Biostatistics Advance Access published January 30, 2012

Biostatistics (2012), **xx**, xx, *pp.* 1–18
doi:10.1093/biostatistics/kxr050

On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial

LU TIAN
Department of Health Research & Policy, Stanford University, Stanford, CA 94305, USA

TIANXI CAI
Department of Biostatistics, Harvard University, Boston, MA 02115, USA

LIHUI ZHAO
Department of Preventive Medicine, Northwestern University, Chicago, IL 60611, USA

LEE-JEN WEI*
Department of Biostatistics, Harvard University, Boston, MA 02115, USA
wei@hsph.harvard.edu

SUMMARY

To estimate an overall treatment difference with data from a randomized comparative clinical study, baseline covariates are often utilized to increase the estimation precision. Using the standard analysis of covariance technique for making inferences about such an average treatment difference may not be appropriate, especially when the fitted model is nonlinear. On the other hand, the novel augmentation procedure recently studied, for example, by Zhang *and others* (2008. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64**, 707–715) is quite flexible. However, in general, it is not clear how to select covariates for augmentation effectively. An overly adjusted estimator may inflate the variance and in some cases be biased. Furthermore, the results from the standard inference procedure by ignoring the sampling variation from the variable selection process may not be valid. In this paper, we first propose an estimation procedure, which augments the simple treatment contrast estimator directly with covariates. The new proposal is asymptotically equivalent to the aforementioned augmentation method. To select covariates, we utilize the standard lasso procedure. Furthermore, to make valid inference from the resulting lasso-type estimator, a cross validation method is used. The validity of the new proposal is justified theoretically and empirically. We illustrate the procedure extensively with a well-known primary biliary cirrhosis clinical trial data set.

Keywords: ANCOVA; Cross validation; Efficiency augmentation; Mayo PBC data; Semi-parametric efficiency.

*To whom correspondence should be addressed.

1. INTRODUCTION

For a typical randomized clinical trial to compare two treatments, generally a summary measure θ_0 for quantifying the treatment effectiveness difference can be estimated unbiasedly or consistently using its simple two-sample empirical counterpart, say $\hat{\theta}$. With the subject's baseline covariates, one may obtain a more efficient estimator for θ_0 via a standard analysis of covariance (ANCOVA) technique or a novel augmentation procedure, which is well documented in Zhang *and others* (2008) and a series of papers (Leon *and others*, 2003; Tsiatis, 2006; Tsiatis *and others*, 2008; Lu and Tsiatis, 2008; Gilbert *and others*, 2009; Zhang and Gilbert, 2010). The ANCOVA approach can be problematic, especially when the regression model is nonlinear, for example, the logistic or Cox model. For this case, the ANCOVA estimator generally does not converge to θ_0, but to a quantity which may be difficult to interpret as a treatment contrast measure. Moreover, in the presence of censored event time observations, this quantity may depend on the censoring distribution. On the other hand, the above augmentation procedure, referred as ZTD, in the literature always produces a consistent estimator for θ_0, provided that the simple estimator $\hat{\theta}$ is consistent.

In theory, the ZTD estimator, denoted by $\hat{\theta}_{\mathrm{ZTD}}$ hereafter, is asymptotically more efficient than $\hat{\theta}$ no matter how many covariates being augmented. In practice, however, an "overly augmented" or "mis-augmented" estimator may have a larger variance than that of $\hat{\theta}$ and in special case may even have undesirable finite sample bias. Recently, Zhang *and others* (2008) showed empirically that the ZTD via the standard stepwise regression for variable selection performs satisfactorily when the number of covariates is not large. In general, however, it is not clear that the standard inference procedures for θ_0 based on estimators augmented by covariates selected via a rather complex variable selection process is appropriate especially when the number of covariates involved is not small relative to the sample size. Therefore, it is highly desirable to develop an estimation procedure to properly and systematically augment $\hat{\theta}$ and make valid inference for the treatment difference using the data with practical sample sizes.

Now, let Y be the response variable, T be the binary treatment indicator, and $\mathbf{Z}$ be a p-dimensional vector of baseline covariates including 1 as its first element and possibly transformations of original variables. The data, $\{(Y_i, T_i, \mathbf{Z}_i), i = 1, \ldots, n\}$, consist of n independent copies of $(Y, T, \mathbf{Z})$, where T and $\mathbf{Z}$ are independent of each other. Let $P(T = 1) = \pi \in (0, 1)$. First, suppose that we are interested in the mean difference: $\theta_0 = E(Y|T = 1) - E(Y|T = 0)$. A simple unadjusted estimator is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{(T_i - \pi)Y_i}{\pi(1-\pi)},$$

which consistently estimates θ_0. To improve efficiency in estimating θ_0, one may employ the standard ANCOVA procedure by fitting the following linear regression "working" model:

$$E(Y|T, \mathbf{Z}) = \theta T + \gamma'\mathbf{Z},$$

where θ and γ are unknown parameters. Since $T \perp \mathbf{Z}$ and $\{(T_i, \mathbf{Z}_i), i = 1, \ldots, n\}$ are independent copies of $(T, \mathbf{Z})$, the resulting ANCOVA estimator is asymptotically equivalent to

$$\hat{\theta} - \hat{\gamma}'\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{(T_i - \pi)\mathbf{Z}_i}{\pi(1-\pi)}\right\}, \tag{1.1}$$

where $\hat{\gamma}$ is the ordinary least square estimator for γ of the model $E(Y|\mathbf{Z}) = \gamma'\mathbf{Z}$. As $n \to \infty$, $\hat{\gamma}$ converges to

$$\gamma_0 = \operatorname{argmin}_{\gamma} E(Y - \gamma'\mathbf{Z})^2.$$

It follows that the ANCOVA estimator is asymptotically equivalent to

$$\hat{\theta} - \gamma_0' \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(T_i - \pi)\mathbf{Z}_i}{\pi(1-\pi)} \right\}. \tag{1.2}$$

In theory, since $\hat{\theta}$ is consistent to θ_0, the ANCOVA estimator is also consistent to θ_0 and more efficient than $\hat{\theta}$ regardless of whether the above working model is correctly specified. Furthermore, as noted by Tsiatis *and others* (2008), the nonparametric ANCOVA estimator proposed by Koch *and others* (1998) and $\hat{\theta}_{\mathrm{ZTD}}$ are also asymptotically equivalent to (1.2) when $\pi = 0.5$. We give details of this equivalence in Appendix A.

The novel ZTD procedure is derived by specifying optimal estimating functions under a very general semi-parametric setting. The efficiency gain from $\hat{\theta}_{\mathrm{ZTD}}$ has been elegantly justified using the semi-parametric inference theory (Tsiatis, 2006). The ZTD is much more flexible than the ANCOVA method in that it can handle cases when the summary measure θ_0 is beyond the simple difference of two group means. On the other hand, the ANCOVA method may only work under above simple linear regression model.

In this paper, we study the estimator (1.1), which augments $\hat{\theta}$ directly with the covariates. The key question is how to choose $\hat{\gamma}$ in (1.1) especially when p is not small with respect to n. Here, we utilize the lasso procedure with a cross validation process to construct a systematic procedure for selecting covariates to increase the estimation precision. The validity of the new proposal is justified theoretically and empirically via an extensive simulation study. The proposal is also illustrated with the data from a clinical trial to evaluate a treatment for a specific liver disease.

2. Estimating the treatment difference via proper augmentation from covariates

For a general treatment contrast measure θ_0 and its simple two-sample estimator $\hat{\theta}$, assume that

$$\hat{\theta} - \theta_0 = n^{-1} \sum_{i=1}^{n} \tau_i(\eta) + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where $\tau_i(\eta)$ is the influence function from the ith observation, η is a vector of unknown parameters, and $i = 1, \ldots, n$. Note that the influence function generally only involves a rather small number of unknown parameters, which is not dependent on $\mathbf{Z}$. Let $\hat{\eta}$ denote the consistent estimator for η. Generally, the above asymptotic expansion is also valid with τ_i being replaced by $\tau_i(\hat{\eta})$. Now, (1.2) can be rewritten as

$$\hat{\theta} - \gamma_0' \left(n^{-1} \sum_{i=1}^{n} \xi_i \right),$$

where $\xi_i = (T_i - \pi)\mathbf{Z}_i / \{\pi(1-\pi)\}, i = 1, \ldots, n$. Then $\hat{\gamma}$ in (1.1) is the minimizer of

$$\sum_{i=1}^{n} \{\tau_i(\hat{\eta}) - \gamma'\xi_i\}^2. \tag{2.1}$$

When the dimension of $\mathbf{Z}$ is not small, to obtain a stable minimizer, one may consider the following regularized minimand:

$$L_\lambda(\gamma) = \sum_{i=1}^{n} \{\tau_i(\hat{\eta}) - \gamma'\xi_i\}^2 + \lambda|\gamma|,$$

where λ is the lasso tuning parameter (Tibshirani, 1996) and $|\cdot|$ denote the L_1 norm for a vector. For any fixed λ, let the resulting minimizer be denoted by $\hat{\gamma}(\lambda)$. The corresponding augmented estimator and its variance estimator are

$$\hat{\theta}_{\text{lasso}}(\lambda) = \hat{\theta} - \hat{\gamma}(\lambda)'\left(n^{-1}\sum_{i=1}^{n}\xi_i\right)$$

and

$$\hat{V}_{\text{lasso}}(\lambda) = n^{-2}\sum_{i=1}^{n}\{\tau_i(\hat{\eta}) - \hat{\gamma}(\lambda)'\xi_i\}^2, \tag{2.2}$$

respectively. Asymptotically, one may ignore the variability of $\hat{\gamma}(\lambda)$ and treat it as a constant when we make inferences about θ_0. However, in some cases, we have found empirically that similar to $\hat{\theta}_{\text{ZTD}}$, $\hat{\theta}_{\text{lasso}}(\lambda)$ is biased partly due to the fact that $\hat{\gamma}(\lambda)$ and $\{\xi_i, i = 1, \ldots, n\}$ are correlated. In the simulation study, we show via a simple example this undesirable finite-sample phenomenon. In practice, such bias may not have real impact on the conclusions about the treatment difference, θ_0, when the study sample size is relatively large with respect to the dimension of $\mathbf{Z}$.

One possible solution to reduce the correlation between $\hat{\gamma}(\lambda)$ and ξ_i is to use a cross validation procedure. Specifically, we randomly split the data into K nonoverlapping sets $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$ and construct an estimator for θ_0:

$$\hat{\theta}_{\text{cv}}(\lambda) = \hat{\theta} - \frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_{(-i)}(\lambda)'\xi_i,$$

where $i \in \mathcal{D}_{k_i}$, $\hat{\gamma}_{(-i)}(\lambda)$ is the minimizer of

$$\sum_{j \notin \mathcal{D}_{k_i}}\{\tau_j(\hat{\eta}_{(-i)}) - \gamma'\xi_j\}^2 + \lambda|\gamma|,$$

and $\hat{\eta}_{(-i)}$ is a consistent estimator for η with all data but not from $\mathcal{D}_{k_i}$. Note that $\hat{\gamma}_{(-i)}(\lambda)$ and ξ_i are independent and no extra bias would be added from $\hat{\theta}_{\text{cv}}(\lambda)$ to $\hat{\theta}$. When $n \gg p$, the variance of $\hat{\theta}_{\text{cv}}(\lambda)$ can be estimated by $\hat{V}_{\text{lasso}}(\lambda)$ given in (2.2). However $\hat{V}_{\text{lasso}}(\lambda)$ tends to underestimate its true variance when p is not small.

Here, we utilize the above cross validation procedure to construct a natural variance estimator:

$$\hat{V}_{\text{cv}}(\lambda) = n^{-2}\sum_{i=1}^{n}\{\tau_i(\hat{\eta}_{(-i)}) - \hat{\gamma}'_{(-i)}(\lambda)\xi_i\}^2.$$

In Appendix B, we justify that this estimator is better than $\hat{V}_{\text{lasso}}(\lambda)$. Moreover, when λ is close to zero and p is large, that is, one almost uses the standard least square procedure to obtain $\hat{\gamma}_{(-i)}(\lambda)$, the above variance estimate can be modified slightly for improving its estimation accuracy (see Appendix B for details). A natural "optimal" estimator using the above lasso procedure is $\hat{\theta}_{\text{opt}} = \hat{\theta}_{\text{cv}}(\hat{\lambda})$, where $\hat{\lambda}$ is the penalty parameter value, which minimizes $\hat{V}_{\text{cv}}(\lambda)$ over a range of λ values of interest. As a referee kindly pointed out, when θ_0 is the mean difference, one may replace (2.1) by the simple least squared objective function

$$\sum_{i=1}^{n}\left\{\frac{T_i - \pi}{\pi(1-\pi)}\right\}^2(Y_i - \gamma'\mathbf{Z}_i)^2$$

without the need of estimating the influence function.

3. Applications

In this section, we show how to apply the new estimation procedure to various cases. To this end, we only need to determine the initial estimate $\hat{\theta}$ for the contrast measure of interest and its corresponding first-order expansion in each application. First, we consider the case that the response is continuous or binary and the group mean difference is the parameter of interest. Here,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\pi} - \frac{(1-T_i)Y_i}{1-\pi}\right\}.$$

In this case, it is straightforward to show that

$$\hat{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i(Y_i - \hat{\mu}_1)}{\pi} - \frac{(1-T_i)(Y_i - \hat{\mu}_0)}{1-\pi}\right\} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where $\eta = (\mu_1, \mu_0)'$, $\hat{\mu}_1 = \sum_{i=1}^{n} T_i Y_i/\pi n$, and $\hat{\mu}_0 = \sum_{i=1}^{n}(1-T_i)Y_i/(1-\pi)n$.

Now, when the response is binary with success rate p_j for the treatment group j, $j = 0, 1$, but $\theta_0 = \log\{p_1(1-p_0)/p_0/(1-p_1)\}$, then

$$\hat{\theta} = \log(\hat{p}_1) - \log(1-\hat{p}_1) - \log(\hat{p}_0) + \log(1-\hat{p}_0),$$

where $\hat{p}_1 = \sum_{i=1}^{n} T_i Y_i/\pi n$, and $\hat{p}_0 = \sum_{i=1}^{n}(1-T_i)Y_i/(1-\pi)n$. For this case,

$$\hat{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(Y_i - \hat{p}_1)T_i}{\pi\hat{p}_1(1-\hat{p}_1)} - \frac{(Y_i - \hat{p}_0)(1-T_i)}{(1-\pi)\hat{p}_0(1-\hat{p}_0)}\right\} + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Last, we consider the case when Y is the time to a specific event but may be censored by an independent censoring variable. To be specific, we observe $(\tilde{Y}, \Delta)$ where $\tilde{Y} = Y \wedge C$, $\Delta = I(Y < C)$, C is the censoring time, and $I(\cdot)$ is the indicator function. A most commonly used summary measure for quantifying the treatment difference in survival analysis is the ratio of two hazard functions. The two sample Cox estimator is often used to estimate such a ratio. However, when the proportional hazards assumption between two groups is not valid, this estimator converges to a parameter which may be difficult to interpret as a measure of the treatment difference. Moreover, this parameter depends on the censoring distribution. Therefore, it is desirable to use a model-free summary measure for the treatment contrast. One may simply use the survival probability at a given time t_0 as a model-free summary for survivorship. For this case, $\theta_0 = P(Y > t_0|T = 1) - P(Y > t_0|T = 0)$ and $\hat{\theta} = \hat{S}_1(t_0) - \hat{S}_0(t_0)$, where $\hat{S}_j(\cdot)$ is the Kaplan–Meier estimator of the survival function of Y in group j, $j = 0, 1$. For this case,

$$\hat{\theta} - \theta_0 = -n^{-1}\sum_{i=1}^{n}\left[\frac{T_i}{\pi}\int_0^{t_0}\frac{\hat{S}_1(t_0)d\hat{M}_{i1}(s)}{\sum_{j=1}^{N} I(\tilde{Y}_j \geqslant s)T_j} - \frac{1-T_i}{1-\pi}\int_0^{t_0}\frac{\hat{S}_0(t_0)d\hat{M}_{i0}(s)}{\sum_{j=1}^{N} I(\tilde{Y}_j \geqslant s)(1-T_j)}\right] + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where

$$\hat{M}_{ij}(s) = I(T_i = j)\left[I(\tilde{Y}_i \leqslant s)\Delta_i - \int_0^{s} I(\tilde{Y}_i \geqslant u)d\hat{\Lambda}_j(u)\right],$$

and $\hat{\Lambda}_j(\cdot)$ is the Nelson–Alan estimator for the cumulative hazard function of Y in group j (Flemming and Harrington, 1991).

To summarize a global survivorship beyond using t-year survival rates, one may use the mean survival time. Unfortunately, in the presence of censoring, such a measure cannot be estimated well. An alternative is to use the so-called restricted mean survival time, that is, the area under the survival function up to time point t_0. The corresponding consistent estimator is the area under the Kaplan–Meier curve. For this case, $\theta_0 = E(Y \wedge t_0 | T = 1) - E(Y \wedge t_0 | T = 0)$ and

$$\hat{\theta} = \int_0^{t_0} \hat{S}_1(s)\mathrm{d}s - \int_0^{t_0} \hat{S}_0(s)\mathrm{d}s,$$

For this case,

$$\begin{aligned}\hat{\theta} - \theta_0 &= n^{-1}\sum_{i=1}^{n}\left[-\frac{T_i}{\pi}\int_0^{t_0}\left\{\frac{\int_s^{t_0}\hat{S}_1(t)\mathrm{d}t}{\sum_{j=1}^{N} I(\tilde{Y}_j \geqslant s)T_j}\right\}\mathrm{d}\hat{M}_{i1}(s)\right.\\ &\qquad \left.+\frac{1-T_i}{1-\pi}\int_0^{t_0}\left\{\frac{\int_s^{t_0}\hat{S}_0(t)\mathrm{d}t}{\sum_{j=1}^{N} I(\tilde{Y}_j \geqslant s)(1-T_j)}\right\}\mathrm{d}\hat{M}_{i0}(s)\right] + o_p(\frac{1}{\sqrt{n}}).\end{aligned}$$

4. A SIMULATION STUDY

We conducted an extensive simulation study to examine the finite sample performance of the new estimates $\hat{\theta}_{\mathrm{cv}}(\lambda)$ and $\hat{\theta}_{\mathrm{opt}}$ for θ_0. First, we investigate whether $\hat{V}_{\mathrm{cv}}(\lambda)$ estimates the true variance of $\hat{\theta}_{\mathrm{cv}}(\lambda)$ well under various practical settings. We also examine the finite sample properties for the interval estimation procedure based on the optimal $\hat{\theta}_{\mathrm{opt}}$. To this end, we consider the following models for generating the underlying data:

1. the linear regression model with continuous response

$$Y = m_T(\mathbf{Z}) + N(0, 1);$$

2. the logistic regression model with binary response

$$\mathrm{P}(Y = 1|T, \mathbf{Z}) = [1 + \exp\{-m_T(\mathbf{Z})\}]^{-1};$$

3. the Cox regression model with survival response

$$Y = \epsilon_0 \exp\{m_T(\mathbf{Z})\},$$

where ϵ_0 and censoring time are generated from the unit exponential distribution and $U(0, 3)$, respectively, and we are interested in survival curves over the time interval $[0, t_0] = [0, 2.5]$.

Throughout we let $n = 200$ and generate $(Z_{[1]}, \ldots, Z_{[100]})'$ from multivariate normal distribution with mean 0, variance 1, and a compound symmetry covariance $\wp$ chosen to be either 0 or 0.5. For each generated data set, the 20-fold cross validation is used to calculate $\hat{\theta}_{\mathrm{cv}}(\lambda)$ and $\hat{V}_{\mathrm{cv}}(\lambda)$ over a sequence of tuning parameters $\{\lambda_1, \lambda_2, \ldots, \lambda_{100}\}$, where λ_1 is chosen such that $\hat{\gamma}(\lambda_1) = 0$ for all simulated data sets, $\lambda_k = 10^{-3/98}\lambda_{k-1}$ for $k = 2, \ldots, 99$ and $\lambda_{100} = 0$. In the first set of simulation, we set

$$m_0(\mathbf{Z}) = \sum_{j=1}^{20}\frac{j}{20}Z_{[j]}, \quad m_1(\mathbf{Z}) = 1 + \sum_{j=1}^{20}\frac{j}{20}Z_{[j]}.$$

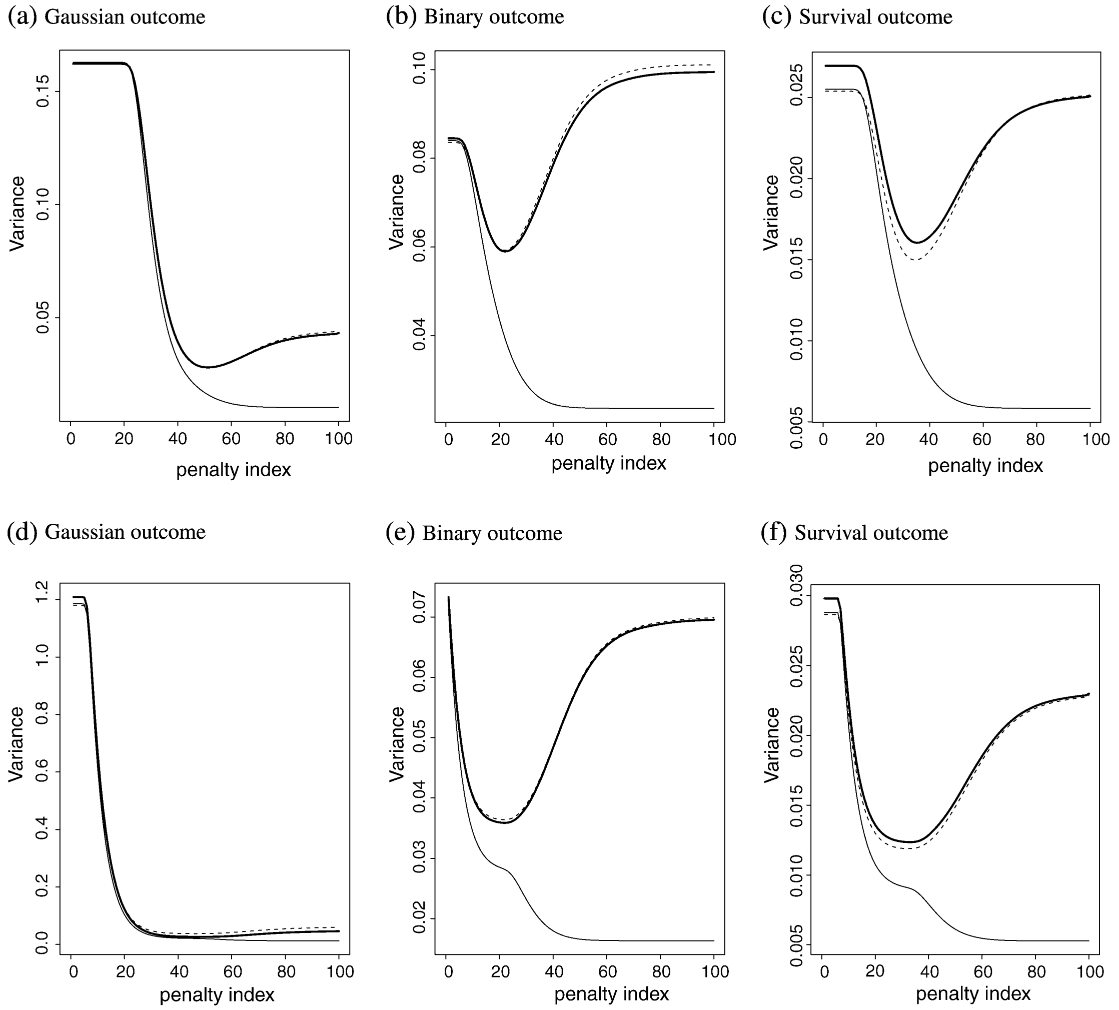

Fig. 1. Comparing various estimates for $\hat{\theta}_{\text{cv}}(\lambda)$ at $\{\lambda_1, \ldots, \lambda_{100}\}$: the empirical variance of $\hat{\theta}_{\text{cv}}(\lambda)$ (black curve); $\hat{V}_{\text{cv}}(\lambda)$ (dashed curve); $\hat{V}_{\text{lasso}}(\lambda)$ (grey curve); (a–c) for independent coviariate; (d–f) for dependent covariate.

All the results are summarized based on 5000 replications. In Figure 1, we present the average of $\hat{V}_{\text{cv}}(\lambda)$, the average of $\hat{V}_{\text{lasso}}(\lambda)$, and the empirical variance of $\hat{\theta}_{\text{cv}}(\lambda)$ when $\wp = 0$ for continuous, binary, and survival responses, respectively. The results suggest that $\hat{V}_{\text{cv}}(\lambda)$ approximates the true variance of $\hat{\theta}_{\text{cv}}(\lambda)$ very well; while $\hat{V}_{\text{lasso}}(\lambda)$ obtained without cross validation tends to severely underestimate the true variance. When the covariates are correlated with $\wp = 0.5$, the corresponding results are presented in Figure 1. The results are consistent with the case with $\wp = 0$.

Next, we examine the performance of the optimal estimator $\hat{\theta}_{\text{opt}} = \hat{\theta}_{\text{cv}}(\hat{\lambda})$, where $\hat{\lambda}$ is chosen to be the minimizer of $\hat{V}_{\text{cv}}(\lambda)$, $\lambda \in \{\lambda_1, \ldots, \lambda_{100}\}$. For each simulated data set, we construct a 95% confidence intervals (CI) based on $\hat{\theta}_{\text{opt}}$ and $\hat{V}_{\text{opt}} = \hat{V}_{\text{cv}}(\hat{\lambda})$. We summarized results from the 5000 replications based on the empirical bias, standard error, and coverage level and length of the constructed CIs. For comparisons, we also obtain those values based on the simple estimator $\hat{\theta}$, $\hat{\theta}_{\text{ZTD}}$, and $\hat{\theta}_{\text{cv}}(\lambda_0)$ along with their variance

Table 1. *The empirical bias, standard error, and coverage levels and lengths for the 0.95 CI based on $\hat{\theta}$, $\hat{\theta}_{\text{opt}}$, $\hat{\theta}_{\text{cv}}(\lambda_0)$, and $\hat{\theta}_{\text{ZTD}}$*

Response	Estimator	Independent covariates				Correlated covariates			
		$m_T(\mathbf{Z}) = \sum_{j=1}^{20} jZ_{[j]}/20 + T$							
		BIAS	ESE	EAL (10^{-3})	ECL (%)	BIAS	ESE	EAL (10^{-3})	ECL (%)
Continuous	$\hat{\theta}$	0.007	0.403	1.580 (1.1[†])	94.9	−0.005	1.100	4.264 (3.0)	94.4
	$\hat{\theta}_{\text{opt}}$	0.002	0.169	0.648 (0.6)	94.2	0.001	0.166	0.743 (1.9)	97.0
	$\hat{\theta}_{\text{cv}}(\lambda_0)$	0.002	0.167	0.652 (0.6)	94.7	0.001	0.163	0.749 (1.9)	97.3
	$\hat{\theta}_{\text{ZTD}}$	0.003	0.204	0.622 (0.6)	87.2	−0.001	0.359	0.749 (1.8)	72.6
Binary	$\hat{\theta}$	0.009	0.291	1.136 (0.2)	95.1	0.004	0.271	1.047 (0.3)	94.6
	$\hat{\theta}_{\text{opt}}$	0.003	0.245	0.946 (0.7)	94.6	0.004	0.191	0.745 (0.5)	95.2
	$\hat{\theta}_{\text{cv}}(\lambda_0)$	0.003	0.243	0.953 (0.7)	94.9	0.003	0.189	0.747 (0.5)	95.5
	$\hat{\theta}_{\text{ZTD}}$	−0.011	0.259	0.822 (0.7)	88.9	−0.005	0.201	0.508 (0.7)	78.9
Survival	$\hat{\theta}$	0.003	0.164	0.626 (0.2)	94.1	0.005	0.173	0.665 (0.1)	94.5
	$\hat{\theta}_{\text{opt}}$	0.001	0.127	0.476 (0.4)	93.7	0.005	0.112	0.426 (0.3)	93.9
	$\hat{\theta}_{\text{cv}}(\lambda_0)$	0.001	0.127	0.479 (0.4)	94.0	0.005	0.111	0.427 (0.3)	94.2
	$\hat{\theta}_{\text{ZTD}}$	0.004	0.141	0.457 (0.4)	89.5	0.005	0.122	0.401 (0.3)	89.8
		$m_T(\mathbf{Z}) = (T+1)\sum_{j=1}^{20}\{(-1)^T(Z_{[j]}^2 - 1)/2 + jZ_{[j]}/20\} + 2T$							
Continuous	$\hat{\theta}$	0.019	0.876	3.476 (2.6)	94.9	0.009	1.502	5.499 (4.4)	93.0
	$\hat{\theta}_{\text{opt}}$	0.002	0.533	2.084 (2.0)	94.4	−0.038	1.188	4.618 (8.4)	93.9
	$\hat{\theta}_{\text{cv}}(\lambda_0)$	0.016	0.530	2.097 (2.0)	94.8	0.069	1.191	4.685 (8.7)	94.5
	$\hat{\theta}_{\text{ZTD}}$	−0.159	0.583	2.068 (2.2)	91.1	0.390	1.305	4.193 (7.1)	88.9
Binary	$\hat{\theta}$	0.023	0.288	1.130 (0.2)	94.3	−0.001	0.290	1.140 (0.3)	95.4
	$\hat{\theta}_{\text{opt}}$	0.017	0.242	0.935 (0.7)	94.7	−0.003	0.188	0.753 (0.6)	95.4
	$\hat{\theta}_{\text{cv}}(\lambda_0)$	0.021	0.240	0.941 (0.7)	95.0%	0.002	0.187	0.757 (0.6)	95.7
	$\hat{\theta}_{\text{ZTD}}$	−0.023	0.265	0.855 (0.8)	88.8	−0.006	0.201	0.546 (0.7)	82.8
Survival	$\hat{\theta}$	−0.003	0.173	0.659 (0.1)	93.7	0.010	0.173	0.663 (0.1)	94.6
	$\hat{\theta}_{\text{opt}}$	−0.005	0.141	0.531 (0.4)	93.6	0.005	0.114	0.431 (0.3)	94.4
	$\hat{\theta}_{\text{cv}}(\lambda_0)$	−0.002	0.140	0.534 (0.4)	93.8	0.007	0.114	0.433 (0.3)	94.6
	$\hat{\theta}_{\text{ZTD}}$	−0.023	0.157	0.515 (0.4)	89.4	0.014	0.120	0.411 (0.3)	91.4

BIAS, empirical bias; ESE, empirical standard error of the estimator; EAL, empirical average length; and ECL: empirical coverage level.

[†]The Monte Carlo standard error in estimating the average length.

estimators, where λ_0 is the minimizer of the empirical variance of $\hat{\theta}_{\text{cv}}(\lambda_0)$. In all the numerical studies, the forward subset selection procedure coupled with BIC is used to select variables for the efficiency augmentation in the ZTD procedure. The results are summarized in Table 1. The coverage levels for $\hat{\theta}_{\text{opt}}$ are close to the nominal counterparts and the interval lengths are almost identical to those based on the estimate with the true optimal λ_0. On the other hand, the simple estimate $\hat{\theta}$ tends to have substantially wider interval estimates than $\hat{\theta}_{\text{opt}}$, $\hat{\theta}_{\text{cv}}(\lambda_0)$, and $\hat{\theta}_{\text{ZTD}}$. The empirical standard error of $\hat{\theta}_{\text{ZTD}}$ is slightly greater than that of $\hat{\theta}_{\text{opt}}$ or $\hat{\theta}_{\text{cv}}(\lambda_0)$, which implies the advantages of lasso procedure. More importantly, the naive variance estimator of $\hat{\theta}_{\text{ZTD}}$ may severely underestimate the true variance and thus results in much more liberal confidence interval estimation procedure, which potentially can be corrected via cross validation. In summary, for all cases studied, the augmented estimators can substantially improve the efficiency of $\hat{\theta}$ in terms of narrowing the average length of the confidence interval of θ_0 and $\hat{\theta}_{\text{opt}}$-based inference is more reliable than that based on $\hat{\theta}_{\text{ZTD}}$. Furthermore, in the variance estimation for $\hat{\theta}_{\text{opt}} = \hat{\theta}_{\text{cv}}(\hat{\lambda})$, the variability in $\hat{\lambda}$ may cause slightly downward bias, which is almost negligible in our empirical studies. Last, all estimators considered here are almost unbiased in the first set of simulation.

For the second set of simulation, we repeat the above numerical study with

$$m_0(\mathbf{Z}) = \sum_{j=1}^{20} \left\{ (Z_{[j]}^2 - 1) + \frac{j}{10} Z_{[j]} \right\}$$

and

$$m_1(\mathbf{Z}) = \sum_{j=1}^{20} \left\{ -(Z_{[j]}^2 - 1) + \frac{j}{10} Z_{[j]} \right\} + 2.$$

We augment the simple estimator by $\mathbf{Z} = (Z_{[1]}, \ldots, Z_{[40]}, Z_{[1]}^2, \ldots, Z_{[40]}^2)'$. The corresponding results are reported in Figure 2(a–f) and Table 1. The results are similar to those from the first set of simulation study except that for the continuous outcome, the empirical bias of $\hat{\theta}_{\text{ZTD}}$ is not trivial relative to the corresponding standard error. On the other hand, the estimate $\hat{\theta}_{\text{opt}}$ is almost unbiased for all cases as ensured by the cross validation procedure. Note that without knowing the practical meanings of the response, the absolute magnitude of the bias alone is difficult to interpret and a seemingly substantial bias relative to the standard error may still be irrelevant in practice. However, the presence of such a bias still poses a risk in making statistical inference on marginal treatment effect. In further simulations (not reported), we have found that the bias cannot be completely eliminated by increasing sample size or including quadratic transformation in $\mathbf{Z}$. Last, we would like to pointed out the presence of bias is a uncommon finite sample phenomenon and does not undermine the asymptotical validity of ZTD and similar procedures. For example, under the aforementioned setup if we reduce the dimension of $\mathbf{Z}$ to 10 and increase the sample size to 500, then the bias becomes essentially 0.

For the third set of simulation, we examine the potential efficiency loss due to not including important nonlinear transformations of baseline covariates in the efficiency augmentation. To this end, we simulate continuous, binary and survival outcomes as in the previous stimulation study with

$$m_0(\mathbf{Z}) = \sum_{j=1}^{20} \left\{ \frac{Z_{[j]}^2 - 1}{2} + \frac{j}{20} Z_{[j]} \right\}$$

and

$$m_1(\mathbf{Z}) = \sum_{j=1}^{20} \left\{ -(Z_{[j]}^2 - 1) + \frac{j}{10} Z_{[j]} \right\} + 2.$$

We augment the efficiency of the initial estimator first by $\mathbf{Z}_1 = (Z_{[1]}, \ldots, Z_{[100]})'$ and second by $\mathbf{Z}_2 = (Z_{[1]}, \ldots, Z_{[100]}, Z_{[1]}^2, \ldots, Z_{[20]}^2)'$. In Table 2, we present the empirical bias and standard error of $\hat{\theta}_{\text{opt}}$ based on 5000 replications. As expected, the empirical performance of the estimator augmented by $\mathbf{Z}_2$ is superior to that of its counterpart using $\mathbf{Z}_1$. The gains in efficiency for binary and survival outcomes are less significant than that for continuous outcome, which is likely due to the fact that the influence function of $\hat{\theta}$ is neither a linear nor a quadratic function of $Z_{[j]}, j = 1, \ldots, 100$ in the binary or survival setting.

In the fourth set of simulation, we examine the "null model" setting in which none of the covariates are related to the response. To this end, we generate continuous responses Y from the normal distribution

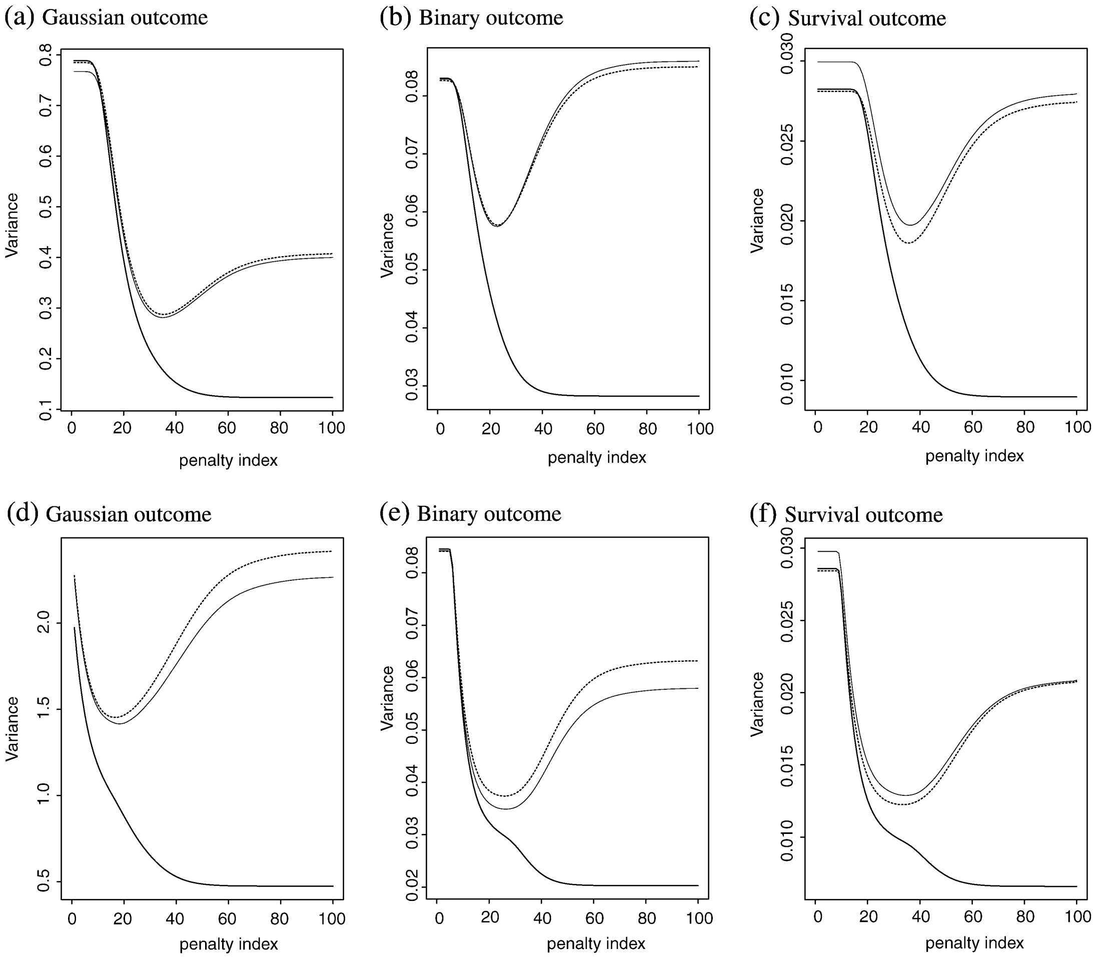
(a) Gaussian outcome
(b) Binary outcome
(c) Survival outcome
(d) Gaussian outcome
(e) Binary outcome
(f) Survival outcome

Fig. 2. Comparing various estimates for $\hat{\theta}_{\text{cv}}(\lambda)$ at $\{\lambda_1, \ldots, \lambda_{100}\}$: the empirical variance of $\hat{\theta}_{\text{cv}}(\lambda)$ (black curve); $\hat{V}_{\text{cv}}(\lambda)$ (dashed curve); $\hat{V}_{\text{lasso}}(\lambda)$ (grey curve) ; (a–c) for independent coviariate; (d–f) for dependent covariate.

Table 2. *The empirical bias and standard error of $\hat{\theta}_{\text{opt}}$ augmented by $\mathbf{Z}_1$ and $\mathbf{Z}_2$*

Response	Augmentation vector	Independent covariates		Correlated covariates	
		BIAS	ESE	BIAS	ESE
Continuous	$\mathbf{Z}_1$	−0.024	0.770	−0.085	1.831
	$\mathbf{Z}_2$	−0.020	0.745	−0.035	1.492
Binary	$\mathbf{Z}_1$	−0.001	0.261	−0.004	0.239
	$\mathbf{Z}_2$	0.001	0.258	−0.002	0.226
Survival	$\mathbf{Z}_1$	0.037	0.156	0.004	0.133
	$\mathbf{Z}_2$	0.037	0.154	0.003	0.124

BIAS, empirical bias; ESE, empirical standard error.

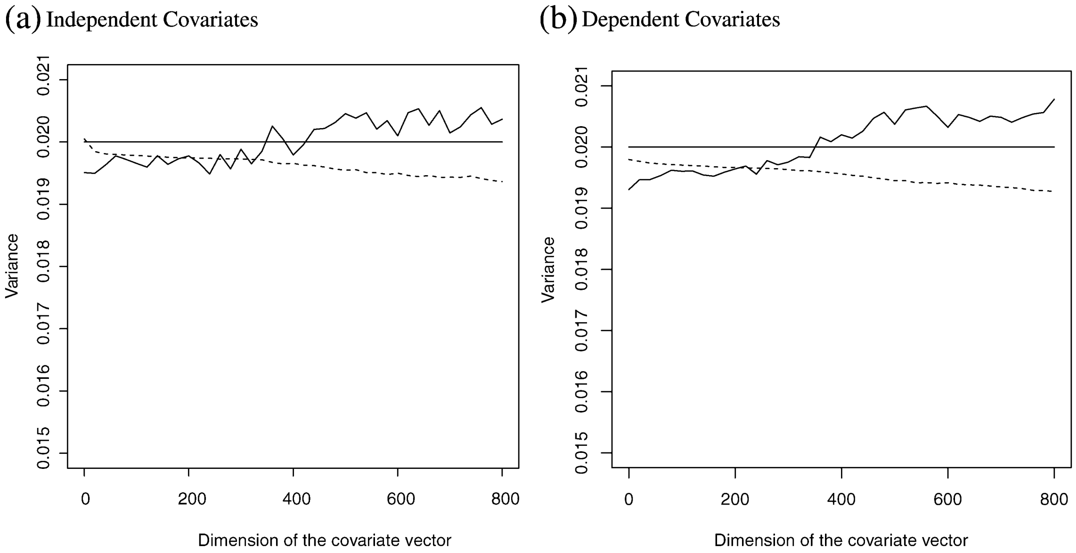

Fig. 3. Empirical variance of $\hat{\theta}_{\text{opt}}$ (wiggly solid curve) and its variance estimator (dashed curve) in the presence of high-dimensional noise covariates. The horizontal solid curve presents the optimal variance level.

$N(0, 1)$ for $T = 0$ and $N(1, 1)$ for $T = 1$. The covariate $\mathbf{Z}$ is from a standard multivariate normal distribution generated independent of Y. For each generated data set, we obtain the optimal estimator $\hat{\theta}_{\text{opt}}$ and its variance estimator as in the previous simulation study. Based on 3000 replications, we estimate the empirical variance of $\hat{\theta}_{\text{opt}}$ and the average of the variance estimator for given combination of n and p. To examine the effect of "overadjustment", we let $p = 0, 20, 40, \ldots, 780$ and 800 while fixing the sample size n at 200. In Figure 3, we present the empirical average for $\hat{V}_{\text{cv}}(\hat{\lambda})$ (dashed curve) and the empirical variance of $\hat{\theta}_{\text{opt}}$ (solid curve). The optimal estimator is the naive estimator $\hat{\theta}$ without any covariate-based augmentation in this case. The figure demonstrates that the variance of $\hat{\theta}_{\text{opt}}$ increases very slowly with the dimension p and is still near the optimal level even with 800 noise covariates. The variance estimator slightly underestimates the true variance and the downward bias increases with the dimension p, which could be attributable to the fact that we use $\hat{V}_{cv}(\hat{\lambda}) = \min_{\lambda}\{\hat{V}_{\text{cv}}(\lambda)\}$ as the variance estimator without any adjustments. On the other hand, the bias remains rather low ($<6\%$ of the empirical variance) such that the valid inference on θ_0 can still be made over the entire range of p. In Figure 3, we represent the similar results with noise covariates generated from dependent multivariate normal distribution as in the previous simulation studies.

5. An example

We illustrate the new proposal with the data from a clinical trial to compare D-penicillmain and placebo for patients with primary biliary cirrhosis (PBC) of liver (Therneau and Grambsch, 2000). The primary endpoint is the time to death. The trial was conducted between 1974 and 1984. For illustration, we use the difference of two restricted mean survival time up to $t_0 = 3650$ (days) as the primary parameter θ_0 of interest. Moreover, we consider 18 baseline covariates for augmentation: gender, stages (1, 2, 3, and 4), presence of ascites, edema, hepatomegaly or enlarged liver, blood vessel malformations in the skin, log-transformed age, serum albumin, alkaline phosphotase, aspartate aminotransferase, serum bilirubin, serum cholesterol, urine copper, platelet count, standardized blood clotting time, and triglycerides. There

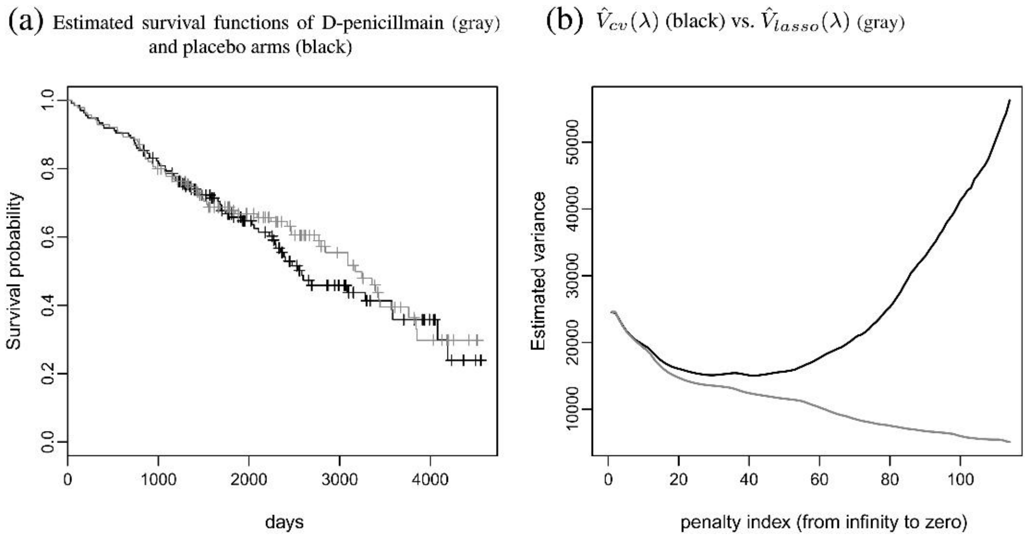

Fig. 4. Analysis results for PBC data.

are 276 patients with complete covariate information (136 and 140 in control and D-penicillmain arms, respectively). The data used in our analysis are given in the Appendix D.1 of Flemming and Harrington (1991). Figure 4 provides the Kaplan–Meier curves for the two treatment groups. The simple two sample estimate $\hat{\theta}$ is 115.2 (days) with an estimated standard error $\hat{V}$ of 156.6 (days). The corresponding 95% confidence interval for the difference is (−191.8, 422.1) (days). The optimal estimate $\hat{\theta}_{\text{opt}}$ augmented additively with the above 18 coavariates is 106.3 with an estimated standard error $\hat{V}_{\text{opt}}$ of 121.4. These estimates were obtained via a 23-fold cross validation (note that $276 = 23 \times 12$) described in Section 2. The corresponding 95% CI is (−131.8, 344.4). To examine the effect of K on the result, we repeated the analysis with 92-fold cross validation ($n = 276 = 92 \times 3$) and the optimal estimator barely changes (108.3 with a 95% CI of (−128.5, 345.1)). In our limited experience, the estimation result is not sensitive to $K \geqslant \max(20, n^{1/2})$.

To examine how robust the new proposal is with respect to different augmentations. We consider a case which includes the above 18 covariates but also their quadratic terms as well as all their two-way interactions. The dimension of **Z** is 178 for this case. The resulting optimal $\hat{\theta}_{\text{opt}}$ is 110.1 with an estimated standard error of 122.6. Note the resulting estimates are amazingly close to those based on the augmented procedure with 18 covariates only.

To examine the advantage of using the cross validation for the standard error estimation, in Figure 4, we plot $\hat{V}_{\text{cv}}(\lambda)$ and $\hat{V}_{\text{lasso}}(\lambda)$ over the order of 100 λ's, which were generated using the same approach as in Section 4. Note that $\hat{V}_{\text{lasso}}(\lambda)$ is substantially smaller than $\hat{V}_{\text{cv}}(\lambda)$, especially when λ approaches to 0, that is, there is no penalty for the L_2 loss function. For $\hat{\theta}_{\text{opt}}$, $\hat{V}_{\text{lasso}}$ is about 20% smaller than its cross validated counterpart.

It has been shown via numerical studies that the ZTD performs well via the standard stepwise regression by ignoring the sampling variation of the estimated weights when the dimension of **Z** is not large with respect to n. However, it is not clear how the ZTD augmentation performs with a relatively high-dimensional covariate vector **Z**. It would be interesting to compare the ZTD and the new proposal with the PBC data. To this end, we implement ZTD augmentation procedure using (1) baseline covariates ($p = 18$); (2) baseline covariates and their quadratic transformations as well as all their two-way

Table 3. *Comparisons between the new and ZTD estimate with the data from the Mayo Clinic PBC clinical trial (SE: estimated standard error)*

p	The new optimal procedure		ZTD	
	Estimate	SE	Estimate	SE
5	92.0	121.5	96.3	119.4
18	106.3	121.4	126.4	111.7
178	110.1	122.6	65.3	114.6

BIAS, empirical bias; ESE, empirical standard error.

interactions ($p = 178$); and (3) only five baseline covariates: edema and log-transformed age, serum albumin, serum bilirubin, and standardized blood clotting time, which were selected in building a multivariate Cox regression model to predict the patient's survival by Therneau and Grambsch (2000). Note that the ZTD procedure augments the following estimating equations for θ_0:

$$\sum_{i=1}^{n} \frac{(1-T_i)\tilde{\Delta}_i}{\hat{K}_0(\tilde{Y}_i \wedge t_0)}[\tilde{Y}_i \wedge t_0 - a_{t_0}] = 0,$$

$$\sum_{i=1}^{n} \frac{T_i\tilde{\Delta}_i}{\hat{K}_1(\tilde{Y}_i \wedge t_0)}[\tilde{Y}_i \wedge t_0 - a_{t_0} - \theta] = 0,$$

where a_{t_0} is the restricted mean for the comparator and θ is the treatment difference, $\tilde{\Delta}_i = I(Y_i \wedge t_0 < C_i)$, and $\hat{K}_j(\cdot)$ is the Kaplan–Meier estimate for the survival function of censoring time C in group $T = j$, $j = 0, 1$. In Table 3, we present the resulting ZDT point estimates and their corresponding standard error estimates for the above three cases. Here, we used the standard forward stepwise regression procedure to select the augmentation covariates with the entry Type I error rate of 0.10 (Zhang *and others*, 2008; Zhang and Gilbert, 2010). It appears that using the entire data set for selecting covariates and making inferences about θ_0 may introduce nontrivial bias and an overly optimistic standard error estimate when p is large. On the other hand, the new procedure does not lose efficiency and yields similar result as ZTD procedure when p is small.

6. Remarks

The new proposal performs well even when the dimension of the covariates involved for augmentation is not large. The new estimation procedure may be implemented for improving estimation precision regardless of the marginal distributions of the covariate vectors between two treatment groups being balanced. On the other hand, to avoid post ad hoc analysis, we strongly recommend that the investigators prespecify the set of all potential covariates for adjustment in the protocol or the statistical analysis plan before the data from the clinical study are unblinded.

The stratified estimation procedure for the treatment difference is also commonly used for improving the estimation precision using baseline covariate information. Specifically, we divide the population into K strata based on baseline variables, denoted by $\{\mathbf{Z} \in B_1\}, \ldots, \{\mathbf{Z} \in B_K\}$, the stratified estimator is

$$\hat{\theta}_{\text{str}} = \frac{\sum_{k=1}^{K} \hat{\theta}_k w_k}{\sum_{k=1}^{K} w_k},$$

where $\hat{\theta}_k$ and w_k are corresponding simple two sample estimator for the treatment difference and the weight for the kth stratum, $k = 1, \ldots, K$. In general, the underlying treatment effect may vary across strata and consequently the stratified estimator may not converge to θ_0. If θ_0 is the mean difference between two groups and w_k is the size of the kth stratum, $\hat{\theta}_{\text{str}}$ is a consistent estimator for θ_0. Like the ANCOVA, the stratified estimation procedure may be problematic. On the other hand, one may use the indicators $\{I(\mathbf{Z} \in B_1), \ldots, I(\mathbf{Z} \in B_K)\}'$ to augment $\hat{\theta}$ to increase the precision for estimating the treatment difference θ_0.

In this paper, we follow the novel approach taken, for example, by Zhang *and others* (2008) for augmenting the simple two sample estimator but present a systematic practical procedure for choosing covariates for making valid inferences about the overall treatment difference. When p is large, there are several advantages over other approaches for augmenting $\hat{\theta}$ with covariates. First, it avoids the complex variable selection step in two arms separately as proposed in Zhang *and others* (2008). Second, compared with other variable selection methods such as the stepwise regression, the lasso method directly controls the variability of $\hat{\gamma}$, which improves the empirical performance of the augmented estimator. Third, the cross validation step enables more accurate estimation of the variance of the augmented estimator. When λ increases from 0 to $+\infty$, the resulting estimator varies from the fully augmented estimator using all the components of $\mathbf{Z}_i$ to $\hat{\theta}$. The lasso procedure also possesses superior computational efficiency with high-dimensional covariates to alternatives. Last, since $\hat{\theta}_{\text{ZTD}}$ can also be viewed as a generalized method of moment estimator with

$$\begin{pmatrix} \theta - \hat{\theta}_0 \\ n^{-1} \sum_{i=1}^{n} \xi_i \end{pmatrix} \approx 0$$

as moment conditions (Hall, 2005), the cross validation method introduced here may be extended to a much broader context than the current setting.

It is important to note that if a permuted block treatment allocation rule is used for assigning patients to the two treatment groups, the augmentation method proposed in the paper can be easily modified. For instance, for the K-fold cross validation process, one may choose the sets $\{\mathcal{D}_k, k = 1, \ldots, K\}$ so that each permuted block would not be in different sets.

For assigning patients to the treatment groups, a stratified random treatment allocation rule is also often utilized to ensure a certain level of balance between the two groups in each stratum. For this case, a weighted average θ_0 of the treatment differences θ_{k0} with weight $w_k, k = 1, \ldots, K$, across K strata may be the parameter of interest for quantifying an overall treatment contrast. Let $\hat{\theta}_k$ be the simple two sample estimator for θ_{k0} and $\hat{w}_k$ be the corresponding empirical weight for w_k. Then the weight average $\hat{\theta} = \sum_k \hat{w}_k \hat{\theta}_k / \sum_k \hat{w}_k$ is the simple estimator for θ_0. For the kth stratum, one may use the same approach as discussed in this paper to augment $\hat{\theta}_k$, let the resulting optimal estimator be denoted by $\hat{\theta}_{\text{opt},k}$. Then we can use the weighted average $\sum_k \hat{w}_k \hat{\theta}_{\text{opt},k} / \sum_k \hat{w}_k$ to estimate θ_0. On the other hand, for the case with the dynamic treatment allocation rules (see, e.g., Pocock and Simon, 1975), it is not clear how to obtain a valid variance estimate even for the simple two sample estimator $\hat{\theta}$ (Shao *and others*, 2010). How to extend the augmentation procedure to cases with more complicated treatment allocation rule warrants further research.

APPENDIX A

Asymptotical equivalence between ZTD and ANCOVA

When the group mean is the parameter of interest, the naive estimator for θ_0 can viewed as the root of the estimating equation

$$\sum_{i=1}^{n} \begin{pmatrix} T_i \\ 1-T_i \end{pmatrix} S_0(\theta, a, Y_i, T_i) = \sum_{i=1}^{n} \begin{pmatrix} T_i \\ 1-T_i \end{pmatrix} (Y_i - a - T_i\theta) = 0,$$

where $a = E(Y|T=0)$ is a nuisance parameter. In the ZTD augmentation procedure, one may augment this simple estimating equation via following steps:

- Obtain the initial estimator

  $$\begin{pmatrix} \hat{\theta} \\ \hat{a} \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \frac{(T_i-\pi)Y_i}{\pi(1-\pi)} \\ \frac{(1-T_i)Y_i}{1-\pi} \end{pmatrix}$$

  from the original estimating equation

- Obtain $\hat{\beta}_1$ and $\hat{\beta}_0'$ by minimizing the objective function

  $$\sum_{i=1}^{n} T_i\{S_0(\hat{\theta}, \hat{a}, Y_i, T_i) - \beta_1'\mathbf{Z}_i\}^2$$

  and

  $$\sum_{i=1}^{n}(1-T_i)\{S_0(\hat{\theta}, \hat{a}, Y_i, T_i) - \beta_0'\mathbf{Z}_i\}^2,$$

  respectively. In other words, using $\hat{\beta}_j'\mathbf{Z}$ to approximate $E\{S_0(\theta_0, a_0; Y, T)|\mathbf{Z}, T = j\}$.

- Solve the augmented estimating equations

  $$\sum_{i=1}^{n} \begin{pmatrix} T_i \\ 1-T_i \end{pmatrix} S_0(\theta, a, Y_i, T_i) - \sum_{i=1}^{n}(T_i - \pi) \begin{pmatrix} \hat{\beta}_1'\mathbf{Z}_i \\ -\hat{\beta}_0'\mathbf{Z}_i \end{pmatrix} = 0$$

  to obtain $\hat{\theta}_{\text{ZTD}}$.

The resulting $\hat{\theta}_{\text{ZTD}}$ is always asymptotically more efficient than the naive counterpart and a simple sandwich variance estimator can be used to consistently estimate the variance of the new estimator. It has been shown that $\hat{\theta}_{\text{ZTD}}$ is asymptotically the most efficient one from the class of the estimators

$$\mathcal{A} = \left\{ \hat{\theta}_\gamma = \hat{\theta} - \gamma' \left\{ n^{-1} \sum_{i=1}^{n} \frac{(T_i - \pi)\mathbf{Z}_i}{\pi(1-\pi)} \right\} \,\middle|\, \gamma \in R^p \right\},$$

whose members are all consistent for θ_0 and asymptotically normal. When $\pi = 0.5$

$$\hat{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n}\{2(2T_i - 1)Y_i - \theta_0\},$$

the optimal weight minimizing the variance of

$$\hat{\theta} - \gamma'\frac{1}{n}\sum_{i=1}^{n} 2(2T_i - 1)\mathbf{Z}_i$$

is simply

$$[E\{2(2T_i - 1)\mathbf{Z}_i\}^{\otimes 2}]^{-1} E[2(2T_i - 1)\mathbf{Z}_i\{2(2T_i - 1)Y_i - \theta_0\}] = [E(\mathbf{Z}_i^{\otimes 2})]^{-1} E(\mathbf{Z}_i Y_i) = \gamma_0.$$

Therefore, $\hat{\theta}_{\text{ZTD}}$ is asymptotically equivalent to the commonly used ANCOVA estimator. This equivalence is noted in Tsiatis *and others* (2008).

APPENDIX B

Justification of the cross validation based variance estimator for $\hat{\theta}_{\rm cv}(\lambda)$

To justify the cross validation based variance estimator, first consider the expansion

$$\hat{\theta}_{\rm cv}(\lambda) = \left\{\hat{\theta} - \gamma_0'\left(n^{-1}\sum_{i=1}^{n}\xi_i\right)\right\} - n^{-1}\sum_{i=1}^{n}\{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}'\xi_i.$$

The variance of $\hat{\theta}_{\rm cv}(\lambda)$ can be expressed as $V_{11} + V_{22} - 2V_{12}$, where

$$V_{11} = E\left\{\hat{\theta} - \gamma_0'\left(n^{-1}\sum_{i=1}^{n}\xi_i\right)\right\}^2,$$

$$V_{22} = \frac{1}{n^2}E\left[\sum_{i=1}^{n}\{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}'\xi_i\right]^2,$$

and

$$V_{12} = \frac{1}{n}E\left[\left\{\hat{\theta} - \gamma_0'\left(n^{-1}\sum_{i=1}^{n}\xi_i\right)\right\}\sum_{i=1}^{n}\{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}'\xi_i\right].$$

First,

$$\begin{aligned}
V_{12} =& \frac{1}{n^2}E\left[\sum_{i=1}^{n}(\tau_i(\hat{\eta}) - \gamma_0'\xi_i)\sum_{i=1}^{n}\{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}'\xi_i\right]\\
\approx& \frac{1}{n^2}\sum_{i\neq j}E[(\tau_i(\hat{\eta}) - \gamma_0'\xi_i)\{\hat{\gamma}_{(-j)}(\lambda) - \gamma_0\}']E\xi_j + \frac{1}{n^2}\sum_{i=1}^{n}E[(\tau_i(\hat{\eta}) - \gamma_0'\xi_i)\{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}'\xi_i]\\
\approx& \frac{1}{n^2}\sum_{i=1}^{n}E\{\hat{\gamma}_{(-i)}(\lambda) - \gamma_0\}'E[(\tau_i(\hat{\eta}) - \gamma_0'\xi_i)\xi_i] \approx 0.
\end{aligned}$$

Therefore, the variance of the augmented estimator $\hat{\theta}_{\rm cv}(\lambda)$ is approximately

$$\begin{aligned}
&V_{11} + V_{22}\\
&= \frac{1}{n}[E\{(\tau_i(\hat{\eta}) - \gamma_0'\xi_i)^2\} + E\{(\hat{\gamma}_{(-i)}(\lambda) - \gamma_0)'\xi_i\}^2] + \frac{(n-1)}{n}E[\xi_1'\{\hat{\gamma}_{(-1)}(\lambda) - \gamma_0\}\xi_2'\{\hat{\gamma}_{(-2)}(\lambda) - \gamma_0\}]\\
&\approx \hat{V}_{\rm cv}(\lambda) + \frac{(n-1)}{n}E[\xi_1'\hat{\gamma}_{(-1)}(\lambda)\xi_2'\hat{\gamma}_{(-2)}(\lambda)].
\end{aligned}$$

In our experience, $d(\lambda) = E[\xi_1'\hat{\gamma}_{(-1)}(\lambda)\xi_2'\hat{\gamma}_{(-2)}(\lambda)] = O(n^{-2})$ is very small compared with $\hat{V}_{\rm cv}(\lambda) = O(n^{-1})$ and is negligible, when λ is not close 0. Therefore, in general, $\hat{V}_{\rm cv}(\lambda)$ serves as a satisfactory estimator for the variance of $\hat{\theta}_{\rm cv}(\lambda)$. For small λ, to explicitly estimate $d(\lambda)$, the covariance between $\xi_1'\hat{\gamma}_{(-1)}(\lambda)$ and $\xi_2'\hat{\gamma}_{(-2)}(\lambda)$, one may use

$$\hat{d}(\lambda) = \frac{2(K^2-1)}{n(n-1)K}\sum_{1\leqslant i<j\leqslant n}\xi_i'\left\{\frac{K-1}{K}\hat{\gamma}_{(-j)}(\lambda) - \hat{\gamma}(\lambda)\right\}\xi_j'\left\{\frac{K-1}{K}\hat{\gamma}_{(-i)}(\lambda) - \hat{\gamma}(\lambda)\right\} \qquad (6.1)$$

as an ad hoc jackknife-type estimator, where $\hat{\gamma}(\lambda)$ is the lasso solution based on the entire data set. To justify the approximation, first note that when λ is close to 0,

$$\hat{\gamma}(\lambda) - \gamma_0 \approx \sum_{i=1}^{n} \Upsilon_i \quad \text{and} \quad \hat{\gamma}_{(-i)}(\lambda) - \gamma_0 \approx \frac{K}{K-1} \sum_{i \notin \mathcal{D}_{k_i}} \Upsilon_i,$$

where Υ_i is the mean zero influence function from the ith observation for $\hat{\gamma}(\lambda)$. Therefore,

$$d(\lambda) = E[\xi_1' \hat{\gamma}_{(-1)}(\lambda) \xi_2' \hat{\gamma}_{(-2)}(\lambda)] \approx \left(1 - \frac{1}{K^2}\right) E[\xi_1' \Upsilon_2 \xi_2' \Upsilon_1],$$

which can be approximated by $\hat{d}(\lambda)$ and one may use $\hat{V}_{\text{cv}}(\lambda) + (n-1)\hat{d}(\lambda)/n$ as the variance estimator for the augmented estimator. Note that the difference between $\hat{V}_{\text{cv}}$ and its modified version appears to be negligible in all the numerical studies presented in the paper.

Acknowledgements

The authors are grateful to the editor and reviewers for their insightful comments. *Conflict of Interest:* None declared.

Funding

National Institutes of Health (R01 AI052817, RC4 CA155940, U01 AI068616, UM1 AI068634, R01 AI024643, U54 LM008748, R01 HL089778).

References

Flemming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Gilbert, P. B., Sato, M., Sun, X. and Mehrotra, D. V. (2009). Efficient and robust method for comparing the immunogenicity of candidate vaccines in randomized clinical trials. *Vaccine* **27**, 396–401.

Hall, A. (2005). *Generalized Method of Moments (Advanced Texts in Econometrics)*. London: Oxford University Press.

Koch, G., Tangen, C., Jung, J. and Amara, I. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine* **17**, 1863–1892.

Leon, S., Tsiatis, A. and Davidian, M. (2003). Semiparametric efficiency estimation of treatment effect in a pretest-posttest study. *Biometrics* **59**, 1046–1055.

Lu, X. and Tsiatis, A. (2008). Improving efficiency of the log-rank test using auxiliary covariates. *Biometrika* **95**, 676–694.

Pocock, S. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31**, 102–115.

Shao, J., Yu, X. and Zhong, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika* **97**, 347–360.

Therneau, T. and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Tsiatis, A., Davidian, M., Zhang, M. and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine* **27**, 4658–4677.

ZHANG, M. AND GILBERT, P. B. (2010). Increasing the efficiency of prevention trials by incorporating baseline covariates. *Statistical of Communications in Infectious Diseases* **2**. http://www.bepress.com/scid/vol2/iss1/art1. doi:10.2202/1948–4690.1002.

ZHANG, M., TSIATIS, A. AND DAVIDIAN, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64**, 707–715.

[*Received July 17, 2011; revised November 20, 2011; accepted for publication December 2, 2011*]