Statistica Sinica Preprint No: SS-2021-0268	
Title	Bootstrap Adjustment to Minimum p-Value Method for Predictive Classification
Manuscript ID	SS-2021-0268
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0268
Complete List of Authors	Na Li, Yanglei Song, Devon Lin and Dongsheng Tu
Corresponding Author	Dongsheng Tu
E-mail	dtu@ctg.queensu.ca

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

Statistica Sinica

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 a different treatment effect than that of other patients. However, this method suffers from 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 noncentered random process, the mean function (i.e., bias) of which diverges as the sample size goes to infinity, even under the null hypothesis. The proposed bootstrap statistic matches the diverging bias asymptotically, and we apply the high-dimensional Gaussian approximation results to establish the asymptotic size validity and the power consistency under local alternatives. The proposed method is applied to a data set from a clinical trial on advanced colorectal cancer.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

p-value Method, Multiplier Residual Bootstrap, Non-centered Process.

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 or the expression level or status of a gene, which we refer to as a biomarker, to make clinical decisions. There are two types of problems. The first is to classify patients with respect to a clinical outcome of interest, regardless of the types of treatments, for risk stratification. The second is to distinguish patients by their degree of benefit or harm to a particular treatment for guidance on its adoption. In the clinical literature, the former is referred to as *prognostic classification*, and the latter is referred to as *predictive classification*. The biomarkers used for these classifications are called prognostic and predictive biomarkers, respectively (Ballman, 2015).

This study focuses on identifying and assessing *predictive biomarkers* based on data from clinical trials that test an experimental treatment against a standard control. For example, consider the CO.17 trial conducted by the Canadian Cancer Trials Group (Jonker et al., 2007), which is a randomized phase III trial that compares cetuximab plus best supportive care (BSC) with BSC alone in patients with metastatic epidermal growth

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

factor receptor-positive colorectal cancer. Investigators identified the K-ras gene as a predictive biomarker with respect to various clinical outcomes, such as the change in the global health status score from baselines at 8 and 16 weeks after randomization. They found that patients with wild-type K-ras benefited more from the cetuximab treatment than did patients with mutated K-ras (Karapetis et al., 2008). As a result, the cetuximab treatment is now restricted to patients with a tumor bearing wild-type K-ras.

We consider 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. Specifically, let Y be a continuous clinical outcome of interest, U be a binary treatment indicator, and X be a continuous biomarker. We assume

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

where c_0 is an unknown cutpoint in the range $[\ell, u]$, $I(\cdot)$ is the indicator function, and $\alpha_0, \beta_0, \gamma_0$, and λ_0 are unknown parameters. Then, $\beta_0 + \lambda_0$ and β_0 represent the treatment effect, that is, the difference in the expected outcome between the experiment and the standard treatment groups, in the two subsets defined by the cutpoint c_0. Thus, λ_0 measures the *differential* treatment effect between these two subsets. An important task

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

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 γ_0 is nonzero. 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 is known to be c, for example, based on subject knowledge or from previous studies, we can use classical test statistics such as the Wald test statistic $M_{n,c}$ based on a sample of size n to test H_0. However, this is unrealistic for many clinical applications. In practice, the *minimum p-value method* is often used to assess both prognostic and predictive biomarkers when the cutpoint is unknown. The basic idea of this method in the context of predictive classification can be described as follows. If c is the true value of c_0, the associated Wald statistic $M_{n,c}$ has a limiting standard normal distribution $N(0,1)$. Thus, $2\{1-\Phi(|M_{n,c}|)\}$ is 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 $\tilde{c}_0$ as a value that achieves the minimum p-value or, equivalently, the maximal absolute Wald statistic,

$$\tilde{c}_0 \;=\; \underset{c\in[\ell,u]}{\operatorname{argmin}}\, 2\{1-\Phi(|M_{n,c}|)\} \;=\; \underset{c\in[\ell,u]}{\operatorname{argmax}}\, |M_{n,c}|. \tag{1.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

Then, the p-value from the minimum p-value method is

$$p_{n,mp} = 2\{1 - \Phi(M_n)\}, \quad \text{where} \quad M_n = |M_{n,\tilde{c}_0}| = \sup_{c\in[\ell,u]} |M_{n,c}|. \tag{1.3}$$

Although this method is simple and appealing to practitioners, its type-I error is substantially inflated, because the definition in (1.3) does not take into account that the cutpoint is estimated from the data. Thus, resulting analyses are not, in general, recognized in the 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 to adjust the p-value calculated from this method are urgently needed.

The main contribution of this study is to propose bootstrap methods to adjust the p-value defined in (1.3) under both *random designs* and *fixed designs* (see, e.g., Freedman (1981) for a discussion), and to establish the asymptotic size validity and power consistency. Note that the critical values obtained using the proposed bootstrap methods lead to proper type-I error control under *both* identifiable ($\gamma_0 \neq 0$) and non-identifiable cases ($\gamma_0 = 0$). These two cases are both practically important, and usually there is no convincing reason to assume one over the other. The literature (reviewed below) provides valid tests for each separate case, but not for both.

Specifically, under the random design, both the biomarker X and the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

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 $H_0 : \lambda_0 = 0$. Using standard empirical process 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) is *deterministic*, is nonstandard 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. Furthermore, although Chernozhukov, Chetverikov, and Kato (2014, 2016) have developed a Gaussian approximation and bootstrap tools for noncentered, non-Donsker empirical processes, $\{M_{n,c} : c \in [\ell, u]\}$ is not an empirical process. To address the problem 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 observe that M_n in (1.3) can be viewed as the supremum of the absolute

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

values of $\{M_{n,c} : c \in \mathcal{C}_n\}$, with $\mathcal{C}_n = \{X_1, \ldots, X_n\} \cap [\ell, u]$, a *noncentered* random vector, the dimension of which grows with n. Then, 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 $\{M^*_{n,c} : c \in \mathcal{C}_n\}$. Furthermore, because 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 body of literature on the adjustment of the minimum p-value method; see, for example, Miller and Siegmund (1982), Jespersen (1986), and Lausen and Schumacher (1992) and a comprehensive review by Mazumdar and Glassman (2000). In addition, Fan et al. (2017) recently considered 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 they assume that $\gamma_0 = 0$. In both the literature related to prognostic classification and in the work of Fan et al. (2017), the setup is nonstandard in the sense that c_0 is not identifiable under the null (Davies, 1977, 1987; Andrews,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

2001). Nonetheless, under the random design, the test statistics converge in distribution under their respective nulls. In this study, we do not assume the model is identifiable; in particular, we allow c_0 to be nonidentifiable. 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, 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.

There is also a large body of literature on the change-point or cutpoint 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 (4.1). If $\gamma_0 \neq 0$, under the random design, Koul et al. (2003) show 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 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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

the failure of conventional bootstrap methods does not contradict our work, because we study a *distinct* problem for a similar model; see Section S5 in the Supplementary Material. 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 and the minimum p-value method. In Sections 3 and 4, we propose paired and multiplier residual bootstrap methods for random and fixed designs, respectively. We present our simulation results in Section 5 to corroborate our theory, and an analysis of an advanced colorectal cancer data set in Section 6. Section 7 concludes the paper.

2. Problem Formulation

Let (Y_i, U_i, X_i), for $i \in [n] := \{1, \ldots, n\}$, 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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

on the covariates X_i and U_i:

$$Y_i = \alpha_0 + \beta_0 U_i + \gamma_0 X_{i,c_0} + \lambda_0 X_{i,c_0} U_i + \epsilon_i \quad \text{for } i \in [n], \tag{2.1}$$

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 $\epsilon_1, \ldots, \epsilon_n$ denote observation noise, which are assumed to be independent and identically distributed (i.i.d.) with mean zero and unknown variance σ^2. If $\gamma_0 = \lambda_0 = 0$, c_0 in (2.1) is not identifiable, in which case, we assume $c_0 = -\infty$, without loss of generality. If either $\gamma_0 \neq 0$ or $\lambda_0 \neq 0$, we assume that c_0 is a priori known to be in an interval $[\ell, u]$. Note that we make no assumption on 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 $\mathrm{RSS}_c(\theta)$ the associated residual sum of squares, that is,

$$\mathrm{RSS}_c(\theta) = \sum_{i=1}^{n} (Y_i - Z_{i,c}^T \theta)^2, \quad \text{where} \quad Z_{i,c} = (1, U_i, X_{i,c}, X_{i,c} U_i)^T.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

Then, for each $c \in \mathbb{R}$, the least squares estimator $\hat{\lambda}_c$ of λ_0 and its estimated variance $\hat{v}_c^2/n$ can be written as

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

$$\text{where} \quad \hat{\theta}_c = n^{-1}\hat{Q}_{n,c}^{-1}\sum_{i=1}^{n} Z_{i,c}Y_i, \quad \hat{Q}_{n,c} = n^{-1}\sum_{i=1}^{n} Z_{i,c}Z_{i,c}^T, \tag{2.2}$$

and $d = (0,0,0,1)^T$. Here, $\hat{\theta}_c$ is the least squares estimator for θ_0 for a fixed c, in the sense of minimizing $\mathrm{RSS}_c(\theta)$.

Recall that the task of a predictive classification is to test the null hypothesis $H_0 : \lambda_0 = 0$. If a particular c is treated as the true value for c_0, the Wald test statistic for $H_0 : \lambda_0 = 0$ is $M_{n,c}$ in (2.3). 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:

$$M_n = |M_{n,\tilde{c}_0}| = \sup_{c\in[\ell,u]} |M_{n,c}|, \quad \text{where} \quad M_{n,c} = \sqrt{n}\hat{\lambda}_c/\hat{v}_c. \tag{2.3}$$

Remark 1. In computing the test statistic M_n in (2.3), it suffices to take the supremum over the distinct values of the sample $X_1, \ldots, X_n$, that is, those $c \in \mathcal{C}_n = \{X_1, \ldots, X_n\} \cap [\ell, u]$.

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,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

that is, $\text{VAR}(\epsilon_1|X_1, U_1) = \sigma^2$. Nonetheless, our proposed procedure in Section 3 applies 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 $c \in [\ell, u]$:

$$n^{1/2}\hat{\lambda}_c = n^{-1/2}\widetilde{d}_c^T \sum_{i=1}^{n} Z_{i,c} Z_{i,c_0}^T \theta_0 + n^{-1/2}\widetilde{d}_c^T \sum_{i=1}^{n} Z_{i,c}\epsilon_i = I_{n,c} + II_{n,c}, \quad (2.4)$$

where $\widetilde{d}_c = \hat{Q}_{n,c}^{-1} d$, and $\hat{Q}_{n,c}$ and d are defined in (2.2). Under both the random and fixed designs, the second term is centered, that is, $E(II_{n,c}) = 0$, for $c \in [\ell, u]$. As discussed in the Introduction, the main challenge lies in analyzing the fixed design.

Specifically, under the fixed design, (U_i, X_i), for $i \in [n]$, are *deterministic* vectors. Thus, $I_{n,c}$ is a deterministic sequence that may diverge as $n \to \infty$. In fact, if (U_i, X_i), for $i \in [n]$, is one realization of an i.i.d. 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 and fixed designs in Sections 3 and 4, respectively. Note that by a conditional argument, a test that has a valid size under the fixed design

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

setup also does so under the random design if ϵ_1 is independent of 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), for $i \in [n]$, in (2.1) are i.i.d. 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 ℓ_∞-norm. Theorem 1 states that $\{M_{n,c} : c \in [\ell, u]\}$, appearing in the maximally selected Wald test statistic (2.3), converges weakly in $\ell^\infty([\ell, u])$ to a tight, centered Gaussian process. A definition of functional weak convergence and proofs for Theorems 1 and 2 can be found in Section S1 in the Supplementary Material.

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, $\mathbf{G} = \{G_c : c \in [\ell, u]\}$, such that $\{M_{n,c} : c \in [\ell, u]\}$ converges weakly in $\ell^\infty([\ell, u])$ to $\mathbf{G}$.*

Because the Gaussian process $\mathbf{G}$ has a complicated covariance structure, we propose the following paired bootstrap method to obtain the asymptotically valid p-value for the test statistic M_n in (2.3).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

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 (2.2), with $(Z_{i,c}, Y_i)$ replaced by $(Z_{i,c}^*, Y_i^*)$, where $Z_{i,c}^* = (1, U_i^*, X_{i,c}^*, U_i^* X_{i,c}^*)$ and $X_{i,c}^* = I(X_i^* \leq c)$. Furthermore, define the following bootstrap version of the maximally selected Wald test statistic, M_n^*:

$$M_n^* = \sup_{c \in [\ell, u]} |M_{n,c}^* - M_{n,c}|, \quad \text{where } M_{n,c}^* = \sqrt{n}\hat{\lambda}_c^*/\hat{v}_c, \quad \text{for } c \in [\ell, u]. \tag{3.1}$$

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

$$p_{n,pb}^* = 1 - F_{n,pb}^*(M_n). \tag{3.2}$$

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 (2.2), with $(Z_{i,c}, Y_i)$ replaced by $\left(Z_{i,c}^*, Y_i^*\right)$, is often used in the denominator to standardize $\hat{\lambda}_c^*$ in (3.1). Here, we propose using $\hat{v}_c$ mainly to simplify the proof.

The following theorem establishes the asymptotic validity of the adjusted p-value in (3.2) obtained using the proposed paired bootstrap method

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

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* $\mathbf{G}$ *as in Theorem 1. Consequently, for any significance level $\xi \in (0, 1)$,*

$$\lim_{n\to\infty} \mathrm{pr}\left(p^*_{n,pb} \le \xi\right) = \xi. \tag{3.3}$$

4. Multiplier Residual Bootstrap for the Fixed Design

In this section, we consider the fixed design setup, where (U_i, X_i), for $i \in [n]$, are deterministic and the randomness comes only from the observation noise ϵ_i, for $i \in [n]$. As a result, the first term $I_{n,c}$ in (2.4) 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 (2.3). Recall the definitions of $\mathrm{RSS}_c(\theta)$, $\hat\theta_c$, and $\hat v_c^2/n$ in Section 2.1.

Step 1. Define the profile least squares estimator $\hat c_0$ for c_0 and $\hat\sigma^2$ for σ^2 as follows:

$$\hat c_0 = \operatorname*{argmin}_{c\in[\ell,u]} \mathrm{RSS}_c(\hat\theta_c), \qquad \hat\sigma^2 = \mathrm{RSS}_{\hat c_0}(\hat\theta_{\hat c_0})/(n-4). \tag{4.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

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

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

$$Y_i^* = \hat{\alpha}_0 + \hat{\beta}_0 U_i + \hat{\gamma}_0 X_{i,\hat{c}_0} + \hat{\sigma}\zeta_i, \tag{4.2}$$

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

STEP 3. For a fixed c, define the least squares estimator, $\hat{\lambda}_c^*$, for the bootstrap sample (Z_i, Y_i^*), for $i \in [n]$, using (2.2), with Y_i replaced by Y_i^*. Furthermore, define the bootstrap test statistic, M_n^*, as follows:

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

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

$$p_{n,mrb}^* = 1 - F_{n,mrb}^*(M_n). \tag{4.3}$$

The conditional (on Y_i) distribution, $F_{n,mrb}^*$, of M_n^* can be estimated using the bootstrap, that is, by repeatedly generating independent realizations of the multipliers ζ_i, for $i \in [n]$. In computing the profile least squares estimator $\hat{c}_0$ for c_0, as in Remark 1, it suffices to consider $c \in \mathcal{C}_n$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

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), for $i \in [n]$, are deterministic and may depend on n. That is, we consider the triangle array setup, where $X_i = X_{n,i}$ and $U_i = U_{n,i}$, for $i \in [n]$, but for notational simplicity, we omit the dependence on n. However, the distributions of ϵ_i, for $i \in [n]$, and (c_0, θ_0) in (2.1) 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 $r \in (4, \infty]$, 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 $E\left(e^{t\epsilon_1}\right) \le \exp\left(\rho^2 t^2/2\right)$, for $t \in \mathbb{R}$. If $r < \infty$, assume that $E(|\epsilon_1|^r) < \infty$.

(A.2) There exist a nondecreasing function $F : [\ell, u] \to (0, 1)$ and constants $\eta_0 \in (0, 1/2 - 1/r)$ and $0 < p < 1$ such that, as $n \to \infty$,

$$n^{\frac{1}{2}-\eta_0}|p_n - p| + n^{\frac{1}{2}-\eta_0} \sup_{c\in[\ell,u]} \left[|s_{n,c} - F(c)| + |q_{n,c} - pF(c)|\right] \to 0.$$

If $\gamma_0 \neq 0$, c_0 is identified in the model (2.1), in which case, we further impose the following assumption. This is *not assumed* if $\gamma_0 = \lambda_0 = 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

(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 $n \to \infty$,

$$n^{\eta_1} \sup_{|c-c_0| \le K n^{-\eta_1}} |s_{n,c} - s_{n,c_0} - \{F(c) - F(c_0)\}| \to 0,$$

$$n^{\eta_1} \sup_{|c-c_0| \le K n^{-\eta_1}} |q_{n,c} - q_{n,c_0} - p\{F(c) - F(c_0)\}| \to 0.$$

Assumption (A.1) requires the noise ϵ_1 to have a sub-Gaussian tail if $r = \infty$, and a finite rth 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 rates, respectively, of $s_{n,c}$, $q_{n,c}$, and p_n to $F(c)$, $pF(c)$, and 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,*

$$\lim_{n\to\infty} n^q \sup_{\xi \in (0,1)} \left|\mathrm{pr}(p^*_{n,mrb} \le \xi) - \xi\right| = 0,$$

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 $r = \infty$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

Proof. Here, we outline the strategy. The complete proof is deferred to Section S2.1 in the Supplementary Material. Recall in (4.2) that $\hat{\alpha}_0, \hat{\beta}_0,$ and $\hat{\gamma}_0$ are the first three components of $\hat{\theta}_{\hat{c}_0}$ associated with $\hat{c}_0$ in (4.1); let $\hat{\theta}_0 = (\hat{\alpha}_0, \hat{\beta}_0, \hat{\gamma}_0, 0)^T$. Similarly to (2.4), for each $c \in [\ell, u]$,

$$\sqrt{n}\hat{\lambda}_c^* = n^{-1/2}\widetilde{d}_c^T \sum_{i=1}^n Z_{i,c} Z_{i,\hat{c}_0}^T \hat{\theta}_0 + n^{-1/2}\hat{\sigma}\widetilde{d}_c^T \sum_{i=1}^n Z_{i,c}\zeta_i = I_{n,c}^* + II_{n,c}^*,$$

where $\widetilde{d}_c = \hat{Q}_{n,c}^{-1} d$, and $\hat{Q}_{n,c}$ and d are defined in (2.2).

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

$$\sup_{\xi\in(0,1)} \left|\mathrm{pr}(p_{n,mrb}^* \le \xi) - \xi\right| = \sup_{\xi\in(0,1)} \left|\mathrm{pr}\{M_n \ge (F_{n,mrb}^*)^{-1}(1-\xi)\} - \xi\right|$$

$$= \sup_{\xi\in(0,1)} \left|\mathrm{pr}\left\{M_n \ge (F_{n,mrb}^*)^{-1}(1-\xi)\right\} - \mathrm{pr}_{|\epsilon}\left\{M_n^* \ge (F_{n,mrb}^*)^{-1}(1-\xi)\right\}\right|,$$

where $\mathrm{pr}_{|\epsilon}$ denotes the conditional probability given ϵ_i, for $i \in [n]$. By the triangle inequality, it is upper bounded by $\Upsilon_1 + \Upsilon_2 + \Upsilon_3$, where

$$\Upsilon_1 = \sup_{t>0}\left|\mathrm{pr}\left(\sup_{c\in\mathcal{C}_n}\left|\frac{I_{n,c}+II_{n,c}}{\hat{v}_c}\right| \le t\right) - \mathrm{pr}\left(\sup_{c\in\mathcal{C}_n}\left|\frac{I_{n,c}+(\sigma/\hat{\sigma})II_{n,c}^*}{\hat{v}_c}\right| \le t\right)\right|,$$

$$\Upsilon_2 = \sup_{t>0}\left|\mathrm{pr}\left(\sup_{c\in\mathcal{C}_n}\left|\frac{I_{n,c}+(\sigma/\hat{\sigma})II_{n,c}^*}{\hat{v}_c}\right| \le t\right) - \mathrm{pr}_{|\epsilon}\left(\sup_{c\in\mathcal{C}_n}\left|\frac{I_{n,c}+II_{n,c}^*}{\hat{v}_c}\right| \le t\right)\right|,$$

$$\Upsilon_3 = \sup_{t>0}\left|\mathrm{pr}_{|\epsilon}\left(\sup_{c\in\mathcal{C}_n}\left|\frac{I_{n,c}+II_{n,c}^*}{\hat{v}_c}\right| \le t\right) - \mathrm{pr}_{|\epsilon}\left(\sup_{c\in\mathcal{C}_n}\left|\frac{I_{n,c}^*+II_{n,c}^*}{\hat{v}_c}\right| \le t\right)\right|.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

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

First, $\Upsilon_1 \leq \sup_{A \in \mathcal{A}^{re}} |\text{pr}\,(II_n \in A) - \text{pr}\,((\sigma/\hat{\sigma})II_n^* \in A)|$, 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 that of 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 zero as $n \to \infty$.

Second, $\Upsilon_2 \leq \sup_{A \in \mathcal{A}^{re}} \left|\text{pr}_{|\epsilon}\,(II_n^* \in A) - \text{pr}\,((\sigma/\hat{\sigma})II_n^* \in A)\right|$. Conditional on ϵ_i, for $i \in [n]$, 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 $n \to \infty$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

The third term Υ_3 is clearly bounded by

$$\sup_{s_1,s_2\in\mathbb{R}^{n_c}} \{\mathrm{pr}_{|\epsilon}\left([II_n^*, -II_n^*] \le [s_1, s_2] + \Delta_n\right) - \mathrm{pr}_{|\epsilon}\left([II_n^*, -II_n^*] \le [s_1, s_2]\right)\},$$

where $\Delta_n = \sup_{c\in\mathcal{C}_n} \left|I_{n,c} - I_{n,c}^*\right|$, 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 Δ_n decays at a polynomial rate as $n \to \infty$. Finally, we apply 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 because of the availability of improved rates in the high-dimensional central limit theorem (Chernozhukov et al., 2019, Theorem 2.1) if the noise ϵ_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. Because the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

minimizer of the p-values of the Wald test statistics is not unique, we define

$$\tilde{c}_0 = \underset{c \in \mathcal{C}_n}{\operatorname{argmax}} \left| M_{n,c} \right|, \tag{4.4}$$

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, we show that if the alternative holds, that is, $\lambda_0 \neq 0$, and c_0 is between the first and third "quantiles" of F, then $\tilde{c}_0$ is consistent for c_0. A more general discussion and the proofs for Lemma 1 and Theorem 4 can be found in Section S3 of the Supplementary Material.

Lemma 1. *Assume the alternative holds, that is,* $\lambda_0 \neq 0$. *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:

$$H_{1,n} : \lambda_0 = \lambda_{0,n}, \text{ with } \liminf_{n\to\infty} n^{1/2-\eta_0} |\lambda_{0,n}| > 0, \ \limsup_{n\to\infty} |\lambda_{0,n}| < \infty, \tag{4.5}$$

where η_0 appears in (A.2). For simplicity, assume α_0, β_0, and γ_0 do not vary with n. Thus, the local alternatives approach the null at a rate of $n^{-1/2+\eta_0}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

Theorem 4. *Suppose that Assumptions* (A.1) *and* (A.2) *hold for some* $r \in (4, \infty]$*, and that the local alternatives in* (4.5) *hold. For any significance level* $\xi \in (0, 1)$*, we have* $\lim_{n\to\infty} \mathrm{pr}(p^*_{n,mrb} \leq \xi) = 1$.

Theorem 4 shows that the proposed test is consistent under the local alternatives in (4.5), 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 Discussion of 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 Supplementary Material, 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 $F_0'(c_0) > 0$, Assumptions (A.2)–(A.3) hold *almost surely*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

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 differentable at c_0, with $F_0'(c_0) > 0$. Denote by F_0^{-1} its quantile function, that is, $F_0^{-1}(q) = \inf\{x : F(x) \geq q\}$. Furthermore, 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 $\mathrm{mod}(i, L)$ the remainder of dividing i by L. If

$$X_i = F_0^{-1}\left(i/n\right), \qquad U_i = \Pi_{\mathrm{mod}(i,L)}, \qquad i \in [n],$$

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, for $i \in [n]$ (resp. $\{U_i\}$) is generated as a realization of an i.i.d. sequence and $U_i, i \in [n]$ (resp. $\{X_i\}$) is generated according to the deterministic binary pattern (resp. F_0^{-1}) described above.

4.4 Profile least squares estimation-based test

Given the profile least squares estimator $\hat{c}_0$ in (4.1), an alternative approach to test $H_0 : \lambda_0 = 0$ is to use the p-value associated with the Wald statistic

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

at $\hat{c}_0$, that is,

$$p_{n,pf} = 2\left\{1 - \Phi(|M_{n,\hat{c}_0}|)\right\}. \tag{4.6}$$

Lemma 2 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 (2.1). The proof of the lemma can be found in Section S4 of the Supplementary Material.

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 $\gamma_0 \neq 0$, because c_0 is not identified when $\gamma_0 = \lambda_0 = 0$. Furthermore, 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 $p^*_{n,mrb}$ in (4.3), based on the proposed multiplier residual bootstrap, does not require c_0 to be identified. As such, when $\gamma_0 = \lambda_0 = 0$, the estimator $\hat{\gamma}_0$ and $\hat{\sigma}^2$ in (4.2) are still consistent for $\gamma_0 = 0$ and σ^2, and thus regardless of the value of $\hat{c}_0$, the effect of $\hat{\gamma}_0 X_{i,\hat{c}_0}$ on Y^*_i in (4.2) is asymptotically negligible.

Remark 5. The size validity of $p_{n,pf}$ is because of the property that $n^{\eta_1}|\hat{c}_0 -$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

$c_0|$ converges to zero almost surely (η_1 appears in Assumption (A.3)), which is not enjoyed by $\tilde{c}_0$ in (4.4). 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. Furthermore, as discussed in Subsection 2.2, because the term $I_{n,c}$ in (2.4) is a sequence of deterministic numbers that may diverge, $p_{n,mp}$ may be arbitrarily small under the null.

5. Simulation Studies for the Fixed Design

In this section, we conduct simulation studies to evaluate the performance of the proposed multiplier residual bootstrap with the p-value $p^*_{n,mrb}$ in (4.3). We then compare this performance with that of 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 (4.6), in terms of their empirical size and power under the fixed design setup. In the following tables, "MRB", "PF", and "MP" represent 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 are presented in Section S6.1 of the Supplementary Material.

We generate $X_1, \ldots, X_n$ independently from a uniform distribution on $(0, 1)$, and $U_1, \ldots, U_n$ from a Bernoulli distribution with success probability 0.5. Under the fixed design, once one realization is generated, it is *shared in*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

all repetitions. For each repetition, the responses $Y_1, \ldots, Y_n$ are generated using (2.1), where $\epsilon_1, \ldots, \epsilon_n$ are generated independently from some distribution F_ϵ. 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 a normal distribution with mean a and variance b, and $t(4)$ is a t-distribution with four degrees of freedom. The empirical size and power of the three tests, defined as the proportion of rejections under H_0 and H_1, respectively, are calculated with 2000 repetitions at the 5% level. Table 1 is for the identifiable cases, and Table 2 shows the non-identifiable cases. Note that in Section S6.2 of the Supplementary Material, we consider additional choices for F_ϵ that have heavy tails as the t-distribution.

Table 1 presents the empirical size and power of the three tests under the *identifiable* cases. The table clearly shows that the empirical sizes of the tests based on the bootstrap adjustment, $p^*_{n,mrb}$, are close to the nominal 5% level, whereas there is an almost seven times inflation if we use the unadjusted version, $p_{n,mp}$. In addition, if the effect of $I(X_i \leq c_0)$ (i.e., γ_0 in (2.1)) is small, the profile least squares estimation-based tests, $p_{n,pf}$, control the type-I error (cf., $\theta^{(2)}$ in Table 1) poorly, whereas they do not affect the approach based on $p^*_{n,mrb}$. In addition, the tests based on the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

bootstrap adjustment, $p^*_{n,mrb}$, perform reasonably well even, when the noise distribution F_ϵ has a heavy tail or is nonsymmetric such as the t-distribution $F^{(2)}_\epsilon$ and the mixture distribution $F^{(3)}_\epsilon$.

As expected, the empirical power of the tests based on the bootstrap adjustments, $p^*_{n,mrb}$, is not as large as that of the minimum p-value approach, $p_{n,mp}$, or that of the profile least squares estimation approach, $p_{n,pf}$, both of which fail to control the type-I error properly. When the sample size is moderate (say ~ 300), the gap is mild.

Table 2 presents the empirical size and power 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 size, whereas the test based on the bootstrap adjustments $p^*_{n,mrb}$ behaves satisfactorily, as in the identifiable cases. The empirical power is close for the three tests.

6. Application to a Colorectal Cancer Data set

In this section, we apply the multiplier residual bootstrap test to data from the CO.17 trial mentioned in the Introduction, which randomized 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, used to assess the effect of a treatment on the palliation of symptoms

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

Table 1: The empirical size and power (in percentage) when testing H_0 : $\lambda_0 = 0$ at the 5% level under the identifiable case. Here, $\theta^{(1)} = (0, 1, 3, 0)^T$ and $\theta^{(2)} = (2, 1.5, 1, 0)^T$ denote the size, and $\theta^{(3)} = (0, 1, 3, 2)^T$ and $\theta^{(4)} = (2, 1.5, 1, 2)^T$ denote the power.

			$F_\epsilon^{(1)}$			$F_\epsilon^{(2)}$			$F_\epsilon^{(3)}$		
n	θ_0	c_0	MRB	PF	MP	MRB	PF	MP	MRB	PF	MP
100	$\theta^{(1)}$	0.3	6.4	4.7	34.8	7.1	6.4	30.6	7.1	7.1	32.4
		0.5	7.1	6.1	31.3	6.2	5.6	28.1	7.7	6.4	35.8
	$\theta^{(2)}$	0.3	5.2	21.2	30.9	6.5	19.3	36.3	6.0	17.5	30.7
		0.5	6.7	16.6	33.9	6.7	17.1	33.1	7.4	14.7	33.5
300	$\theta^{(1)}$	0.3	5.1	4.5	34.5	5.6	5.1	31.6	5.2	5.6	35.8
		0.5	5.4	4.7	34.6	5.9	5.5	29.9	4.9	5.6	25.3
	$\theta^{(2)}$	0.3	5.3	10.0	37.0	5.7	8.5	36.4	5.1	10.6	38.9
		0.5	5.2	10.4	31.9	4.3	8.2	37.3	5.2	9.1	37.2
100	$\theta^{(3)}$	0.3	39.4	66.3	71.5	43.1	62.2	68.9	66.7	64.2	82.9
		0.5	58.7	69.7	85.5	64.1	68.7	79.6	60.4	69.6	78.5
	$\theta^{(4)}$	0.3	53.2	66.7	80.9	59.1	62.9	82.1	51.1	59.4	79.3
		0.5	59.6	69.1	89.9	63.1	67.3	84.3	60.4	67.4	83.6
300	$\theta^{(3)}$	0.3	94.3	96.2	99.0	94.1	98.1	98.8	96.5	96.3	99.5
		0.5	97.8	98.6	99.5	96.4	98.9	99.7	97.9	99.0	99.8
	$\theta^{(4)}$	0.3	93.2	95.8	98.9	92.3	96.3	98.5	95.4	96.7	99.1
		0.5	97.1	99.1	99.8	97.6	98.7	99.6	97.2	98.8	99.7

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

Table 2: The empirical size and power (in percentage) when testing $H_0 : \lambda_0 = 0$ at the 5% level under the non-identifiable case. Here, $\theta^{(5)} = (0, 1, 0, 0)^T$ and $\theta^{(6)} = (2, 1.5, 0, 0)^T$ denote the size, and $\theta^{(7)} = (0, 1, 0, 2)^T$ and $\theta^{(8)} = (2, 1.5, 0, 2)^T$ denote the power.

			$F_\epsilon^{(1)}$			$F_\epsilon^{(2)}$			$F_\epsilon^{(3)}$		
n	θ_0	c_0	MRB	PF	MP	MRB	PF	MP	MRB	PF	MP
300	$\theta^{(5)}$	0.3	5.7	31.8	37.5	5.5	29.6	39.7	5.6	32.1	36.4
		0.5	5.7	31.1	38.0	5.4	29.4	39.2	5.4	29.4	38.1
	$\theta^{(6)}$	0.3	5.5	31.8	37.7	5.8	31.2	37.6	6.0	29.9	39.0
		0.5	6.1	29.6	39.3	5.2	30.3	37.4	6.1	29.1	37.7
	$\theta^{(7)}$	0.3	93.6	96.6	99.4	93.3	97.0	99.5	92.9	97.4	99.5
		0.5	95.8	99.1	99.6	97.3	99.0	99.7	97.0	98.8	99.9
	$\theta^{(8)}$	0.3	95.1	96.7	99.5	94.2	97.1	99.1	93.9	94.9	99.5
		0.5	97.1	98.8	99.9	95.9	98.9	99.7	97.0	98.7	99.6

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

and the minimization of toxicity from the perspective of the patients. In CO.17, QoL is assessed using the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ)-C30. The prespecified primary objectives of the QoL analysis were to compare two treatment groups in terms of the change scores of the Physical Function Scale (PFS) and Global Health Status (GHS), two important subscales of EORTC QlQ-C30, from baselines at eight 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 the PFS and GHS, based on biomarkers other than the Kras gene studied previously. In addition to 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 potential predictive biomarkers. To conserve space, we present only the results based on the PFS at 16 weeks after the randomization.

We first apply the minimum p-value method to estimate the cutpoints for each of the candidate biomarkers, which are 6.07 for EREG, 229 for LDH, and 108 for ALKPH. Table 3 presents the mean of the changes in the PFS scores from the baseline at 16 weeks after the randomization for

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

the patients treated by cetuximab plus BSC and BSC alone, as well as the difference between the two treatment groups and the associated p-value from the Wilcoxon test in the two subgroups defined by the estimated cutpoints for each biomarker. The table shows that patients with a higher expression level of Epiregulin (smaller EREG value) and treated by cetuximab plus BSC had a highly significant better physical function (PF) at the 0.05 level compared with those who were treated by BSC only. However, the PF was comparable between the two treatment groups for patients with a lower expression of Epiregulin (greater value of EREG). The p-value for the interaction between the treatment and the EREG expression status from the minimum p-value method is 0.024, indicating that cetuximab should not be offered to patients with a lower Epiregulin expression level. The p-value of the interaction test from the multiplier residual bootstrap method, shown in Table 4, is 0.219, which suggests that the differential treatment effects in the two EREG groups may be overstated by the unadjusted minimum p-value method, and there is insufficient evidence to support the conclusion that the treatment effects in terms of the change in the PF are different between patients with lower and higher EREG levels after the bootstrap adjustment is applied.

Similar conclusions can be drawn from the results of the analyses for

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

LDH. The test of the interaction between the treatment and the ALKPH level is not statistically significant when using either the unadjusted minimum p-values method or the multiplier residual bootstrap method, but the p-value from the latter method is more than four times that of 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 PF scale at 16 weeks.

		Cetuximab+BSC		BSC		
Factor	Value	n	Mean	n	Mean	Difference	p-value
EREG	≤ 6.07	47	-0.35	28	-15.53	15.18	0.002
	> 6.07	39	-9.74	24	-10.06	0.32	0.947
LDH	≤ 229	61	-3.55	40	-14.92	11.36	0.003
	> 229	21	-11.74	9	-11.11	-0.63	0.941
ALKPH	≤ 108	44	-4.69	28	-17.73	13.04	0.007
	> 108	42	-4.21	22	-7.88	3.67	0.480

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

Table 4: The estimated cutpoint $\tilde{c}_0$ and corresponding p-values from the multiplier residual bootstrap and unadjusted minimum p-value method based on the change score in the PF scale at 16 weeks.

	$\tilde{c}_0$	$p_{n,mp}$	$p^*_{n,mrb}$
EREG	6.07	0.024	0.219
LDH	229	0.080	0.474
ALKPH	108	0.164	0.724

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 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 (2.1) for clarity of presentation, and also because it demonstrates two salient features of the minimum p-value method and its adjustment for linear models. First, without assuming the cutpoint is identifiable, 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.

In future research, we will study predictive classification problems based on other types of clinical outcomes, such as binary and time-to-event out-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

comes. Nonlinear models, such as generalized linear models for binary outcomes and Cox proportional hazards models for time-to-event outcomes, will be required to formalize the problems. From a preliminary investigation, for these nonlinear models, the maximally selected Wald test statistic diverges at a rate $\sqrt{n}$, even under the random design. Thus, it is not clear how to 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 an appropriate bootstrap calibration, the resulting test has a valid size control, regardless of the identifiability of the cutpoint.

Supplementary Material

The proofs and additional simulation results are presented in the online Supplementary Material.

Acknowledgments

We would like to thank the associate editor and two reviewers for their careful review and constructive comments. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

This research was also supported, in part, by Compute Canada (`www.computecanada.ca`).

References

Andrews, D. W. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica 69*(3), 683–734.

Ballman, K. V. (2015). Biomarker: Predictive or prognostic. *J. Clin. Oncol. 33*(33), 3968–3971.

Blok, E. J., C. C. Engels, G. Dekker-Ensink, E. M.-K. Kranenbarg, H. Putter, V. T. Smit, G.-J. Liefers, J. P. Morden, J. M. Bliss, R. C. Coombes, J. M. Bartlett, J. R. Kroep, C. J. van de Velde, and P. J. Kuppen (2018). Exploration of tumour-infiltrating lymphocytes as a predictive biomarker for adjuvant endocrine therapy in early breast cancer. *Breast Cancer Res. Treat. 171*(1), 65–74.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist. 41*(6), 2786–2819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist. 42*(4), 1564–1597.

Chernozhukov, V., D. Chetverikov, and K. Kato (2016). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings. *Stoch. Proces. Appl. 126*(12), 3632–3651.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

Chernozhukov, V., D. Chetverikov, and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Prob. 45*(4), 2309–2352.

Chernozhukov, V., D. Chetverikov, K. Kato, and Y. Koike (2019). Improved central limit theorem and bootstrap approximations in high dimensions. *arXiv*, 1912.10529.

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.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist. 7*(1), 1–26.

Fan, A., R. Song, and W. Lu (2017). Change-plane analysis for subgroup detection and sample size calculation. *J. Am. Statist. Assoc. 112*(518), 769–778.

Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist. 9*(6), 1218 – 1228.

Gavanji, P., B. E. Chen, and W. Jiang (2018). Residual bootstrap test for interactions in biomarker threshold models with survival data. *Stat. Biosci. 10*(1), 202–216.

Götte, H., M. Kirchner, and M. Kieser (2020). Adjustment for exploratory cut-off selection in randomized clinical trials with survival endpoint. *Biom. J. 62*(3), 627–642.

He, P. (2014). Identifying cut points for biomarker defined subset effects in clinical trials with survival endpoints. *Contemp. Clin. Trials 38*(2), 333–337.

Jespersen, N. (1986). Dichotomizing a continuous covariate in the cox regression model. Tech-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

nical report, Statistical Research Unit, University of Copenhagen.

Jiang, W., B. Freidlin, and R. Simon (2007). Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J. Natl. Cancer Inst. 99*(13), 1036–1043.

Jonker, D. J., C. S. Karapetis, C. Harbison, C. J. O'Callaghan, D. Tu, R. J. Simes, D. P. Malone, C. Langer, N. Tebbutt, T. J. Price, J. Shapiro, L. L. Siu, R. P. Wong, G. Bjarnason, M. J. Moore, J. R. Zalcberg, and S. Khambata-Ford (2014). Epiregulin gene expression as a biomarker of benefit from cetuximab in the treatment of advanced colorectal cancer. *Br. J. Cancer 110*(3), 648–655.

Jonker, D. J., C. J. O'Callaghan, C. S. Karapetis, J. R. Zalcberg, D. Tu, H.-J. Au, S. R. Berry, M. Krahn, T. Price, R. J. Simes, N. C. Tebbutt, G. van Hazel, R. Wierzbicki, C. Langer, and M. J. Moore (2007). Cetuximab for the treatment of colorectal cancer. *N. Engl. J. Med. 357*(20), 2040–2048.

Karapetis, C. S., S. Khambata-Ford, D. J. Jonker, C. J. O'Callaghan, D. Tu, N. C. Tebbutt, R. J. Simes, H. Chalchal, J. D. Shapiro, S. Robitaille, T. J. Price, L. Shepherd, H.-J. Au, C. Langer, M. J. Moore, and J. R. Zalcberg (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N. Engl. J. Med. 359*(17), 1757–1765.

Koul, H. L., L. Qian, and D. Surgailis (2003). Asymptotics of M-estimators in two-phase linear regression models. *Stoch. Proces. Appl. 103*(1), 123–154.

Lausen, B. and M. Schumacher (1992). Maximally selected rank statistics. *Biometrics 48*(1),

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

73–85.

Li, J. and B. Jin (2018). Multi-threshold accelerated failure time model. *Ann. Statist. 46*(6A), 2657–2682.

Mallik, A., B. Sen, M. Banerjee, and G. Michailidis (2011). Threshold estimation based on a p-value framework in dose-response and regression settings. *Biometrika 98*(4), 887–900.

Mazumdar, M. and J. R. Glassman (2000). Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statist. Med. 19*(1), 113–132.

Miller, R. and D. Siegmund (1982). Maximally selected chi square statistics. *Biometrics 38*(4), 1011–1016.

Mukherjee, D., M. Banerjee, and Y. Ritov (2020). Asymptotic normality of a linear threshold estimator in fixed dimension with near-optimal rate. *arXiv*, 2001.06955.

Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In M. V.D. and S. G. (Eds.), *Geometric Aspects of Functional Analysis: Israel Seminar 2001-2002*, Volume 1807, pp. 169–187. Springer Berlin Heidelberg.

Seijo, E. and B. Sen (2011). Change-point in stochastic design regression and the bootstrap. *Ann. Statist. 39*(3), 1580–1607.

Shao, J. and D. Tu (2012). *The Jackknife and Bootstrap*. Springer Science & Business Media.

Van Der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes: With*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0268

Applications to Statistics. Springer-Verlag New York.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist. 14*(4), 1261–1295.

Yu, P. (2014). The bootstrap in threshold regression. *Econometric Theory 30*(3), 676–714.

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

E-mail: na.li@queensu.ca

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

E-mail: yanglei.song@queensu.ca

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

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