A method for estimating treatment effects in clinical data associated with personalized medicine: a replication study

by

Zimeng Xie

Senior Honors Thesis

Department of Biostatistics

University of North Carolina at Chapel Hill

April 6, 2016

Approved by:

_____

Dr. Michael Hudgens, Thesis Advisor

_____

Dr. Jane Monaco, Reader

_____

Dr. Eric Bair, Reader

A method for estimating treatment effects in clinical data associated with personalized medicine: a replication study

Introduction

Personalized medicine is a medical model that separates patients into different groups - with medical decisions, practices, interventions and/or products being tailored to the individual patient based on their predicted response or risk of disease (1). Owing to advances in the field of genetics and genomics, the medical community is able to gain understanding of various diseases on a molecular level. As a result, the approach of personalized medicine, an innovative approach that takes into account differences in people's genes, environments and lifestyles, is of great interest to doctors and policymakers alike (2). Numerous governmental agencies, such as National Institutes of Health (NIH) and U.S. Food and Drug Administration (FDA), are actively engaged in research and regulatory activities associated with personalized medicine: The Human Genome Project and thousands of follow-on studies are helping scientists to develop gene-targeted treatments (2, 3); President Obama launched the Precision Medicine Initiative in January 2015, which also seeks to identify genetically-based drivers of disease in order to develop new, more effective treatments (2). However, the concept of personalized medicine is not limited to genetics and has broadened to encompass various personalization measures, such as differential dosing (4).

Usually, there are large numbers of covariates in datasets associated with personalized medicine. To develop strategies for performing analysis on these datasets, it is critical to identify the interactions between treatment and baseline covariates in randomized clinical trials. One commonly used approach is subgroup analysis, where treatment and control arms are compared in different predefined subgroups, such as male and female subjects. More rigorously, the treatment-covariates interactions can be analyzed in a multivariate regression analysis where the product of the treatment variable and a set of baseline covariates are included in the regression model, as demonstrated by Tian *et al.* (5).

In this text, the methodology developed by Tian *et al.* is verified by the author via a number of numerical simulations: an arbitrary collection of random variables are generated to represent baseline covariates, and "true" treatment effect for each subject is calculated using a preset formula. By coding

the treatment variable as ± 1 and fitting the products of the treatment variable and baseline covariates (which essentially are treatment/covariate interaction terms) in a LASSO regression model, a score could be constructed to select a subgroup of patients who may benefit from a specific treatment. Subsequently, the method is applied to the collection of random variables to determine treatment scores for all subjects. Finally, Spearman's correlation coefficients between treatment scores and treatment effect are calculated to evaluate the performance of the score. A high Spearman's correlation coefficient suggests that the calculated score is a good predictor of "true" treatment effect, and vice versa.

Methods

In this section, we define $T = \pm 1$ as the binary treatment indicator. $Y^{(1)}$ and $Y^{(-1)}$ are defined as the potential outcome for patients who received treatment $T = 1$ and -1 respectively. We only observe $Y = Y^{(T)}, T$ and $\mathbf{Z}$, a q-dimensional baseline covariate vector. We assume that treatment is randomly assigned to a patient, which means that T and $\mathbf{Z}$ are independent. Furthermore, we let $\mathbf{W}$ be a p-dimensional function of baseline covariates $\mathbf{Z}$. $\mathbf{W}(\mathbf{Z}_i)$ is denoted as $\mathbf{W}_i$ in the rest of this article. For simplicity, we also assume that $P(T = 1) = P(T = -1) = 1/2$. Finally, we assume that Y is a continuous response. We examined two different methods in the article authored by Tian *et al.*: the full regression method and the modified covariate method.

Full Regression Method

When Y is a continuous response, a simple multivariate linear regression model could be constructed as follows to include interactions between the treatment and the covariates:

$$Y = \beta_0^{'}\mathbf{W}(\mathbf{Z}) + \gamma_{\mathbf{0}}^{'}\mathbf{W}(\mathbf{Z}) \cdot T/2 + \epsilon \tag{1}$$

where $\beta_0^{'}$ and $\gamma_0^{'}$ are transposed vectors of coefficients in the linear model and ϵ is the mean zero random error. According to Tian *et al.*, the interaction term $\gamma_{\mathbf{0}}^{'}\mathbf{W}(\mathbf{Z}) \cdot T$ models the differential treatment effect across the population and the linear combination of $\gamma_{\mathbf{0}}^{'}\mathbf{W}(\mathbf{Z})$ can be used for identifying potential subgroup(s) of patients who may benefit from the treatment. Since the vector $\mathbf{W}(\mathbf{Z})$ contains an intercept, the main effect for treatment is always included in the model. Specifically, under model (1), we have

$$\mathbf{\Delta}(\mathbf{z}) = E(Y^{(1)} - Y^{(-1)}|\mathbf{Z} = \mathbf{z}) = \gamma_{\mathbf{0}}^{'}\mathbf{W}(\mathbf{Z}) \tag{2}$$

that is, $\mathbf{\Delta}(\mathbf{z}) = \gamma_{\mathbf{0}}^{'}\mathbf{W}(\mathbf{Z})$ measures the causal treatment effect for patients with the baseline covariate $\mathbf{z}$.

With observed data, γ_0 can be estimated via the ordinary least squares method.

Modified Covariate Method

Alternatively, we consider the following model with only interaction terms:

$$Y = \gamma_{\mathbf{0}}^{'}\mathbf{W}(\mathbf{Z}) \cdot T/2 + \epsilon \tag{3}$$

where ϵ is random error. That is, we perform a regular linear regression using the product of each component of $\mathbf{W}_i$ and one-half the treatment indicator (T = $\pm$1), without including an intercept. The detailed steps for carrying out this operation is as follows:

1. Transform the covariate

$$Z_i \rightarrow \mathbf{W_i} = \mathbf{W}(\mathbf{Z_i}) \rightarrow \mathbf{W_i^*} = \mathbf{W_i} \cdot T_i/2 \tag{4}$$

2. Fit linear regression

$$Y \sim \gamma_{\mathbf{0}}^{'}\mathbf{W}^* \tag{5}$$

based on modified observations without intercept

$$(\mathbf{W_i^*}, Y_i) = \{(\mathbf{W_i} \cdot T_i)/2, Y_i\}, i = 1, 2, \cdots, N. \tag{6}$$

3. $\hat{\gamma}^{'}\mathbf{W}(\mathbf{z})$ could be used to stratify patients for individualized treatment selection.

(Refer to Tian *et al.*, Journal of American Statistical Association, December 2014, Volume 109, for more detailed information)

Simulation Studies

In this section, the author performed independent numerical studies with similar data settings to validate the performance of both methods as demonstrated by Tian *et al.* with continuous responses. For each set of simulation, we generated a training dataset and a validation dataset with sample sizes of $N_1 = 100$ and $N_2 = 300$, respectively. For both datasets, we generated independent Gaussian samples from the regression model

$$Y = \left(\beta_0 + \sum_{j=1}^{p} \beta_j Z_j\right)^2 + \left(\gamma_0 + \sum_{j=1}^{p} \gamma_j Z_j + 0.8 Z_1 Z_2\right) T + \sigma_0 \cdot \epsilon, \tag{7}$$

where the covariates $(Z_1, Z_2, \cdots, Z_p)$ follow a mean zero multivariate normal distribution with a compound symmetric variance-covariance matrix, $(1-\rho)\mathbf{I}_p + \rho \mathbf{1}'\mathbf{1}$, and $\epsilon \sim N(0,1)$. We let $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \cdots, \gamma_p) = (0.4, 0.8, -0.8, 0.8, -0.8, 0, \cdots, 0), \sigma_0 = \sqrt{2}, N_1 = 100$, and $p = 50$ and 1000 representing low and high-dimensional cases, respectively. The treatment was generated as 1 or -1 with equal probability (essentially a transformed Bernoulli random variable). We consider 4 different data settings:

1. $\beta_0 = (\sqrt{6})^{-1}, \beta_1 = \beta_2 = 0, \beta_3 = \beta_4 = \cdots = \beta_{10} = (2\sqrt{6})^{-1}$, and $\beta_{11} = \cdots = \beta_p = 0$. $\rho = 0$.
2. $\beta_0 = (\sqrt{6})^{-1}, \beta_1 = \beta_2 = 0, \beta_3 = \beta_4 = \cdots = \beta_{10} = (2\sqrt{6})^{-1}$, and $\beta_{11} = \cdots = \beta_p = 0$. $\rho = 1/3$.
3. $\beta_0 = (\sqrt{3})^{-1}, \beta_1 = \beta_2 = 0, \beta_3 = \beta_4 = \cdots = \beta_{10} = (2\sqrt{3})^{-1}$, and $\beta_{11} = \cdots = \beta_p = 0$. $\rho = 0$.
4. $\beta_0 = (\sqrt{3})^{-1}, \beta_1 = \beta_2 = 0, \beta_3 = \beta_4 = \cdots = \beta_{10} = (2\sqrt{3})^{-1}$, and $\beta_{11} = \cdots = \beta_p = 0$. $\rho = 1/3$.

According to Tian *et al.*, settings 1 and 2 represent cases with relatively small main effects, where the variations in responses contributable to the main effect, interaction, and random error were about 37.5%, 37.5%, and 25%, respectively, when the covariates were correlated. Settings 3 and 4 represent cases with relatively large main effects, where the variations in responses contributable to the main effect, interaction, and random error were about 75%, 15%, and 10%, respectively, when the covariates were correlated.

For each of the simulated dataset, we applied both the full regression method and the modified covariate method:

• Full regression: we fitted the model specified in formula (1), with complete main effects and covariate-treatment interaction terms. LASSO was used to select the variables.

• Modified covariate: we fitted the model specified in formula (3), with the modified covariate $\mathbf{W}^* = (1, \mathbf{Z})' \cdot T/2$. LASSO was also used to select the variables.

For both methods, LASSO penalty parameter was selected by 10-fold cross validation. The outputs of both methods are estimated scores, $\hat{\gamma}' \mathbf{W}(\mathbf{z})$, where $\hat{\gamma}'$ is the estimated coefficients for the covariate-treatment interaction terms (corresponds to γ's in formula (1)) in the full regression method and the estimated coefficients for modified covariates (corresponds to γ's in formula (3)) in the modified covariate method. The performance of the resulting score is evaluated by estimating the Spearman's rank correlation coefficient between the estimated score and the "true" treatment effect

$$\mathbf{\Delta}(\mathbf{z}) = E(Y^{(1)} - Y^{(-1)}|\mathbf{Z} = \mathbf{z}) = 1.6 \times (0.5 + Z_1 - Z_2 + Z_3 - Z_4 + Z_1 Z_2)$$

in the validation set of size $N_2 = 300$ generated above. Based on 500 sets of simulations for the low-dimensional case and 200 sets of simulations for the high-dimensional case, we plotted the boxplots of the rank correlation coefficients between the estimated scores $\hat{\gamma}^{'}\mathbf{W}(\mathbf{Z})$ and $\mathbf{\Delta}(\mathbf{Z})$ under all 4 data settings, and compared the boxplots to the boxplots in the Tian *et al.* article to complete the validation process of their methods.

Results

The boxplots of Spearman's rank correlation coefficients for both methods are as follows. Figures 1, 2, 3 and 4 represents data settings 1, 2, 3 and 4, respectively. Empty boxplots indicate high-dimensional ($p = 1000$) covariates, whereas filled boxplots indicate low-dimensional ($p = 50$) covariates.

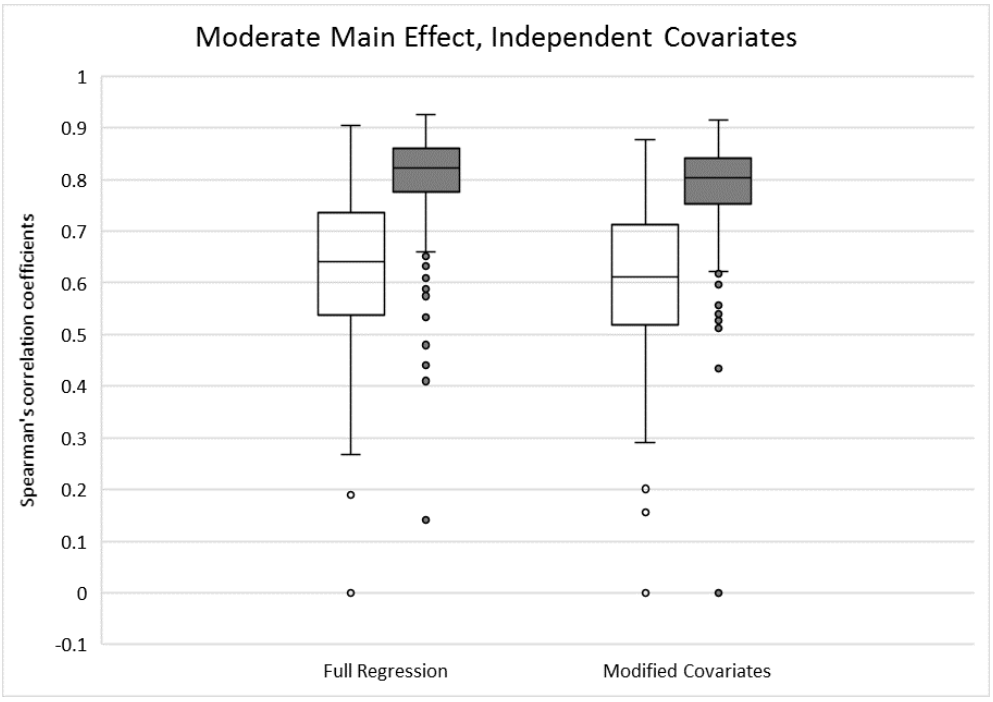

Figure 1: Moderate main effect, independent covariates. Empty and filled boxplots indicate high- and low- dimensional cases respectively.

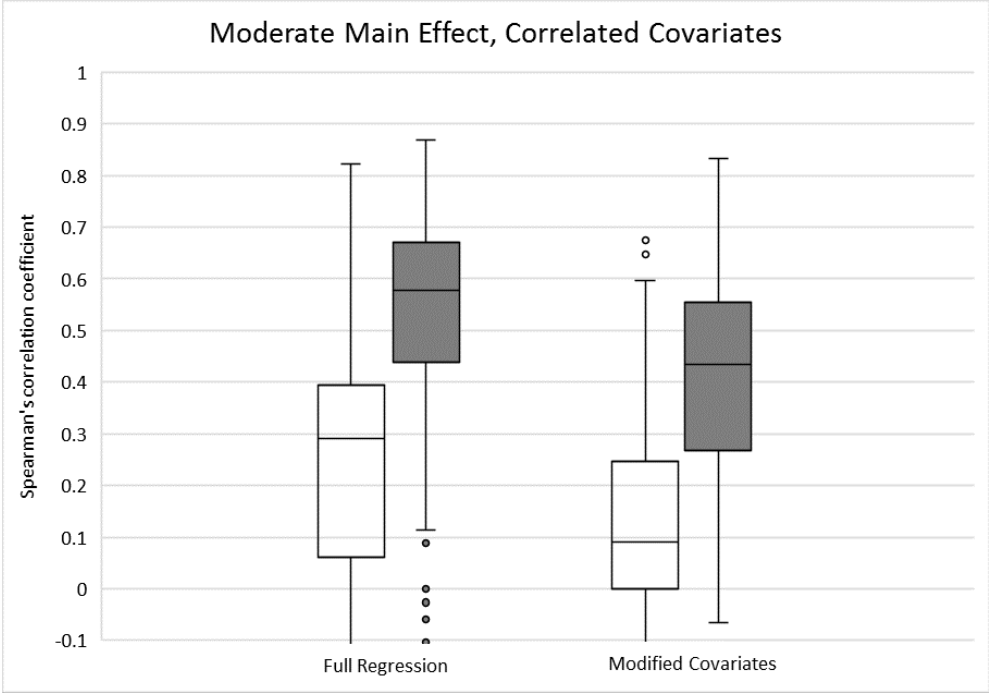

Figure 2: Moderate main effect, correlated covariates. Empty and filled boxplots indicate high- and low-dimensional cases respectively.

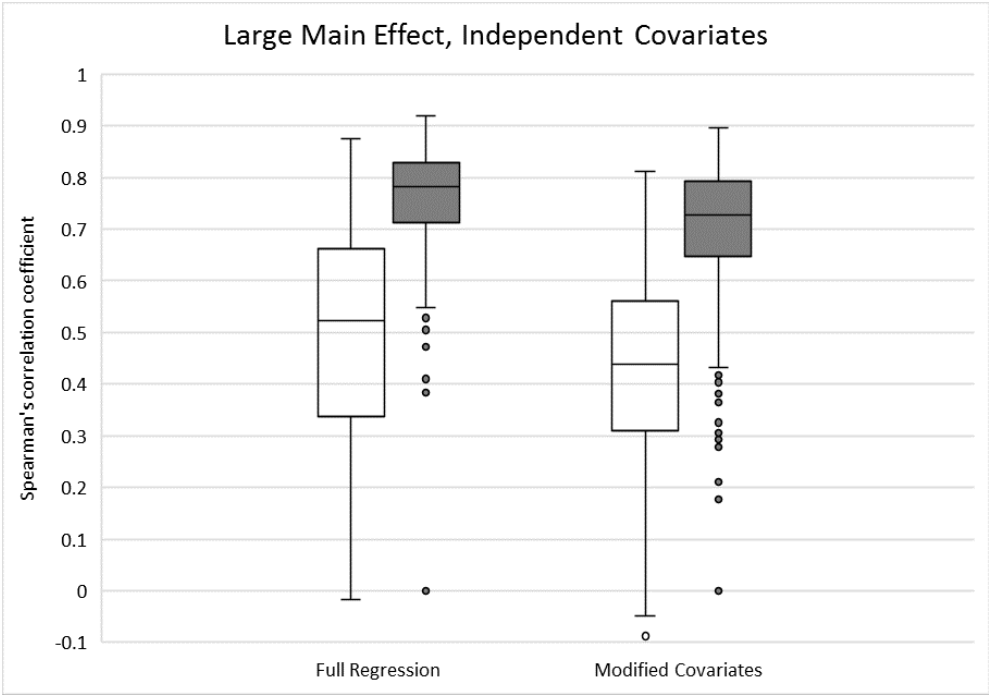

Figure 3: Big main effect, independent covariates. Empty and filled boxplots indicate high- and low-dimensional cases respectively.

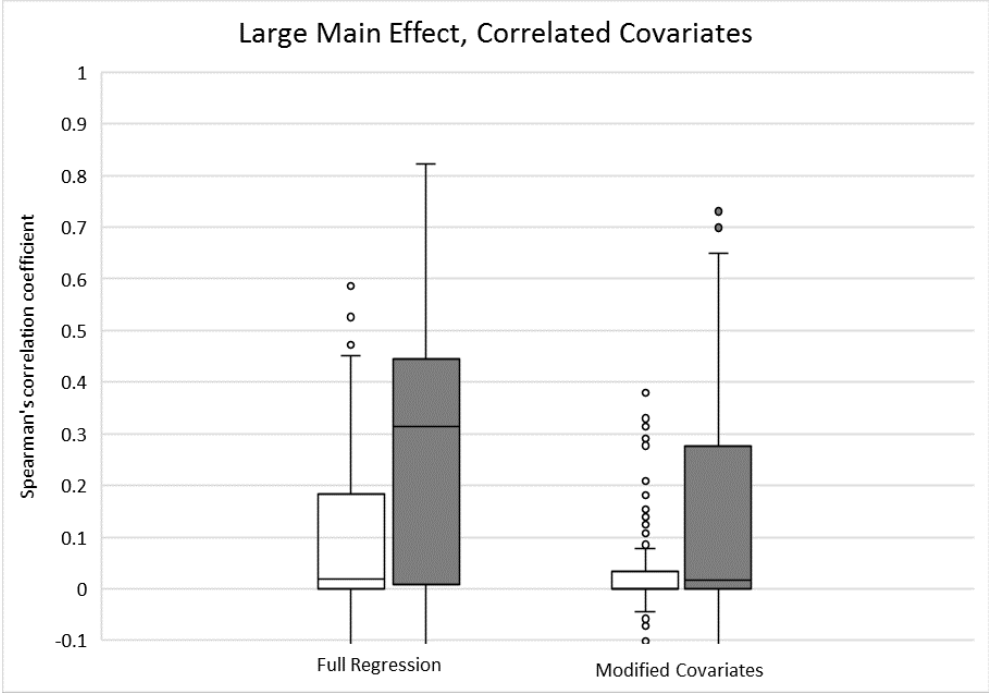

Figure 4: Big main effect, correlated covariates. Empty and filled boxplots indicate high- and low-dimensional cases respectively.

Discussion

In the results section, the author examined both methods in Tian *et al.*'s article, the full regression method and modified covariates method, for estimating treatment effect based on a number of baseline covariates by replicating the numerical studies that they have performed.

For the full regression method, treatments are coded as $T = \pm 1$ and all baseline covariates and covariate/treatment interaction terms are included in a regular linear model. The article by Tian *et al.* demonstrated that this method is inferior to the modified covariates method by showing that Spearman's correlation coefficients between estimated scores and "true" treatment effect under this method are lower than that of modified covariates method. However, using independently generated data, the author was not able to replicate Tian *et al.*'s results for the full regression method despite careful imitation of their data settings; in fact, the opposite result was achieved (see Figures 1-4). It is clear that additional investigation will be required to fully understand the reason of this contradiction.

For the modified covariates method, the general idea is to use $\mathbf{W}(\mathbf{Z}) \cdot T/2$ as new covariates in a regression model to predict the outcome. This provides an efficient approach for constructing a score to estimate the individualized treatment effect as a function of given covariates. Using independently

generated data, the author was able to generate boxplots that are highly similar to those produced by Tian *et al* for the modified covariates method. Thus, this method is likely to be valid, and it provides a casual but theoretically sound tool to assign different treatments in clinical trials with a personalized medicine approach. However, the superiority of this method could not be replicated.

As a limitation, this method is primarily designed for analyzing data from randomized clinical trials. When applied to an observational study, where the covariates and treatment assignment are likely not independent, the constructed score might not apply. According to the boxplots, it is also evident that the performance of the method, especially when applying to correlated high-dimensional data, is limited. Furthermore, although the modified covariates method aims to estimate the individualized treatment effect with casual interpretation, the method is not immune to common problems encountered in high-dimensional data analysis such as multiple testing and over-fitting according to Tian *et al.* Therefore, the method is just an exploratory tool, and it is important to withhold a validation set for verification of the estimated interactions.

References

1. Stratified, personalised or P4 medicine: a new direction for placing the patient at the centre of healthcare and health education (Technical report). Academy of Medical Sciences. May 2015. Retrieved 1 Apr 2016.

2. FACT SHEET: President Obama's Precision Medicine Initiative. The White House. Jan 2015. Retrieved 1 Apr 2016.

3. All About The Human Genome Project (HGP). National Human Genome Research Institute. Oct 2015. Retrieved 1 Apr 2016.

4. French, Benjamin et al. "Statistical Design of Personalized Medicine Interventions: The Clarification of Optimal Anticoagulation through Genetics (COAG) Trial." Trials 11 (2010): 108. PMC. Web. Retrieved 1 Apr 2016.

5. Tian, Lu et al. "A Simple Method for Estimating Interactions between a Treatment and a Large Number of Covariates." Journal of the American Statistical Association 109.508 (2014): 1517-1532. PMC. Web. Retrieved 1 Apr 2016.