, n; r = 0, 1, 2, \ldots$), are

$$
\begin{aligned}
\delta_i^{(r)} &= E(\delta_i | y_i, x_i, z_i, g^{(r)}, \theta^{(r)}) = P(\delta_i = 1 | y_i, x_i, z_i, g^{(r)}, \theta^{(r)}) \\
&= \frac{P(y_i | \delta_i = 1, x_i, z_i, g^{(r)}, \theta^{(r)}) P(\delta_i = 1 | x_i, z_i, g^{(r)}, \theta^{(r)})}{P(y_i | x_i, z_i, g^{(r)}, \theta^{(r)})} \\
&= \frac{\lambda^{(r)} \phi\left(y_i - \beta'^{(r)} x_i - g^{(r)}(z_i) - \eta^{(r)}\right)}{\lambda^{(r)} \phi\left(y_i - \beta'^{(r)} x_i - g^{(r)}(z_i) - \eta^{(r)}\right) + (1 - \lambda^{(r)}) \phi\left(y_i - \beta'^{(r)} x_i - g^{(r)}(z_i)\right)}.
\end{aligned}
$$

Step 2. In the M-step for θ, compute

$$
\theta^{(r+1)} = \arg\sup_{\theta \in \Theta} H_n(\theta, g^{(r)} | \theta^{(r)}, g^{(r)}).
$$

This step can be computed by standard optimization packages. Especially,

$$
\lambda^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \delta_i^{(r)}.
$$

Step 3. For fixed $(\theta^{(r+1)}, \delta_i^{(r+1)})$ compute

$$
g^{(r+1)}(\cdot) = \arg\max_{g \in \mathcal{G}} H_n(\theta^{(r+1)}, g | \theta^{(r)}, g^{(r)}).
$$

This step computes the nonparametric maximum likelihood estimate of $\hat{g}$ under shape restriction, which is non-trivial, we describe it below.

11.2.2.1 Computation of $g^{(r+1)}$

The pool adjacent violators algorithm (PAVA, see for example, Best and Chakravarti (1990)) is a convenient computational tool to perform such order restricted maximization or minimization, and is available in R. Patrick et al. (2009) gives a review of the algorithm history and computational aspects. In particular, the computation of $\hat{g}(z_i) = \hat{g}_i$ is as follows.

$$g^{(r+1)}(\cdot) = \arg\max_{g \in \mathcal{G}} H_n(\boldsymbol{\theta}^{(r+1)}, g | \boldsymbol{\theta}^{(r)}, g^{(r)})$$

$$= \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \left(\delta_i^{(r)} (y_i - \beta^{(r)} x_i - \eta^{(r)} - g_i)^2 \right.$$

$$\left. + (1 - \delta_i^{(r)}) (y_i - \boldsymbol{\beta}^{(r)} \boldsymbol{x}_i - g_i)^2 \right)$$

$$= \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{n} (y_i - \beta^{(r)} x_i - \eta^{(r)} \delta_i^{(r)} - g_i)^2$$

Generally, let $v_i = y_i - \boldsymbol{\beta}' \boldsymbol{x}_i - \delta_i \eta$, $w_i = 1$, then

$$\hat{g} = \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{n} w_i (v_i - g_i)^2$$

The above is the standard form of isotonic regression procedure, and $\hat{g}$ can be computed using the R-function *isoreg*$(\cdot)$.

11.2.3   Asymptotic Results of the Estimates

Zhou et al. (2019) derived asymptotic results for $\hat{\boldsymbol{\theta}}$ and $\hat{g}(\cdot)$, as presented below. Detailed regularity conditions and proofs can be found there.

Theorem 11.1 *Under regularity conditions, as $n \to \infty$*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \overset{a.s.}{\to} 0, \quad \int |\hat{g}(z) - g_0(z)| dz \overset{a.s.}{\to} 0.$$

Denote $\overset{D}{\to}$ for convergence in distribution.

Theorem 11.2 *Under regularity conditions, as $n \to \infty$,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{D}{\to} N(\boldsymbol{0}, I^{*-1}(\boldsymbol{\theta}_0 | g_0)),$$

where $I^*(\boldsymbol{\theta}_0 | g_0) = E[\ell^*(X, Z | \boldsymbol{\theta}_0, g_0)\ell^{*'}(X, Z | \boldsymbol{\theta}_0, g_0)]$ is the efficient Fisher informa-tion matrix of $\boldsymbol{\theta}$ for fixed g_0, and $\ell^*(X, Z | \boldsymbol{\theta}_0, g_0)$ is the efficient score for $\boldsymbol{\theta}$.

Let $\mathbb{B}(\cdot)$ be the two-sided Brownian motion originating from zero: a mean zero Gaussian process on R with $\mathbb{B}(0) = 0$, and $E(\mathbb{B}(s) - \mathbb{B}(h))^2 = |s - h|$ for all s, $h \in R$.

Theorem 11.3 *Denote $\dot{g}_0(z) = dg_0(z)/dz$ and density of z as q(z). Assume q(z) > 0. Under regularity conditions, as $n \to \infty$,*

$$n^{1/3}(\hat{g}_n(z) - g_0(z)) \xrightarrow{D} \left(\frac{4\dot{g}_0(z)}{q(z)}\right)^{1/3} \arg\max_{h \in R}\{\mathbb{B}(h) - h^2\}.$$

11.3 Testing the Null Hypothesis and the Classification Rules

11.3.1 Test the Null Hypothesis

After the model parameters are estimated, we need to test the existence of subgroups, which is formulated as testing the null hypothesis $H_0 : \eta = 0$ vs the alternative $H_1 : \eta \neq 0$. For parametric model, commonly used test statistic including the likelihood ratio statistic, score statistic and the Wald statistic, and the three statistics are asymptotically chi-squared distributed and equivalent. However, in our case when $\eta = 0$, λ is non-identifiable in the model, although the other parameters are still identifiable and estimable. In this case, the likelihood ratio statistic cannot be applied. So we use the Wald statistic.

Denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with $dim(\boldsymbol{\theta}) = d$ and $dim(\boldsymbol{\theta}_1) = d_1$, and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ is the MLE of $\boldsymbol{\theta}$ under the full model. Consider the null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$. The Wald test statistic is

$$W_n = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0})' Var^{-1}(\hat{\boldsymbol{\theta}}_1)(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0}).$$

If $Cov(\hat{\boldsymbol{\theta}}_1)$ is known, then asymptotically $W_n \sim \chi^2_{d_1}$. If $Cov(\hat{\boldsymbol{\theta}}_1)$ is estimated, asymptotically $W_n/d_1 \sim F_{d_1, n-d}$. For our problem, $\boldsymbol{\theta}_1 = \eta$, $\boldsymbol{\theta}_{1,0} = 0$, we treat $Cov(\hat{\eta})$ to be known, so $W_n = \hat{\eta}_n Var^{-1}(\hat{\eta}_n)\hat{\eta}_n \sim \chi^2_1$ asymptotically, and if $W_n > \chi^2_1(1 - \alpha)$, which is the upper $(1 - \alpha)$-th quantile of the χ^2_1 distribution, then H_0 is rejected.

11.3.2 The Classification Rule

After the existence of subgroup is justified, or the null hypothesis above is rejected, we need to classify the subjects. There are different classification rules. In subgroup analysis, the correct classification of the treatment favorable subgroup is of significant clinical meaning, so we use the Neyman-Pearson rule in Yuan et al. (2018, 2020) as it can control the miss-classification error for the treatment favorable subgroup.

To be specific, for each subject i, denote the i-th likelihood ratio

$$LR(y_i, \boldsymbol{x}_i) = \frac{f(y_i, \boldsymbol{x}_i, z_i | \hat{\boldsymbol{\theta}}, \delta = 1)}{f(y_i, \boldsymbol{x}_i, z_i | \hat{\boldsymbol{\theta}}, \delta = 0)} \approx \frac{\phi(y_i - \hat{\boldsymbol{\beta}}' \boldsymbol{x}_i - \hat{g}(z_i) - \hat{\eta})}{\phi(y_i - \hat{\boldsymbol{\beta}}' \boldsymbol{x}_i - \hat{g}(z_i))}.$$

Parallel to the NP uniformly most powerful test procedure for testing the simple hypothesis $H_0 : \eta = 0$ vs. $H_1 : \eta \neq 0$. For given significance level α, the optimal classification rule is: classify the i-th subject to subgroup S_1 if

$$LR(y_i, \boldsymbol{x}_i, z_i) \geq K(\alpha), \text{ with } K(\alpha) \text{ determined by } P_{H_0}\big(LR(Y, X, Z) \geq K(\alpha)\big) = \alpha,$$

or, with $\epsilon = y - \hat{\boldsymbol{\beta}}' \boldsymbol{x} - \hat{g}(z_i)$ generated under H_0,

$$P_{H_0}\left(\frac{\phi(y_i - \hat{\boldsymbol{\beta}}' \boldsymbol{x}_i - \hat{g}(z_i) - \hat{\eta})}{\phi(y_i - \hat{\boldsymbol{\beta}}' \boldsymbol{x}_i - \hat{g}(z_i))} \geq K(\alpha)\right) = \alpha.$$

We can find approximate solution for $K(\alpha)$. For simulated data, let $\{LR_j : j = 1, \ldots, n_0\}$ be the LR_j's of patients from the treatment unfavorable subgroup (for simulated data, the subgroup memberships are known), then set $K(\alpha)$ is estimated by the $(1 - \alpha)$-th upper quantile of $LR_1, \ldots, LR_{n_0}$, it is the cut-off beyond which patients will be classified to the treatment favorable subgroup, even though they are from the treatment unfavorable subgroup.

However, for real data $\{(y_i, \boldsymbol{x}_i, z_i) : i = 1, \ldots, n\}$, the subgroup memberships are unknown, we cannot use the above method to decide $K(\alpha)$, instead we obtain it by the following way. Set $LR_i = \phi(\epsilon_i - \hat{\eta})/\phi(\epsilon_i)$, let

$$Q_n(t) = \sum_{i=1}^{n} w_{ni} I(LR_i \leq t), \quad w_{ni} = (1 - \hat{\delta}_i) / \sum_{j=1}^{n} (1 - \hat{\delta}_j)$$

be a weighted empirical distribution of the LR_i's under the null hypothesis. Note that $1 - \hat{\delta}_i$ is the estimated membership of subject i belonging to group 0, corresponding to the null hypothesis, and $1 - \hat{\delta}_i$ scaled by $\sum_{j=1}^{n}(1 - \hat{\delta}_j)$ makes the w_{ni}'s a set of actual weights. So intuitively, $Q_n(\cdot)$ is a reasonable estimate of the distribution of the LR_i's under the null hypothesis. We set $K(\alpha) = Q_n^{-1}(1 - \alpha)$ to be the $(1 - \alpha)$-th upper quantile of Q_n.

For coming patient with covariate $\boldsymbol{x}$ but without response y, we define

$$LR(\boldsymbol{x}, z) = E_{H_0}\left(\frac{\phi(y - \hat{\boldsymbol{\beta}}' \boldsymbol{x} - \hat{\eta})}{\phi(y - \hat{\boldsymbol{\beta}}' \boldsymbol{x})} \bigg| \boldsymbol{x}, z\right) \approx \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\phi(y_i - \hat{\boldsymbol{\beta}}' \boldsymbol{x} - \hat{g}(z_i) - \hat{\eta})}{\phi(y_i - \hat{\boldsymbol{\beta}}' \boldsymbol{x} - \hat{g}(z_i))},$$

where y_i ($i = 1, \ldots, n_0$) are the responses of the subjects already in the trail, and being classified to group 0, and classify this patient to group 1 if $LR(\boldsymbol{x}, z) > K(\alpha)$, with $K(\alpha)$ given above.

11.4   Simulation Study and Application

11.4.1   Simulation Study

We simulate four examples with non-linear effect of z_i to y_i. We simulate $n = 1000$ i.i.d. data with 1-dimensional response y_i's and with covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})$. We first generate the covariates, sample the $\boldsymbol{x}_i$'s from the 3-dimensional normal distribution with mean vector $\boldsymbol{\mu} = (3.1, 1.8, -0.5)'$ and a given covariance matrix Γ. sample the z_i's from the normal distribution with mean $\mu = 0$ and $\sigma^2 = 1$. The ε_i are also sampled from normal distribution with mean $\mu = 0$ and $\sigma^2 = 1$. We will display estimation results with four different choices of $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \eta_0, \lambda_0)$ and four choices of $g_0(\cdot)$ below. What is more, we fixed a point $(0, 0)$ for the non-linear effect.

Example 1 $g_0(z) = 6 \times Exponential(z + 2) - 6 \times Expnential(0 + 2)$;

Example 2 $g_0(z) = 5 \times Beta((z + 2)/4, 5, 1) - 5 \times Beta((0 + 2)/4, 5, 1)$;

Example 3 $g_0(z) = 6 \times I(z < 0) \times ((N(z, 0, 0.5)) - N(0, 0, 0.5)) + 6 \times I(z \geq 0) \times (N(z, 0, 0.2) - N(0, 0, 0.2)))$;

Example 4 $g_0(z) = 3 \times I(z < 0) \times (Beta((z + 2)/4, 0.2, 0.2) - Beta((0 + 2)/4, 0.2, 0.2)) + 7 \times I(z \geq 0) \times (Beta((z + 2)/4, 0.7, 0.7) - Beta((0 + 2)/4, 0.7, 0.7))$.

The estimated $\hat{g}$ and g_0 are shown in Fig. 11.1.

The parameter estimates from the proposed model are displayed in Tables 11.1, 11.2, 11.3 and 11.4, along with the estimates from commonly used linear model as comparison. The estimated standard errors are displayed as [se].

The hypothesis testing results from both partial linear and linear model are given in Table 11.5, and the classification results using the partial linear model are in Table 11.6.

From Table 11.5 we see that the partial linear model gives reasonable estimates, while the estimates from the linear model is not reasonable, may due to the fact that it seriously over-estimate the effect η for small value of it.

From Table 11.6, it is seen that the mis-classification error for the treatment favorable subgroup is well controlled around the specified level $\alpha = 0.05$, and the overall classification error depends on the effect size η. It is small when η is large and vice versa. Note that for $\eta = 0.95$ and 1.70, the N-P error is larger than 0.05 this is because the estimate of η is not that accurate when the true value of η is small.

Interpretation of the Results From Tables 11.1, 11.2, 11.3 and 11.4, we see that when the effect η of treatment favorable subgroup is tiny, the biases of the estimates

Example1

Example2
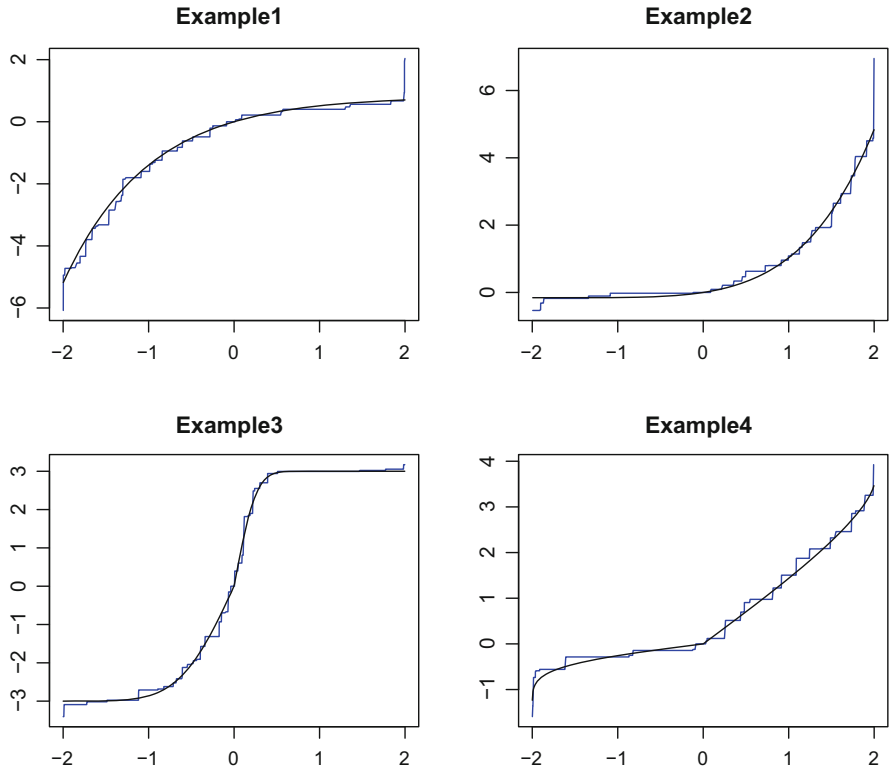
Example3

Example4

Fig. 11.1 Solid line: true $g_0(\cdot)$; Step line: estimate $\hat{g}(\cdot)$

Table 11.1 Parameter estimates under two models (example 1)

θ	β	η	λ
θ_0	$(1.300, 1.200, -1.600)$	1.650	0.700
Partial linear	$(1.291, 1.202, -1.598)$	1.619	0.723
[se]	[0.094, 0.087, 0.039]	[0.134]	[0.060]
95% CP	(0.944, 0.945, 0.949)	0.942	0.941
Linear model	$(0.995, 1.275, -1.520)$	2.053	0.653
[se]	[0.084, 0.100, 0.043]	[0.101]	[0.048]
95% CP	(0.042, 0.885, 0.552)	0.017	0.840
θ_0	$(1.200, -1.400, 3.200)$	7.740	0.300
Partial linear	$(1.198, -1.401, 3.202)$	7.740	0.300
[se]	[0.061, 0.068, 0.030]	[0.070]	[0.014]
95% CP	(0.946, 0.953, 0.949)	0.953	0.955
Linear model	$(0.966, -1.344, 3.263)$	7.742	0.300
[se]	[0.054, 0.081, 0.034]	[0.085]	[0.014]
95% CP	(0.009, 0.896, 0.541)	0.948	0.952

Table 11.2 Parameter estimates under two models (example 2)

θ	β	η	λ
θ_0	$(-1.300, 2.200, 1.400)$	1.700	0.700
Partial linear	$(-1.300, 2.202, 1.398)$	1.678	0.718
[se]	[0.085, 0.085, 0.039]	[0.126]	[0.055]
95% CP	(0.952, 0.947, 0.955)	0.951	0.946
Linear model	$(-1.033, 2.137, 1.328)$	2.029	0.549
[se]	[0.084, 0.103, 0.045]	[0.084]	[0.059]
95% CP	(0.105, 0.909, 0.643)	0.018	0.272
θ_0	$(-2.300, -1.400, 2.100)$	8.390	0.800
Partial linear	$(-2.295, -1.400, 2.098)$	8.392	0.801
[se]	[0.059, 0.070, 0.030]	[0.079]	[0.012]
95% CP	(0.942, 0.947, 0.942)	0.951	0.940
Linear model	$(-2.061, -1.458, 2.037)$	8.396	0.801
[se]	[0.061, 0.084, 0.035]	[0.099]	[0.012]
95% CP	(0.031, 0.897, 0.557)	0.949	0.939

Table 11.3 Parameter estimates under two models (example 3)

θ	β	η	λ
θ_0	$(-1.300, 1.400, 2.000)$	1.500	0.650
Partial linear	$(-1.282, 1.396, 1.995)$	1.454	0.666
[se]	[0.135, 0.090, 0.047]	[0.122]	[0.080]
95% CP	(0.955, 0.952, 0.951)	0.938	0.942
Linear model	$(-1.270, 1.389, 1.995)$	2.263	0.476
[se]	[0.083, 0.113, 0.048]	[0.079]	[0.046]
95% CP	(0.936, 0.945, 0.945)	0	0.023
θ_0	$(1.600, -1.200, 1.300)$	5.800	0.400
Partial linear	$(1.622, -1.204, 1.294)$	5.801	0.400
[se]	[0.108, 0.076, 0.039]	[0.067]	[0.015]
95% CP	(0.944, 0.941, 0.945)	0.948	0.949
Linear model	$(1.648, -1.209, 1.286)$	5.811	0.406
[se]	[0.076, 0.115, 0.046]	[0.090]	[0.015]
95% CP	(0.909, 0.956, 0.940)	0.948	0.918

from the linear model are much larger than those with the proposed partial linear model. That also can be used to explain the results of hypothesis testing with linear model. When the effect of treatment favorable subgroup is small, linear model tend to give an estimate with positive bias. So, type I error here is large and type II error is small. If the effect of treatment favorable subgroup is large, partial linear model and linear model tend to give similiar estimates of parameters.

Table 11.4 Parameter estimates under two models (example 4)

θ	β	η	λ
θ_0	(2.000,2.400,−2.500)	1.300	0.850
Partial linear	(1.977,2.403,−2.493)	1.402	0.853
[se]	[0.183,0.088,0.054]	[0.394]	[0.150]
95% CP	(0.948,0.946,0.952)	0.938	0.946
Linear model	(2.257,2.336,−2.566)	1.358	0.717
[se]	[0.116,0.083,0.041]	[0.190]	[0.122]
95% CP	(0.383,0.873,0.618)	0.931	0.825
θ_0	(−2.500,1.200, 1.700)	8.850	0.300
Partial linear	(−2.481,1.198,1.693)	8.851	0.299
[se]	[0.059,0.069,0.030]	[0.070]	[0.014]
95% CP	(0.940,0.946,0.947)	0.950	0.951
Linear model	(−2.287,1.151,1.643)	8.853	0.299
[se]	[0.049,0.072,0.030]	[0.074]	[0.014]
95% CP	(0.012,0.897,0.536)	0.945	0.951

Table 11.5 Hypothesis test using the partial linear and linear models (example 4)

η_0	Partial linear			Linear model		
	$\hat{\eta}$	Type I error	Power	$\hat{\eta}$	Type I error	Power
0	0.006	0.018		1.485	0.237	
0.02	0.003	0.016		1.496	0.234	
0.5	0.761		0.151	1.545		1
0.75	1.047		0.223	1.608		1
0.9	1.087		0.491	1.659		1
1.0	1.103		0.934	1.700		1
1.1	1.142		0.996	1.738		1
1.3	1.286		1	1.842		1

Table 11.6 Classification results using partial linear model (simulated data)

η_0	$\hat{\eta}$	Decision	Overall error	N-P Error	$K(0.05)$
0.95	0.680	H_1	0.165	0.390	1.093
1.70	1.531	H_1	0.198	0.182	1.000
3.50	3.592	H_1	0.042	0.032	1.000
5.00	5.036	H_1	0.003	0.004	1.000
7.74	7.785	H_1	0	0	1.000

11.4.2 Application to Real Data Problem

Now we analyze the real data ACTG175 with the proposed method. The trial was
conducted by the AIDS Clinical Trials Group (ACTG), which was supported by the
National Institute of Allergy and Infectious Diseases (NIAID). Participants were
enrolled into the study between December 1991 and October 1992, and received

treatment through December 1994. Follow-up and final evaluations of participants took place between December 1994 and February 1995.

The purpose of this data was to investigate whether treatment of HIV infection with one drug (monotherapy) was the same, better than, or worse than treatment with two drugs (combination therapy) in patients under some conditions.Three different drugs were used to conduct this study: (1) zidovudine (AZT), (2) didanosine (ddI), and (3) zalcitabine (ddC). The three drugs are nucleotide analogues that act as reverse transcriptase inhibitors (RT-inhibitors). The original study noted no clear differences between the ddI and AZT + ddI treatments—both appeared to be approximately equal effective in preventing HIV progressing. Treatment with AZT + ddC provided no additional benefit to continued treatment with AZT. However, the results of ACTG 175 together with the results from earlier studies demonstrate that antiretroviral therapy is beneficial to HIV-infected people who have less than 500 CD4+ T cells/mm3. This study also shows, for the first time, that an improvement in survival can be achieved in a sub-population.

We analyze this data using the proposed method on the combined therapy (ZDV+ddI). The number of patients is 522. The response variable is the CD4 counts after 20 weeks of the corresponding treatment, and the covariates are age, baseline CD4 counts, karnofsky score and number of days of previously received antiretroviral therapy. We assume the effect of baseline CD4 counts on the response variable is non-linear.

The analysis results are presented in Tables 11.7 and 11.8. We see that the null hypothesis of no subgroup is rejected, and there is a treatment favorable subgroup which is about 5% of the total patients. This is consistent with the result in Yuan et al. (2020). This case is of particular interest for hypothesis generating for developmental therapeutics. We can examine the small group of patients who are not benefiting from the treatment and identify underlying reasons and study them.

Table 11.7 Parameter estimates under two models (scaled real data)

θ	β	η	λ
Partial model	(0.073,0.0421,−0.105)	2.986	0.009
[se]	[0.042,0.041,0.044]	[0.610]	[0.031]
Linear model	(0.083,0.053,−0.123)	3.106	0.010
[se]	[0.042,0.041,0.043]	[0.585]	[0.004]

Table 11.8 Classification results (under scaled real data)

$\hat{\eta}$	Decision	$K(0.05)$	Group1-percent
2.986	H_1	0.349	0.052

11.5   Conclusion

A partial linear model is proposed for the analysis of subgroups in clinical trial, for the case one of the covariate has monotone non-linear effect on the response. The non-linear part is modeled by a monotone function along with the linear part of other covariates. The semiparametric maximum likelihood is used to estimate model parameters. Simulation study is conducted to evaluate the performance of the proposed method, and results show that the proposed model perform much better than linear models especially when treatment effect is relatively small. Then the model is applied to analyze a real data.

References

Balan TA, Putter H (2019) frailtyEM: an R package for estimating Semipaarametric shared frailty models. J Stat Softw 90(7)

Bertsekas DP (2016) Nonlinear programming, 3rd edn. Athena Scientific, Nashua

Best MJ, Chakravarti N (1990) Active set algorithms for isotonic regression; a unifying framework. Math Program 47:425–439

Bordes L, Chauveau D, Vandekerknove P (2007) A stochastic EM algorithm for a semiparametric mixture model. Comput Stat Data Anal 51:5429–5443

Campbell G (1981) Nonparametric bivariate estimation with ranodmly censored data. Biometrica 68:417–422

Chen J, Zhang D, Davidian M (2002) A Monte Carlo EM algorithm for generalized linear mixed models with ïñĆexible random effects distribution. Biometrics 3(3):347–360

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm, J R Stat Soc Ser B 39:1–38

Fan A, Song R, Lu W (2017) Change-plane analysis for subgroup detection and sample size calculation. J Am Stat Assoc 112:769–778

Fokkema M, Smits N, Zeileis A et al (2018) Detecting treatment-subgroup interactions in clustered data with generalized linear mixed- effects model trees. Behav Res Methods 50:2016–2034

Foster JC, Taylor JMC, Ruberg SJ (2011) Subgroup identification from randomized clinical trial data. Stat Med 30:2867–2880

Friede T, Parsons N, Stallard N (2012) A conditional error function approach for subgroup selection in adaptive clinical trials. Stat Med 31:4309–4320

Groeneboom P, Wellner J (1992) Information bounds and nonparametric maximum likelihood estimation. Birkháuser Verlag, Basel

Hanley JA, Parnes MN (1983) Nonparametric estimation of a multivariate distribution in the presence of censoring. Biometrics 39:129–139

International Breast Cancer Study Group (IBCSG) (2002) Endocrine responsiveness and tailoring adjuvant therapy for postmenopausal lymph node-negative breast cancer: a randomized trial. J Natl Cancer Inst 94:1054–1065

Lipkovich I, Dmitrienko A, Denne J, Enas G (2011) Subgroup identification based on differential effect search (SIDES)—A recursive partitioning method for establishing response to treatment in patient sub-populations. Stat Med 30:2601–2621

Ma S, Huang J (2017) A concave pairwise fusion approach to subgroup analysis. J Am Stat Assoc 112:410–423

Muñoz A (1980) Nonparametric estimation from censored bivariate observations. Technical Report, Department of Statistics, Stanford University

Patrick M, Kurt H, Jan DL (2009) Isotonic optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. J Stat Softw 32(5):1–24

Rothmann MD, Zhang J, Lu L, Fleming TR (2012) Testing in a pre-specified subgroup and the intent-to-treat population. Drug Inf J 46(2):175–179

Ruberg SJ, Chen L, Wang Y (2010) The mean doesn't mean as much any more: finding sub-groups for tailored therapeutics. Clin Trials 7:574–583

Sabine C (2005) AIDS events among individuals initiating HAART: do some patients experience a greater benefit from HAART than others? AIDS 19:1995–2000

Shen J, He X (2015) Inference for subgroup analysis with a structured logistic-normal mixture model. J Am Stat Assoc 110:303–312

Song Y, Chi GY (2007) A method for testing a pre-specified subgroup in clinical trials. Stat Med 26:3535–3549

Tan M, Tian G-L, Ng KW (2009) Bayesian missing data problems: EM, data augmentation and non-iterative computation. Chapman and Hall/CRC, London/Boca Raton

Yuan A, Chen X, Zhou Y, Tan MT (2018) Subgroup analysis with semiparametric models toward precision medicine. Stat Med 37(2):1830–1845

Yuan A, Zhou Y, Tan MT (2020) Subgroup analysis with a nonparametric unimodal symmetric error distribution. Comm Statist Theory Methods. Published online

Zhou Y, Yuan A, Tan MT (2019) Subgroup analysis with semiparametric partial linear regression model. Submitted to Statistical Methods in Medical Research