<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-2021-0268</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
</tr>
<tr>
<td><strong>Corresponding Author</strong></td>
</tr>
<tr>
<td><strong>E-mail</strong></td>
</tr>
</tbody>
</table>

Notice: Accepted version subject to English editing.
Bootstrap Adjustment to Minimum $p$-Value Method for Predictive Classification

Na Li, Yanglei Song, C. Devon Lin and Dongsheng Tu

Queen’s University

Abstract: In medical studies, the minimum $p$-value method is often used to determine a cutpoint of a continuous biomarker for predictive classification and to assess whether a subset of patients may have different treatment effect than other patients. This method, however, suffers from an issue of type I error inflation when the estimated cutpoint is treated as known. In this paper, we propose bootstrap-based procedures to obtain the valid $p$-value for the minimum $p$-value test statistic when the treatment effect is measured by a continuous outcome under both random and fixed designs, regardless of whether the cutpoint is identifiable. In the fixed design case, the test statistic is the supremum of a non-centered random process, whose mean function (i.e. bias) diverges as the sample size goes to infinity even under the null hypothesis. The proposed bootstrap statistic matches asymptotically the diverging bias, and we apply the high-dimensional Gaussian approximation results to establish the asymptotic size validity, as well as the power consistency under local alternatives. The proposed method is applied to a dataset from a clinical trial on advanced colorectal cancer.

Key words and phrases: High-dimensional Gaussian Approximation, Minimum
1. Introduction

In clinical practices, it is common to classify patients into two or more groups based on a demographic, clinical, or genetic variable, such as age and expression level or status of a gene, to which we refer as a biomarker, to make clinical decisions. There are two types of problems: one is to classify patients with respect to a clinical outcome of interest, regardless of the types of treatments, for risk stratification, and the other is to distinguish patients by their degree of benefit or harm to a particular treatment for guidance to its adoption. In the clinical literature, the former is called as *prognostic classification*, while the latter as *predictive classification*, and the biomarker used for these classifications is called respectively as prognostic and predictive biomarker ([Ballman, 2015](#)).

This paper is concerned about the identification and assessment of *predictive biomarkers* based on data from clinical trials for testing an experimental treatment against a standard control. Consider the CO.17 trial conducted by the Canadian Cancer Trials Group ([Jonker et al., 2007](#)) as an example, which is a randomized phase III trial to compare cetuximab plus best supportive care (BSC) with BSC alone in patients with metastatic...
epidermal growth factor receptor-positive colorectal cancer. Investigators identified K-ras gene as a predictive biomarker with respect to various clinical outcomes, such as the change in global health status score from baseline at 8 and 16 weeks after randomization, and found patients with wild-type K-ras benefited more from cetuximab treatment than patients with mutated K-ras [Karapetis et al. 2008]. As a result, the cetuximab treatment is now restricted to patients with tumor bearing wild-type K-ras.

We consider in this paper the problem of predictive classification with respect to a continuous clinical outcome when only the continuous measurements of a biomarker are available but a cutpoint is required to classify patients into two groups. To be more concrete, let $Y$ be a continuous clinical outcome of interest, $U$ a binary treatment indicator, and $X$ a continuous biomarker. We assume

$$E(Y) = \alpha_0 + \beta_0 U + \gamma_0 I(X \leq c_0) + \lambda_0 I(X \leq c_0)U,$$ (1.1)

where $c_0$ is an unknown cutpoint in a range $[\ell, u]$, $I(\cdot)$ is the indicator function, and $\alpha_0, \beta_0, \gamma_0, \lambda_0$ are unknown parameters. Then $\beta_0 + \lambda_0$ and $\beta_0$ represent respectively the treatment effect, i.e., the difference in the expected outcome between the experiment and standard treatment groups, in the two subsets defined by the cutpoint $c_0$, and thus $\lambda_0$ measures the differential treatment effect between these two subsets. One important task
for *predictive classification* is to assess whether the differential treatment effect is significant by testing the null hypothesis $H_0 : \lambda_0 = 0$. Note that we make no assumption about whether $\gamma_0$ is non-zero or not. If $\gamma_0 = \lambda_0 = 0$, the cutpoint is not identifiable in the sense that all values of $c_0$ induce the same distribution on the response $Y$.

If the value of the cutpoint $c_0$ was known to be $c$, e.g., based on subject knowledge or from previous studies, classical test statistics, such as the Wald test statistic $M_{n,c}$ based on a sample of size $n$, can be used to test $H_0$, which is however unrealistic for many clinical applications. In practice, the *minimum p-value method* is often used for the assessment of both prognostic and predictive biomarkers when the cutpoint is unknown. Its basic idea in the context of predictive classification can be described as follows. If $c$ was the true value for $c_0$, the associated Wald statistic $M_{n,c}$ would have a limiting standard normal distribution $N(0, 1)$, and thus $2\{1 - \Phi(|M_{n,c}|)\}$ would be an asymptotically valid $p$-value for testing $H_0 : \lambda_0 = 0$, where $\Phi(\cdot)$ is the distribution function of $N(0, 1)$. When the true value of $c_0$ is unknown, define $\tilde{c}_0$ to be a value that achieves the minimum $p$-value or equivalently the maximal absolute Wald statistic:

$$
\tilde{c}_0 = \arg\min_{c \in [\ell, u]} 2\{1 - \Phi(|M_{n,c}|)\} = \arg\max_{c \in [\ell, u]} |M_{n,c}|. \quad (1.2)
$$
Then the $p$-value from the minimum $p$-value method is

$$p_{n,mp} = 2\{1 - \Phi(M_n)\}, \quad \text{where} \quad M_n = |M_{n,\tilde{c}_0}| = \sup_{c \in [\ell, u]} |M_{n,c}|.$$

(1.3)

Although this method is simple and appealing to practitioners, its Type I error is substantially inflated since the definition in (1.3) does not take into account that the cutpoint is estimated from the data. Due to this issue, resulting analyses are not generally recognized in medical literature and, for those eventually published, for example, [Jonker et al. 2014] and [Blok et al. 2018], the analyses can only be considered as exploratory. Therefore, statistical methods for the adjustment of the $p$-value calculated from this method are urgently needed.

Our main contribution in this paper is to propose bootstrap methods to adjust the $p$-value defined in (1.3) under both the random design and the fixed design (see, e.g., [Freedman 1981] for a discussion), and establish the asymptotic size validity as well as the power consistency. We emphasize that the critical values obtained by the proposed bootstrap methods lead to proper Type I error control under both the identifiable ($\gamma_0 \neq 0$) and non-identifiable case ($\gamma_0 = 0$). These two cases are both practically important, and usually there is no convincing reason to assume one case or the other. However, in the literature (reviewed below), there are valid tests for each separate case, but not for both.
Specifically, under the random design, both the biomarker $X$ and the treatment indicator $U$ are viewed as random variables. If $U$ and $X$ are independent, $E(M_{n,c}) \to 0$ as $n \to \infty$ for each $c \in \mathbb{R}$ under the null $H_0 : \lambda_0 = 0$. By standard empirical processes arguments (Van Der Vaart and Wellner 1996), we show that $\{M_{n,c} : c \in [\ell, u]\}$ converges in distribution to a zero-mean Gaussian process, and propose a paired bootstrap procedure to approximate the distribution of the limiting process.

The adjustment under a fixed design, which assumes $(U, X)$ as deterministic, is non-standard and statistically more challenging. Specifically, the minimum $p$-value test statistic $M_n$ is the supremum of the absolute values of the random process, $\{M_{n,c} : c \in [\ell, u]\}$; its mean function may diverge as $n \to \infty$, which makes the classical functional weak convergence theory (Van Der Vaart and Wellner 1996) inapplicable. Further, although Chernozhukov, Chetverikov, and Kato (2014, 2016) develop the Gaussian approximation and bootstrap tools for non-centered, non-Donsker empirical processes, $\{M_{n,c} : c \in [\ell, u]\}$ is not an empirical process. To address the issue of “diverging bias” under the fixed design, we propose a multiplier residual bootstrap for which the bias of the bootstrap process $\{M^*_{n,c} : c \in [\ell, u]\}$ asymptotically matches that of $\{M_{n,c} : c \in [\ell, u]\}$, and thus also diverges. To establish the size validity of the proposed test, we ob-
serve that $M_n$ in (1.3) can be viewed as the supremum of the absolute values of $\{M_{n,c} : c \in C_n\}$ with $C_n = \{X_1, \ldots, X_n\} \cap [\ell, u]$, a non-centered random vector whose dimension grows with $n$, and we use the high-dimensional Gaussian approximation results (Chernozhukov et al., 2013, 2017, 2019) to show that its distribution is well approximated by the bootstrap counterpart $\{M_{n,c}^* : c \in C_n\}$. Further, since the test statistic $M_n$ possibly diverges under the null, it is questionable whether a test based on $M_n$ will have any power. We show that the proposed procedure is asymptotically consistent under local alternatives.

Next we discuss the relevant statistical literature on the minimum $p$-value method. For the prognostic classification, which in the setup of the model in (1.1) assumes $\lambda_0 = 0$ and tests the null hypothesis $H'_0 : \gamma_0 = 0$, there is a rich literature on the adjustment of the minimum $p$-value method; see, for example, Miller and Siegmund (1982); Jespersen (1986); Lausen and Schumacher (1992) and a comprehensive review by Mazumdar and Glassman (2000). Further, Fan et al. (2017) recently considers testing and identifying a subgroup with an enhanced treatment effect by testing $H_0 : \lambda_0 = 0$ in a model similar to (1.1) but assumes that $\gamma_0 = 0$. In both literature related to prognostic classification and Fan et al. (2017), the setup is non-standard in the sense that $c_0$ is not identifiable under the null (Davies).
1977, 1987; Andrews, 2001); nonetheless, under the random design, the test statistics therein converge in distribution under their respective nulls. In the current paper, we do not assume whether the model is identifiable or not; in particular, we allow \( c_0 \) to be non-identifiable. A further challenge in our setup and analysis is that under the fixed design, the test statistic \( M_n \) is not bounded in probability under the null. In addition, a few studies (Jiang et al., 2007; He, 2014; Gavanji et al., 2018; Götte et al., 2020) consider adjustments to the minimum \( p \)-value statistics for survival endpoints in the context of predictive classification. However, no theoretical justification has been provided for the size validity of the adjusted tests.

We should also mention the rich literature on the change-point or cut-point estimation (Koul et al., 2003; Seijo and Sen, 2011; Mallik et al., 2011; Yu, 2014; Li and Jin, 2018; Mukherjee et al., 2020), which provides a valid test for the null \( H_0 : \lambda_0 = 0 \) under the assumption that the model is identifiable. Specifically, denote by \( \hat{c}_0 \) the profile least squares estimator for \( c_0 \) in (1.1); see its precise definition in (4.11). If \( \gamma_0 \neq 0 \), under the random design, Koul et al. (2003) shows that \( n(\hat{c}_0 - c_0) \) converges in distribution to the minimizer of a compound Poisson process. Seijo and Sen (2011) and Yu (2014) show that the conventional bootstrap methods, such as the paired bootstrap and the residual bootstrap, are inconsistent, and propose valid smoothed
bootstrap methods for constructing confidence intervals for \( c_0 \). Note however that the failure of conventional bootstrap methods does not contradict with our work, as we study a distinct problem for a similar model; see Section \( \text{S5} \) in the Supplementary Materials (SMs, hereafter) for details. In this study, we show that the \( p \)-value based on \( \hat{c}_0 \), \( p_{n,pf} = 2\{1 - \Phi(|M_{n,\hat{c}_0}|)\} \), has an asymptotically valid size, under the assumption \( \gamma_0 \neq 0 \). Without this identifiability assumption, the Type I error of \( p_{n,pf} \) is significantly inflated (see Section \( 5 \)).

The remainder of the paper is organized as follows. In Section \( 2 \) we formally introduce the model, as well as the minimum \( p \)-value method. In Sections \( 3 \) and \( 4 \) we propose paired and multiplier residual bootstrap methods for the random and fixed designs respectively. We present simulation results in Section \( 5 \) to corroborate our theory, and an analysis of an advanced colorectal cancer dataset in Section \( 6 \).

2. Problem Formulation

Let \( (Y_i, U_i, X_i), i \in [n] := \{1, \ldots, n\} \) be a sample of \( n \) observations, where \( Y_i \in \mathbb{R} \) is a continuous outcome, \( U_i \in \{0, 1\} \) is a binary treatment indicator, and \( X_i \in \mathbb{R} \) is a continuous biomarker. For a given \( c \in \mathbb{R} \), define \( X_{i,c} = I(X_i \leq c) \), where \( I(\cdot) \) is the indicator function. Next we expand on the
model in (1.1) and assume the following dependence of the outcome $Y_i$ on the covariates $X_i$ and $U_i$:

$$Y_i = \alpha_0 + \beta_0 U_i + \gamma_0 X_{i,c_0} + \lambda_0 X_{i,c_0} U_i + \epsilon_i \quad \text{for } i \in [n], \quad (2.4)$$

where $c_0 \in \mathbb{R}$ is an unknown cutpoint, $\theta_0 = (\alpha_0, \beta_0, \gamma_0, \lambda_0)^T$ is a vector of unknown regression parameters and $\epsilon_1, \ldots, \epsilon_n$ are observation noise, which are assumed to be independent and identically distributed with mean 0 and unknown variance $\sigma^2$. If $\gamma_0 = \lambda_0 = 0$, $c_0$ in (2.4) is not identifiable, in which case we assume $c_0 = -\infty$ without loss of generality. In the case that either $\gamma_0 \neq 0$ or $\lambda_0 \neq 0$, we assume that $c_0$ is apriori known to be in an interval $[\ell, u]$. Note that we make no assumption about whether $c_0$ is identifiable, and require a valid test to control the Type I error properly in both cases.

2.1 Minimum $p$-value method

For any $c \in \mathbb{R}$ and $\theta \in \mathbb{R}^4$, denote by $\text{RSS}_c(\theta)$ the associated residual sum of squares, i.e.,

$$\text{RSS}_c(\theta) = \sum_{i=1}^{n} (Y_i - Z_{i,c}^T \theta)^2, \quad \text{where } Z_{i,c} = (1, U_i, X_{i,c}, X_{i,c} U_i)^T.$$
Then for each $c \in \mathbb{R}$, the least squares estimator $\hat{\lambda}_c$ of $\lambda_0$ and its estimated variance $\hat{v}_c^2/n$ can be respectively written as:

$$
\hat{\lambda}_c = d^T \hat{\theta}_c, \quad \text{and} \quad \hat{v}_c^2/n = \left(d^T \hat{Q}_n^{-1}d\right) \text{RSS}_c(\hat{\theta}_c)/\{n(n - 4)\},
$$

(2.5)

where $\hat{\theta}_c = \frac{1}{n} \hat{Q}_n^{-1} \sum_{i=1}^n Z_{i,c} Y_i$, $\hat{Q}_n,c = \frac{1}{n} \sum_{i=1}^n Z_{i,c} Z_{i,c}^T$, and $d = (0, 0, 0, 1)^T$. Here $\hat{\theta}_c$ is the least squares estimator for $\theta_0$ for a fixed $c$, in the sense of minimizing $\text{RSS}_c(\theta)$.

Recall that the task of the predictive classification is to test the null hypothesis $H_0 : \lambda_0 = 0$. If a particular $c$ was treated as the true value for $c_0$, the Wald test statistic for $H_0 : \lambda_0 = 0$ would be $M_{n,c}$ in (2.6) below. Recall that $\tilde{c}_0$ in [1,2] achieves the smallest $p$-value or equivalently the largest absolute Wald statistic among $c \in [\ell, u]$. Our goal is to develop a valid test by calibrating the distribution of the following minimum $p$-value test statistic or maximally selected Wald test statistic:

$$
M_n = |M_{n,\tilde{c}_0}| = \sup_{c \in [\ell, u]} |M_{n,c}|, \quad \text{where} \quad M_{n,c} = \sqrt{n} \hat{\lambda}_c/\hat{v}_c.
$$

(2.6)

**Remark 1.** In computing the test statistic $M_n$ in (2.6), it suffices to take the supremum over the distinct values of the sample $X_1, \ldots, X_n$, i.e., those $c \in C_n = \{X_1, \ldots, X_n\} \cap [\ell, u]$.

**Remark 2.** If $c_0$ is known, under the random design setup, $M_{n,c_0}$ has a standard normal limiting distribution only under the homogeneous case,
i.e., $\text{VAR}(\epsilon_1|X_1,U_1) = \sigma^2$. Nonetheless, our proposed procedure in Section 3 is applicable to the heterogeneous case as well.

### 2.2 Challenges in calibrating the distribution of $M_n$

Consider the following decomposition of $n^{1/2}\hat{\lambda}_c$ for each $c \in [\ell,u]$,

$$n^{1/2}\hat{\lambda}_c = n^{-1/2}\tilde{d}_c^T \sum_{i=1}^{n} Z_{i,c}Z_{i,c}^T \theta_0 + n^{-1/2}\tilde{d}_c^T \sum_{i=1}^{n} Z_{i,c}\epsilon_i = I_{n,c} + II_{n,c}, \quad (2.7)$$

where $\tilde{d}_c = \hat{Q}_{n,c}^{-1}d$, and $\hat{Q}_{n,c}$ and $d$ are defined in (2.5). Under both the random and fixed designs, the second term is centered, i.e., $E(II_{n,c}) = 0$ for $c \in [\ell,u]$. As discussed in the Introduction section, the main challenge lies in the analysis of the fixed design.

Specifically, under the fixed design, $(U_i, X_i), i \in [n]$ are deterministic vectors, thus $I_{n,c}$ is a deterministic sequence which may diverge as $n \to \infty$. In fact, if $(U_i, X_i), i \in [n]$ is one realization of an independently and identically distributed sequence – fixed once generated, then by the law of the iterated logarithm, for any $c \neq c_0$, $|I_{n,c}|$ diverges at a rate of $(\log \log(n))^{1/2}$ almost surely. As a result, $\{E(M_{n,c}) : c \in [\ell,u]\}$ possibly diverges as $n \to \infty$ even under the null, and so does the test statistic $M_n$.

Next we develop asymptotically valid bootstrap-based tests under the random design and fixed design, respectively in Sections 3 and 4. We note that by a conditional argument, a test that has a valid size under the fixed
design setup remains to do so under the random design, if \( \epsilon_1 \) is independent from \( X_1 \) and \( U_1 \).

### 3. Paired Bootstrap for the Random Design

First we consider the random design setup, which assumes \((Y_i, U_i, X_i), i \in [n] \) in (2.4) are independent and identically distributed random vectors with \( E(\epsilon_1|U_1, X_1) = 0 \). Denote by \( F \) the distribution function of \( X_1 \), and by \( p = E(U_1) \) the expected value of the binary treatment \( U_1 \). We assume that \( 0 < F(\ell) < F(u) < 1 \), and \( 0 < p < 1 \).

Denote by \( \ell^\infty([\ell, u]) \) the space of bounded functions on \([\ell, u]\) equipped with the \( \ell_\infty \) norm. Theorem 1 below states that \( \{M_{n,c} : c \in [\ell, u]\} \), appearing in the maximally selected Wald test statistic (2.6), converges weakly in \( \ell^\infty([\ell, u]) \) to a tight, centered Gaussian process. A brief review of the definition of the functional weak convergence, as well as the proofs for Theorems 1 and 2 below, can be found in Section S1 in the SMs.

**Theorem 1.** Assume that \( U_1 \) and \( X_1 \) are independent, and that \( H_0 : \lambda_0 = 0 \) holds. There exists a tight, zero mean Gaussian process, \( G = \{G_c : c \in [\ell, u]\} \), such that \( \{M_{n,c} : c \in [\ell, u]\} \) converges weakly in \( \ell^\infty([\ell, u]) \) to \( G \).

Since the Gaussian process \( G \) has a complicated covariance structure, we propose the following paired bootstrap method to obtain the asymptot-
ically valid $p$-value for the test statistic $M_n$ in (2.6).

Let $(Y_1^*, U_1^*, X_1^*)$, \ldots, $(Y_n^*, U_n^*, X_n^*)$ be a random sample with replacement from the data $(Y_1, U_1, X_1), \ldots, (Y_n, U_n, X_n)$. For each $c \in [\ell, u]$, define the bootstrap least squares estimate $\hat{\lambda}_c^*$ using (2.5) with $(Z_{i,c}, Y_i)$ replaced by $(Z_{i,c}^*, Y_i^*)$, where $Z_{i,c}^* = (1, U_{i,c}^*, U_{i,c}^* X_{i,c}^*)$ and $X_{i,c}^* = I(X_i^* \leq c)$. Further, define the following bootstrap version of the maximally selected Wald test statistic, $M_n^*$:

$$M_n^* = \sup_{c \in [\ell, u]} |M_{n,c}^* - M_{n,c}|, \quad \text{where } M_{n,c}^* = \sqrt{n \hat{\lambda}_c^* / \hat{v}_c}, \quad \text{for } c \in [\ell, u]. \quad (3.8)$$

Denote by $F_{n,pb}^*$ the distribution function of the bootstrap test statistic $M_n^*$, conditional on the data $(Y_i, U_i, X_i), i \in [n]$. The adjusted $p$-value based on the paired bootstrap is defined as

$$p_{n,pb}^* = 1 - F_{n,pb}^*(M_n). \quad (3.9)$$

In practice, $F_{n,pb}^*$ is approximated by the empirical distribution of realizations of $M_n^*$ based on $B$ bootstrap samples.

**Remark 3.** In the paired bootstrap literature (Shao and Tu 2012), for each $c$, $\hat{v}_c^*$, which is computed using (2.5) with $(Z_{i,c}, Y_i)$ replaced by $(Z_{i,c}^*, Y_i^*)$, is often used in the denominator to standardize $\hat{\lambda}_c^*$ in (3.8). Here, we propose to use $\hat{v}_c$ mainly for simplicity of the proof.
The following theorem establishes the asymptotic validity of the adjusted $p$-value in (3.9) obtained by the proposed paired bootstrap method under the random design setup.

**Theorem 2.** Assume that $U_1$ and $X_1$ are independent, and that $H_0 : \lambda_0 = 0$ holds. Conditional on $(Y_i, U_i, X_i)$ ($i = 1, 2, \ldots$), for almost every sequence $(Y_i, U_i, X_i)$ ($i = 1, 2, \ldots$), the random process $\{M_{n,c}^{*} - M_{n,c} : c \in [\ell, u]\}$ converges weakly in $\ell^\infty([\ell, u])$ to the same tight, zero mean Gaussian process $G$ as in Theorem 1. Consequently, for any significance level $\xi \in (0, 1)$,

$$\lim_{n \to \infty} \text{pr} \left( p_{n,pb}^{*} \leq \xi \right) = \xi. \quad (3.10)$$

4. **Multiplier Residual Bootstrap for the Fixed Design**

In this section, we consider the fixed design setup, where $(U_i, X_i), i \in [n]$ are deterministic and the randomness only comes from the observation noise $\epsilon_i, i \in [n]$. As a result, the first term $I_{n,c}$ in (2.7) is a deterministic function of $c$, viewed as a bias term that needs to be removed. We propose the following multiplier residual bootstrap (Efron 1979; Wu 1986; Shao and Tu 2012) to obtain the asymptotically valid $p$-value for the test statistic $M_n$ in (2.6). Recall the definitions of $\text{RSS}_c(\theta), \hat{\theta}_c,$ and $\hat{\nu}_c^2/n$ in Section 2.1.
STEP 1. Define the profile least squares estimator $\hat{c}_0$ for $c_0$ and $\hat{\sigma}^2$ for $\sigma^2$:

$$\hat{c}_0 = \arg\min_{c \in [\ell, u]} \text{RSS}_c(\hat{\theta}_c), \quad \hat{\sigma}^2 = \text{RSS}_{\hat{c}_0}(\hat{\theta}_{\hat{c}_0})/(n - 4). \tag{4.11}$$

Thus, $(\hat{c}_0, \hat{\theta}_{\hat{c}_0})$ achieves the smallest residual sum of squares, i.e., minimizing $\text{RSS}_c(\theta)$ over all $(c, \theta) \in [\ell, u] \times \mathbb{R}^4$.

STEP 2. Let $\zeta_1, \ldots, \zeta_n$ be independent and identically distributed standard normal random variables, that are independent of the data $Y_i, i \in [n]$. Define the bootstrap sample as $(Z_i, Y_i^*), i \in [n]$ with

$$Y_i^* = \hat{\alpha}_0 + \hat{\beta}_0 U_i + \hat{\gamma}_0 X_{i, c_0} + \hat{\sigma} \zeta_i, \tag{4.12}$$

where $\hat{\alpha}_0$, $\hat{\beta}_0$, and $\hat{\gamma}_0$ are the first three components of $\hat{\theta}_{c_0}$.

STEP 3. For a fixed $c$, define the least squares estimator, $\hat{\lambda}_c^*$, for the bootstrap sample $(Z_i, Y_i^*), i \in [n]$ using (2.5) with $Y_i$ replaced by $Y_i^*$. Further, define the bootstrap test statistic, $M_n^*$:

$$M_n^* = \sup_{c \in [\ell, u]} |M_{n, c}|, \quad \text{where} \quad M_{n, c} = \sqrt{n} \hat{\lambda}_c^*/\hat{v}_c.$$

STEP 4. Denote by $F_{n, mrb}^*$ the distribution function of $M_n^*$, conditional on the data $Y_i, i \in [n]$, and define the adjusted $p$-value as

$$p_{n, mrb}^* = 1 - F_{n, mrb}^*(M_n). \tag{4.13}$$
The conditional (on $Y_i$'s) distribution, $F_{n,mrb}^*$, of $M_n^*$ can be estimated by the bootstrap, i.e., by repeatedly generating independent realizations of the multipliers $\zeta_i, i \in [n]$. In computing the profile least squares estimator $\hat{c}_0$ for $c_0$, as in Remark 1, it suffices to consider $c \in C_n$.

### 4.1 Asymptotic size validity

Next, we establish the asymptotic validity of the $p$-value obtained from the above multiplier residual bootstrap test procedure. We consider the following asymptotic regime. Assume that $(X_i, U_i), i \in [n]$ are deterministic, and may depend on $n$. That is, we consider the triangle array setup where $X_i = X_{n,i}, U_i = U_{n,i}$ for $i \in [n]$, but for notational simplicity, we omit the dependence on $n$. However, the distribution of $\epsilon_i, i \in [n]$ and $(c_0, \theta_0)$ in (2.4) do not depend on $n$, except when we consider the local alternatives in Subsection 4.2.

Denote $s_{n,c} = \sum_{i=1}^{n} X_{i,c}/n, p_n = \sum_{i=1}^{n} U_i/n, and q_{n,c} = \sum_{i=1}^{n} X_{i,c}U_i/n$.

For some $r \in (4, \infty]$ that needs to be specified later, we impose the following assumptions, with the convention $1/\infty = 0$.

(A.1) If $r = \infty$, assume the existence of a constant $\rho > 0$ such that $E(e^{\rho \epsilon_1}) \leq \exp(\rho^2 t^2/2)$ for $t \in \mathbb{R}$. If $r < \infty$, assume that $E(|\epsilon_1|^r) < \infty$.

(A.2) There exist a non-decreasing function $F : [\ell, u] \rightarrow (0, 1)$, and con-
stants \( \eta \in (0, 1/2 - 1/r) \) and \( 0 < p < 1 \) such that as \( n \to \infty \),

\[
n^{1 - \eta} |p_n - p| + n^{1 - \eta} \sup_{c \in [\ell, u]} \left[ |s_{n,c} - F(c)| + |q_{n,c} - pF(c)| \right] \to 0.
\]

If \( \gamma_0 \neq 0 \), \( c_0 \) is identified in the model (2.4), in which case we further impose the following assumption. It is not assumed if \( \gamma_0 = \lambda_0 = 0 \).

(A.3) \( F \) is differentiable at \( c_0 \) with a positive derivative, and there exists an \( \eta_1 \in (1/2, 1 - 2/r) \) such that for any constant \( K > 0 \), as \( n \to \infty \),

\[
n^{\eta_1} \sup_{|c - c_0| \leq Kn^{-\eta_1}} |s_{n,c} - s_{n,c_0} - \{F(c) - F(c_0)\}| \to 0,
\]

\[
n^{\eta_1} \sup_{|c - c_0| \leq Kn^{-\eta_1}} |q_{n,c} - q_{n,c_0} - p\{F(c) - F(c_0)\}| \to 0.
\]

Assumption (A.1) requires the noise \( \epsilon_1 \) to have a sub-Gaussian tail if \( r = \infty \), and a finite \( r \)-th moment if \( r < \infty \); Remark 4 explains why we consider these two cases separately. Assumptions (A.2) and (A.3) concern the global and local (around \( c_0 \)) convergence rate of \( s_{n,c}, q_{n,c}, p_n \) to \( F(c), pF(c), p \), respectively. We provide examples in Subsection 4.3 that satisfy (A.2) and (A.3).

Next we establish that the distribution of the adjusted \( p \)-value converges to the uniform distribution over \((0, 1)\) uniformly at a polynomial rate.

**Theorem 3.** Let \( r \in (4, \infty] \). Suppose that the null hypothesis \( H_0 : \lambda_0 = 0 \) holds, and that Assumptions (A.1) and (A.2) hold. If \( \gamma_0 \neq 0 \), suppose that
Assumption (A.3) is satisfied. If \( \gamma_0 = 0 \), let \( \eta_1 = 1 \). Then

\[
\lim_{n \to \infty} n^q \sup_{\xi \in (0,1)} \left| \Pr(p^*_{n, mrb} \leq \xi) - \xi \right| = 0,
\]

for any \( q < \min\{1/6 - 1/(3r), 1/3 - 4/(3r), \eta_1 - 1/2, 1/2 - 1/r - \eta_0 \} \) if \( r < \infty \), and \( q < \min\{1/4, \eta_1 - 1/2, 1/2 - \eta_0 \} \) if \( r = \infty \).

Proof. Here we outline the strategy, and defer the complete proof to Section S2.1 in the SMs. Recall in (4.12) that \( \hat{\alpha}_0, \hat{\beta}_0, \hat{\gamma}_0 \) are the first three components of \( \hat{\theta} \hat{c}_0 \), associated with \( \hat{c}_0 \) in (4.11); let \( \hat{\theta}_0 = (\hat{\alpha}_0, \hat{\beta}_0, \hat{\gamma}_0, 0)^T \). Similar to (2.7), for each \( c \in [\ell, u] \),

\[
\sqrt{n} \hat{\lambda}_c = n^{-1/2} \tilde{d}_c^T \sum_{i=1}^{n} Z_{i,c} \hat{Z}_{i,\hat{c}_0} \hat{\theta}_0 + n^{-1/2} \tilde{d}_c^T \sum_{i=1}^{n} Z_{i,c} \zeta_i = I_{n,c}^* + II_{n,c}^*,
\]

where \( \tilde{d}_c = \hat{Q}_{n,c}^{-1} d \), and \( \hat{Q}_{n,c} \) and \( d \) are defined in (2.5).

As in Remark 1, the supremum of \( M_{n,c} \) or \( M_{n,c}^* \) over \( c \in [\ell, u] \) is equal to that over \( c \in C_n = \{X_1, \ldots, X_n\} \cap [\ell, u] \). As a result, conditional on \( Y_i, i \in [n] \), or equivalently on \( \epsilon_i, i \in [n] \), the distribution function \( F_{n, mrb}^* \) of \( M_n^* \) is continuous and strictly increasing on \([0, \infty)\), and we have

\[
\sup_{\xi \in (0,1)} \left| \Pr(p^*_{n, mrb} \leq \xi) - \xi \right| = \sup_{\xi \in (0,1)} \left| \Pr\{M_n \geq (F_{n, mrb}^*)^{-1}(1 - \xi)\} - \xi \right|
\]

\[
= \sup_{\xi \in (0,1)} \left| \Pr\{M_n \geq (F_{n, mrb}^*)^{-1}(1 - \xi)\} - \Pr_{|\epsilon}\{M_n^* \geq (F_{n, mrb}^*)^{-1}(1 - \xi)\} \right|
\]

where \( \Pr_{|\epsilon} \) denotes the conditional probability given \( \epsilon_i, i \in [n] \). By the
triangle inequality, it is upper bounded by $\Upsilon_1 + \Upsilon_2 + \Upsilon_3$, where

\[
\Upsilon_1 = \sup_{t > 0} \left| \Pr \left( \sup_{c \in \mathcal{C}_n} \left| \frac{I_{n,c} + II_{n,c}}{\hat{v}_c} \right| \leq t \right) - \Pr \left( \sup_{c \in \mathcal{C}_n} \left| \frac{I_{n,c} + (\sigma/\hat{\sigma})II^*_n}{\hat{v}_c} \right| \leq t \right) \right|,
\]

\[
\Upsilon_2 = \sup_{t > 0} \left| \Pr \left( \sup_{c \in \mathcal{C}_n} \left| \frac{I_{n,c} + (\sigma/\hat{\sigma})II^*_n}{\hat{v}_c} \right| \leq t \right) - \Pr_{\epsilon} \left( \sup_{c \in \mathcal{C}_n} \left| \frac{I_{n,c} + II^*_n}{\hat{v}_c} \right| \leq t \right) \right|,
\]

\[
\Upsilon_3 = \sup_{t > 0} \left| \Pr_{\epsilon} \left( \sup_{c \in \mathcal{C}_n} \left| \frac{I_{n,c} + II^*_n}{\hat{v}_c} \right| \leq t \right) - \Pr_{\epsilon} \left( \sup_{c \in \mathcal{C}_n} \left| \frac{I_{n,c} + II^*_n}{\hat{v}_c} \right| \leq t \right) \right|.
\]

Denote by $n_c$ the cardinality of $\mathcal{C}_n$, and define $II_n = \{II_{n,c} : c \in \mathcal{C}_n\}$ and $II^*_n = \{II^*_{n,c} : c \in \mathcal{C}_n\}$ to be two $n_c$-dimensional random vectors. Further, denote by $A^{re}$ the collection of all hyper-rectangles in $\mathbb{R}^{n_c}$; that is, $A^{re}$ consists of all sets $A$ of the form $A = \{x \in \mathbb{R}^{n_c} : t_i \leq x_i \leq s_i$, for all $1 \leq i \leq n_c\}$, for some $-\infty \leq t_i \leq s_i \leq \infty$ for $1 \leq i \leq n_c$.

First, $\Upsilon_1 \leq \sup_{A \in A^{re}} |\Pr (II_n \in A) - \Pr ((\sigma/\hat{\sigma})II^*_n \in A)|$, where the random vector $\sigma II^*_n/\hat{\sigma}$ is a zero mean Gaussian vector (of length $n_c$) with the same covariance matrix as $II_n$. Thus, we apply the high-dimensional central limit theorem (Chernozhukov et al., 2019, Theorem 2.1) if $r = \infty$, and (Chernozhukov et al., 2017, Proposition 2.1) if $r < \infty$ to show that it converges to 0 as $n \to \infty$.

Second, $\Upsilon_2 \leq \sup_{A \in A^{re}} |\Pr_{\epsilon} (II^*_n \in A) - \Pr ((\sigma/\hat{\sigma})II^*_n \in A)|$. Conditional on $\epsilon, i \in [n]$, $II^*_n$ is also a centered Gaussian vector, and we show that the supremum difference between the conditional covariance matrix of $II^*_n$ and the covariance matrix of $\sigma II^*_n/\hat{\sigma}$ vanishes at a polynomial rate.
as $n \to \infty$ almost surely. Then we apply the (high-dimensional) Gaussian comparison theorem (Chernozhukov et al., 2019, Corollary 5.1) to show that it vanishes almost surely as $n \to \infty$.

The third term $\Upsilon_3$ can be clearly bounded by

$$\sup_{s_1, s_2 \in \mathbb{R}^{n_c}} \{ \text{pr}_\epsilon ([II_n^*, -II_n^*] \leq [s_1, s_2] + \Delta_n) - \text{pr}_\epsilon ([II_n^*, -II_n^*] \leq [s_1, s_2]) \},$$

where $\Delta_n = \sup_{c \in \mathcal{C}_n} |I_{n,c} - I_{n,c}^*|$, and both the inequalities and the scalar addition are interpreted component-wise. Under the fixed design setup, $I_{n,c}$ and $I_{n,c}^*$ are deterministic; we show that $\Delta_n$ decays at a polynomial rate as $n \to \infty$. Finally, we apply the Nazarov’s inequality (Nazarov, 2003; Chernozhukov et al., 2017) to the Gaussian vector $[II_n^*, -II_n^*]$ of length $2n_c$ to establish that the above term vanishes almost surely as $n \to \infty$.

Remark 4. For the examples in Subsection 4.3 (A.2) holds for any $\eta_0 \in (0, 1/2 - 1/r)$ and (A.3) holds for any $\eta_1 \in (1/2, 1 - 2/r)$. Thus if $r = \infty$, the approximation error for the size of the adjusted $p$-value vanishes at a faster rate than $n^{-q}$ as $n \to \infty$ for any $q < 1/4$.

Note that the rate for the $r = \infty$ case cannot be recovered from the finite $r$ result by letting $r \to \infty$. This is due to the availability of improved rates in the high-dimensional central limit theorem (Chernozhukov et al., 2019, Theorem 2.1) if the noise $\epsilon_1$ has a sub-Gaussian tail.
4.2 Consistency under alternatives

In practice, if the null hypothesis is rejected, the minimum \( p \)-value estimator \( \tilde{c}_0 \) is commonly used as an estimator for the cutpoint \( c_0 \). Since the minimizer of the \( p \)-values of the Wald test statistics is not unique, we define

\[
\tilde{c}_0 = \arg\max_{c \in \mathcal{C}_n} |M_{n,c}| ,
\]

where if there are multiple \( c \in \mathcal{C}_n \) achieving the maximum, we define \( \tilde{c}_0 \) to be the smallest one.

In Lemma 1 below, we show that if the alternative holds, i.e., \( \lambda_0 \neq 0 \), and \( c_0 \) is between the first and third “quantile” of \( F \), then \( \tilde{c}_0 \) is consistent for \( c_0 \). A more general discussion, as well as the proofs for Lemma 1 and Theorem 4 below, can be found in Section S3 of the SMs.

**Lemma 1.** Assume the alternative holds, i.e., \( \lambda_0 \neq 0 \). Suppose that Assumptions (A.1) and (A.2) hold for some \( r \in (4, \infty) \), and that \( F \) is differentiable at \( c_0 \) with \( F'(c_0) > 0 \). If \( F(c_0) \in [1/4, 3/4] \), \( \tilde{c}_0 \) converges to \( c_0 \) in probability as \( n \to \infty \).

Next, we establish the power consistency of the proposed test under the local alternatives:

\[
H_{1,n} : \lambda_0 = \lambda_{0,n}, \quad \text{with} \quad \lim_{n \to \infty} \inf n^{1/2 - \eta_0} |\lambda_{0,n}| > 0, \quad \lim_{n \to \infty} \sup |\lambda_{0,n}| < \infty , \quad (4.15)
\]
where $\eta_0$ appears in (A.2). For simplicity, assume $\alpha_0, \beta_0, \gamma_0$ do not vary with $n$. Thus the local alternatives approach the null at the rate of $n^{-1/2+\eta_0}$.

**Theorem 4.** Suppose that Assumptions (A.1) and (A.2) hold for some $r \in (4, \infty]$, and that the local alternatives in (4.15) hold. For any significance level $\xi \in (0, 1)$, we have $\lim_{n \to \infty} \text{pr}(p_{n,mrb}^* \leq \xi) = 1$.

Theorem 4 shows that the proposed test is consistent under the local alternatives in (4.15), despite the fact that the test statistic $M_n$ may not be bounded in probability under the null. For the examples in Subsection 4.3, Assumption (A.2) holds for any $\eta_0 \in (0, 1/2 - 1/r)$. Thus, if $r = \infty$, the local alternatives are allowed to approach the null $H_0 : \lambda_0 = 0$ at a faster rate than $n^{-q}$ as $n \to \infty$ for any $q < 1/2$.

### 4.3 Discussions on the assumptions

In this subsection we discuss examples for which Assumptions (A.2)(A.3) are satisfied for any $r \in (4, \infty]$.

**Example 1** (Almost All Realizations from an i.i.d. Sequence). Assume $(X_i, U_i) \ (i = 1, 2, \ldots)$ are independently and identically generated from some distribution. Once generated, they are fixed; therefore, the design is considered as fixed. Denote by $F_0$ the distribution function of $X_1$. In Lemma S2.6 in the SMs, we show that if $X_1$ and $U_1$ are independent, $0 <$
$E(U_1) < 1$, $0 < F_0(\ell) < F_0(u) < 1$, and $F_0$ is differentiable at $c_0$ with $F'(c_0) > 0$, Assumptions [(A.2)]-(A.3) hold almost surely with $F = F_0$, $p = E(U_1)$, for any $\eta_0 \in (0, 1/2 - 1/r)$ and $\eta_1 \in (1/2, 1 - 2/r)$.

**Example 2** (Regular Design). Let $F_0$ be a distribution function such that $0 < F_0(\ell) < F_0(u) < 1$ and $F_0$ is differentiable at $c_0$ with $F'(c_0) > 0$. Denote by $F_0^{-1}$ its quantile function, i.e., $F_0^{-1}(q) = \inf\{x : F(x) \geq q\}$. Further, let $\Pi = (\Pi_0, \ldots, \Pi_{L-1}) \in \{0, 1\}^L$ be a deterministic binary vector of length $L \geq 2$ such that $\sum_{k=0}^{L-1} \Pi_k \in (0, L)$. Denote by mod($i, L$) the remainder of dividing $i$ by $L$. If

$$X_i = F_0^{-1}(i/n), \quad U_i = \Pi_{\text{mod}(i,L)}, \quad i \in [n],$$

then Assumptions [(A.2)]-(A.3) hold with $F = F_0$, $p = L^{-1} \sum_{k=0}^{L-1} \Pi_k$, for any $\eta_0 \in (0, 1/2 - 1/r)$ and $\eta_1 \in (1/2, 1 - 2/r)$.

**Example 3** (Combinations). Assumptions [(A.2)]-(A.3) hold almost surely for any $\eta_0 \in (0, 1/2 - 1/r)$ and $\eta_1 \in (1/2, 1 - 2/r)$, if $X_i, i \in [n]$ (resp. $\{U_i\}$) is generated as a realization of an independent and identically distributed sequence and $U_i, i \in [n]$ (resp. $\{X_i\}$) is generated according to the deterministic binary pattern (resp. $F_0^{-1}$) as above.
4.4 Profile least squares estimation based test

Given the profile least squares estimator $\hat{c}_0$ in (4.11), an alternative approach to test $H_0 : \lambda_0 = 0$ would be using the $p$-value associated with the Wald statistic at $\hat{c}_0$, i.e.,

$$p_{n, pf} = 2 \left\{ 1 - \Phi(|M_{n, \hat{c}_0}|) \right\}.$$  \hspace{1cm} (4.16)

Lemma 2 below establishes the asymptotic size validity of the profile least squares estimation based test, $p_{n, pf}$, under the assumption that $c_0$ is identified in (2.4). The proof of the following lemma can be found in Section S4 of the SMs.

**Lemma 2.** Suppose that Assumptions (A.1), (A.2), and (A.3) hold for some $r \in (4, \infty)$, and that $\gamma_0 \neq 0$. Under the null $H_0 : \lambda_0 = 0$, $M_{n, \hat{c}_0}$ converges in distribution to the standard normal distribution as $n \to \infty$.

Despite its simplicity, the size validity of $p_{n, pf}$ requires that $\gamma_0 \neq 0$, as $c_0$ is not identified when $\gamma_0 = \lambda_0 = 0$. Further, even if $\gamma_0 \neq 0$, the Type I error of $p_{n, pf}$ is poorly controlled with a moderate sample size (see Table 1 in Section 5) if the effect of $I(X_i \leq c_0)$ is small. In comparison, the validity of the $p$-value, $p_{n, mr b}^*$ in (4.13), based on the proposed multiplier residual bootstrap, does not require $c_0$ to be identified. To wit, when $\gamma_0 = \lambda_0 = 0$, the estimator $\hat{\gamma}_0$ and $\hat{\sigma}^2$ in (4.12) are still consistent for $\gamma_0 = 0$ and $\sigma^2$, and
thus regardless of the value of \( \hat{c}_0 \), the impact of \( \hat{\gamma}_0 X_i, \hat{c}_0 \) to \( Y_i^* \) in (4.12) is asymptotically negligible.

**Remark 5.** The size validity of \( p_{n,pf} \) is due to the property that \( n^{\eta_1} |\hat{c}_0 - c_0| \) converges to zero almost surely (\( \eta_1 \) appears in Assumption (A.3)), which is not enjoyed by \( \tilde{c}_0 \) in (4.14). As a result, the \( p \)-value based on the minimum \( p \)-value estimator \( \tilde{c}_0, p_{n,mp} \) in (1.3), is not valid. Further, as discussed in Subsection 2.2 since the term \( I_{n,c} \) in (2.7) is a sequence of deterministic numbers which may diverge, \( p_{n,mp} \) may be arbitrarily small under the null.

5. **Simulation Studies for the Fixed Design**

In this section, simulation studies are conducted to evaluate the performance of the proposed multiplier residual bootstrap with the \( p \)-value \( p_{n,mrb}^* \) in (4.13), and compare it with the competing tests including the unadjusted minimum \( p \)-value test with the \( p \)-value \( p_{n,mp} \) in (1.3), and the test based on the profile least squares estimation with the \( p \)-value \( p_{n,pf} \) in (4.16) in terms of empirical sizes and powers, under the fixed design setup. In the following tables, “MRB”, “PF” and “MP” stand for the tests based on \( p_{n,mrb}^* \), \( p_{n,pf} \) and \( p_{n,mp} \), respectively; the bootstrap repetition for \( p_{n,mrb}^* \) is \( B = 2000 \). The results for the random design are qualitatively similar and presented in Section S6.1 of the SMs.
We generate $X_1, \ldots, X_n$ independently from the uniform distribution on $(0, 1)$, and $U_1, \ldots, U_n$ from the Bernoulli distribution with a success probability 0.5. Under the fixed design, once one realization is generated, it is *shared in all repetitions*. For each repetition, the responses $Y_1, \ldots, Y_n$ are generated using (2.4), where $\epsilon_1, \ldots, \epsilon_n$ are generated independently from some distribution $F_{\epsilon}$. We vary the sample size $n$, parameters $\theta_0 = (\alpha_0, \beta_0, \gamma_0, \lambda_0)^T$ and $c_0$, and consider the following noise distributions: $F_{\epsilon}(1) = N(0, 2^2)$, $F_{\epsilon}(2) = 2^{1/2} t(4)$, and $F_{\epsilon}(3) = 0.5 \times N(0.5, 1^2) + 0.5 \times N(-0.5, 2.55^2)$, where $N(a, b)$ denotes the normal distribution with mean $a$ and variance $b$, and $t(4)$ the $t$-distribution with 4 degrees of freedom. The empirical sizes and powers of the three tests, defined as the proportion of rejections under respectively $H_0$ and $H_1$, are calculated with 2000 repetitions at the level 5%. Table 1 is for the identifiable cases, while Table 2 for the non-identifiable cases. Note that in Section S6.2 in the SMs, we consider additional choices for $F_{\epsilon}$, which have heavy tails as the $t$-distribution.

Table 1 presents the empirical sizes and powers of the three tests under the *identifiable* cases. It is clear from the table that the empirical sizes of the tests based on the bootstrap adjustment, $p_{n,mb}^*$, are close to the nominal 5% level, while there is an almost 7 times inflation if the unadjusted version, $p_{n,mp}$, is used. Further, if the effect of $I(X_i \leq c_0)$ (i.e. $\gamma_0$ in (2.4)) is small,
the profile least squares estimation based tests, $p_{n, pf}$, controls poorly the Type I error (cf. $\theta^{(2)}$ in Table 1), whereas it does not affect the approach based on $p_{n, mrb}^*$. In addition, the tests based on the bootstrap adjustment, $p_{n, mrb}^*$, perform reasonably well even when the noise distribution $F_\epsilon$ has a heavy tail or is non-symmetric such as the $t$-distribution $F_\epsilon^{(2)}$ and the mixture distribution $F_\epsilon^{(3)}$.

As expected, the empirical powers of the tests based on the bootstrap adjustments, $p_{n, mrb}^*$, are not as large as those of the minimum $p$-value approach, $p_{n, mp}$, and of the profile least squares estimation approach, $p_{n, pf}$, both of which, however, fail to control the Type I error properly. When the sample size is moderate (say $\sim 300$), the gap is mild.

Table 2 presents the empirical sizes and powers of the three tests under the non-identifiable cases. From the table, the tests based on $p_{n, pf}$ and $p_{n, mp}$ lose control of the empirical sizes whereas the test based on the bootstrap adjustments $p_{n, mrb}^*$ behaves satisfactorily as in the identifiable cases. The empirical powers are close for the three tests.

6. **Application to a Colorectal Cancer Dataset**

In this section we apply the multiplier residual bootstrap test to the data from the CO.17 trial mentioned in the Introduction section, which random-
Table 1: The empirical sizes and powers (in percentage) for testing $H_0: \lambda_0 = 0$ at the level 5% under the identifiable case. Here, $\theta^{(1)} = (0, 1, 3, 0)^T$, $\theta^{(2)} = (2, 1.5, 1, 0)^T$ are for sizes, and $\theta^{(3)} = (0, 1, 3, 2)^T$, $\theta^{(4)} = (2, 1.5, 1, 2)^T$ for powers.

<table>
<thead>
<tr>
<th></th>
<th>$F_{\epsilon}^{(1)}$</th>
<th></th>
<th>$F_{\epsilon}^{(2)}$</th>
<th></th>
<th>$F_{\epsilon}^{(3)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n\theta_{\epsilon}$</td>
<td>$c_0$</td>
<td>MRB</td>
<td>PF</td>
<td>MP</td>
<td>MRB</td>
</tr>
<tr>
<td>100</td>
<td>$\theta^{(1)}$</td>
<td>0.3</td>
<td>6.4</td>
<td>4.7</td>
<td>34.8</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>7.1</td>
<td>6.1</td>
<td>31.3</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>$\theta^{(2)}$</td>
<td>0.3</td>
<td>5.2</td>
<td>21.2</td>
<td>30.9</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>6.7</td>
<td>16.6</td>
<td>33.9</td>
<td>6.7</td>
</tr>
<tr>
<td>300</td>
<td>$\theta^{(3)}$</td>
<td>0.3</td>
<td>5.1</td>
<td>4.5</td>
<td>34.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>5.4</td>
<td>4.7</td>
<td>34.6</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>$\theta^{(4)}$</td>
<td>0.3</td>
<td>5.3</td>
<td>10.0</td>
<td>37.0</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>5.2</td>
<td>10.4</td>
<td>31.9</td>
<td>4.3</td>
</tr>
<tr>
<td>100</td>
<td>$\theta^{(3)}$</td>
<td>0.3</td>
<td>39.4</td>
<td>66.3</td>
<td>71.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>58.7</td>
<td>69.7</td>
<td>85.5</td>
<td>64.1</td>
</tr>
<tr>
<td></td>
<td>$\theta^{(4)}$</td>
<td>0.3</td>
<td>53.2</td>
<td>66.7</td>
<td>80.9</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>59.6</td>
<td>69.1</td>
<td>89.9</td>
<td>63.1</td>
</tr>
<tr>
<td>300</td>
<td>$\theta^{(3)}$</td>
<td>0.3</td>
<td>94.3</td>
<td>96.2</td>
<td>99.0</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>97.8</td>
<td>98.6</td>
<td>99.5</td>
<td>96.4</td>
</tr>
<tr>
<td></td>
<td>$\theta^{(4)}$</td>
<td>0.3</td>
<td>93.2</td>
<td>95.8</td>
<td>98.9</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>97.1</td>
<td>99.1</td>
<td>99.8</td>
<td>97.6</td>
</tr>
</tbody>
</table>
Table 2: The empirical sizes and powers (in percentage) for testing $H_0 : \lambda_0 = 0$ at the level 5% under the non-identifiable case. Here, $\theta^{(5)} = (0, 1, 0, 0)^T$, $\theta^{(6)} = (2, 1.5, 0, 0)^T$ are for sizes, and $\theta^{(7)} = (0, 1, 0, 2)^T$, $\theta^{(8)} = (2, 1.5, 0, 2)^T$ for powers.

<table>
<thead>
<tr>
<th>$\theta_0$</th>
<th>$c_0$</th>
<th>$F_{\epsilon}^{(1)}$</th>
<th>$F_{\epsilon}^{(2)}$</th>
<th>$F_{\epsilon}^{(3)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MRB</td>
<td>PF</td>
<td>MP</td>
</tr>
<tr>
<td>300</td>
<td>$\theta^{(5)}$</td>
<td>0.3</td>
<td>5.7</td>
<td>31.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>5.7</td>
<td>31.1</td>
</tr>
<tr>
<td>$\theta^{(6)}$</td>
<td>0.3</td>
<td>5.5</td>
<td>31.8</td>
<td>37.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>6.1</td>
<td>29.6</td>
</tr>
<tr>
<td>$\theta^{(7)}$</td>
<td>0.3</td>
<td>93.6</td>
<td>96.6</td>
<td>99.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>95.8</td>
<td>99.1</td>
</tr>
<tr>
<td>$\theta^{(8)}$</td>
<td>0.3</td>
<td>95.1</td>
<td>96.7</td>
<td>99.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>97.1</td>
<td>98.8</td>
</tr>
</tbody>
</table>
ized 572 patients with advanced colorectal cancers to receive cetuximab plus BSC or BSC alone. Quality of life (QoL) is an important outcome in cancer clinical trials to assess the effect of a treatment on the palliation of symptoms and minimization of toxicity from perspectives of patients. In CO.17, QoL was assessed by using the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ)-C30. The prespecified primary objectives of QoL analysis were to compare between two treatment groups the change scores of the Physical Function Scale (PFS) and Global Health Status (GHS), two important subscales of EORTC QLQ-C30, from baseline at 8 and 16 weeks after the randomization. In our analyses, we are interested in identifying a subset of patients who may have a better QoL as measured by the change scores in PFS and GHS based on biomarkers other than the Kras gene studied previously. Besides the mRNA expression of the gene epiregulin (EREG) as studied in (Jonker et al., 2014) with respect to a survival outcome, the levels of lactate dehydrogenase (LDH) and alkaline phosphatase (ALKPH) in the blood are also considered as the candidates for a potential predictive biomarker. For the sake of space, we only present the results based on PFS at 16 weeks after the randomization.

We first apply the minimum $p$-value method to find the estimates of
the cutpoints for each of the candidate biomarkers, which are respectively 6.07 for EREG, 229 for LDH, and 108 for ALKPH. Table 3 presents the mean of the changes in PFS scores from baseline at 16 weeks after the randomization for the patients treated respectively by cetuximab plus BSC and BSC alone and also the difference between the two treatment groups and associated $p$-value from the Wilcoxon test in the two subgroups defined by the estimated cutpoints for each biomarker. From this table, one can see that patients with higher expression level of Epiregulin (smaller EREG value) and treated by cetuximab plus BSC had a highly significant better physical function (PF) at 0.05 level compared to those who were treated by BSC only, but PF was comparable between two treatment groups for patients with lower expression of Epiregulin (greater value of EREG). The $p$-value for the interaction between the treatment and EREG expression status from the minimum $p$-value method was 0.024, which would conclude that cetuximab should not be offered to patients with lower Epiregulin expression level. The $p$-value of the interaction test from the multiplier residual bootstrap method, as shown in Table 4, was 0.219, which suggests the differential treatment effects in the two EREG groups may be overstated by the unadjusted minimum $p$-value method and there is no sufficient evidence to support the conclusion that the treatment effects in terms of the change
in physical function are different between the patients with lower and higher EREG levels after the bootstrap adjustment is applied.

Similar conclusions can be drawn from the results of analyses for LDH. The test of interaction between the treatment and ALKPH level was not significant from either the unadjusted minimum \( p \)-values method or the multiplier residual bootstrap method but the \( p \)-value from the later method is more than four times of that from the former method.

Table 3: Subgroup analysis of EREG, LDH and ALKPH based on \( \tilde{c}_0 \) with respect to the change score in the physical function scale at 16 weeks.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Value</th>
<th>Cetuximab+BSC</th>
<th>BSC</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Value</td>
<td>n</td>
<td>Mean</td>
<td>Mean</td>
<td>Difference</td>
<td>p-value</td>
</tr>
<tr>
<td>EREG</td>
<td>( \leq 6.07 )</td>
<td>47</td>
<td>-0.35</td>
<td>28</td>
<td>-15.53</td>
<td>15.18</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>( &gt; 6.07 )</td>
<td>39</td>
<td>-9.74</td>
<td>24</td>
<td>-10.06</td>
<td>0.32</td>
<td>0.947</td>
</tr>
<tr>
<td>LDH</td>
<td>( \leq 229 )</td>
<td>61</td>
<td>-3.55</td>
<td>40</td>
<td>-14.92</td>
<td>11.36</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>( &gt; 229 )</td>
<td>21</td>
<td>-11.74</td>
<td>9</td>
<td>-11.11</td>
<td>-0.63</td>
<td>0.941</td>
</tr>
<tr>
<td>ALKPH</td>
<td>( \leq 108 )</td>
<td>44</td>
<td>-4.69</td>
<td>28</td>
<td>-17.73</td>
<td>13.04</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>( &gt; 108 )</td>
<td>42</td>
<td>-4.21</td>
<td>22</td>
<td>-7.88</td>
<td>3.67</td>
<td>0.480</td>
</tr>
</tbody>
</table>
Table 4: The estimated cutpoint $\tilde{c}_0$ and the corresponding $p$-values from the multiplier residual bootstrap and the unadjusted minimum $p$-value method based on the change score in the physical function scale at 16 weeks.

<table>
<thead>
<tr>
<th></th>
<th>$\tilde{c}_0$</th>
<th>$p_{n,mp}$</th>
<th>$p^*_{n, mrb}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EREG</td>
<td>6.07</td>
<td>0.024</td>
<td>0.219</td>
</tr>
<tr>
<td>LDH</td>
<td>229</td>
<td>0.080</td>
<td>0.474</td>
</tr>
<tr>
<td>ALKPH</td>
<td>108</td>
<td>0.164</td>
<td>0.724</td>
</tr>
</tbody>
</table>

7. Conclusion

In this work, we consider the problem of testing the significance of the interaction term in a linear model with two binary covariates: a treatment variable $U$ and a group indicator $I(X \leq c_0)$, where $X$ is a continuous biomarker and $c_0$ is an unknown cutpoint. We propose bootstrap methods to obtain the valid $p$-value for the minimum $p$-value test statistic under both the random and fixed designs. The extension to a linear model with additional covariates is straightforward, as long as they are independent of $X$ and $U$. We choose to focus on the simple model in (2.4) for clarity of presentation, and also because it already demonstrates two salient features of the minimum $p$-value method and its adjustment for linear models. First,
without assuming whether the cutpoint is identifiable or not, the proposed
adjustment leads to a valid size control in both cases. Second, under the
fixed design, the test statistic is not bounded in probability under the null.

For future research, we will study the predictive classification problems
based on other types of clinical outcomes, such as binary and time-to-event
outcomes. Nonlinear models, such as generalized linear models for binary
outcomes and Cox proportional hazards models for time-to-event outcomes,
will be required to formalise the problems. From preliminary investigation,
for these nonlinear models, the maximally selected Wald test statistic di-
verges at rate $\sqrt{n}$ even under the random design. Thus it is not clear how
we could obtain its critical values or whether the associated test is power
consistent against a fixed alternative. One possible remedy is to use the
profile maximum likelihood estimator $\hat{c}_0$ for the cutpoint, and plug it into
the usual score test statistic. With appropriate bootstrap calibration, the
resulting test has a valid size control regardless of the identifiability of the
cutpoint. We leave the precise statements for future work.

**Supplementary Materials**

The proofs and more simulation results are presented in the SMs.
Acknowledgements

We would like to thank the associated editor and two reviewers for the careful review and constructive comments which led to an improved manuscript. This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). This research is also enabled in part by support provided by Compute Canada (www.computecanada.ca).

References


Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64(2), 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 74(1), 33–43.


Götte, H., M. Kirchner, and M. Kieser (2020). Adjustment for exploratory cut-off selection in


Department of Mathematics and Statistics, Queen’s University, Kingston, ON, K7L 3N6, Canada

E-mail: na.li@queensu.ca

E-mail: yanglei.song@queensu.ca

E-mail: devon.lin@queensu.ca

Departments of Public Health Sciences & Mathematics and Statistics and Canadian Cancer Trials Group, Queen’s University, Kingston, ON, K7L 3N6, Canada

E-mail: dtu@ctg.queensu.ca