Application of Machine Learning methods in baseline covariate adjustment to improve the efficiency of clinical trials

Dimitrios Moliotis

Athens University of Economics and Businness

Department of Statistics

Supervisor: *Professor Vassilios Vasdekis*

September 30, 2020

Overview

Introduction

Randomized clinical trials are considered the gold standard for comparing two or more treatments, regarding an outcome of clinical interest, but they are typically expensive in terms of time and money.

To mitigate this cost, the efficiency of clinical trials can be improved by incorporating baseline information.

The aim of this Master's dissertation is to showcase the use of machine learning algorithms in order to *augment* the simple unadsjusted treatment estimator, by involving a function of baseline covariates.

Introduction to Clinical Trials

Clinical trials are considered one of the best experimental approaches to compare two or more medical treatments. The main goal of a randomized clinical trial, is to assess the effectiveness of an intervention (treatment), as well as to identify possible harms that may arise, as a result of this intervention (Friedman, et al., 2015).

- The analysis of clinical trial data is primarily based on the outcome and the treatment indicator.
- The methods for the analysis of clinical trial data, differ regarding the nature of the outcome variable, such as difference in the means for continuous response, Log-Odds ratio for binary response, Survival Analysis etc.
- Along with them, many baseline covariates may be records for each subject which may have a high association with the outcome.
- In these cases, we can improve the precision and efficiency by *adjusting* our estimators, using this information.

Introduction to Clinical Trials - Key Aspects

A clinical trial refers to an experiment that it is being performed, in the sense that there must be implemented one or more *intervention* techniques (in contrast to observational studies). The key aspects of a typical clinical trial study are:

- The comparison of the effect measure is performed between two or more different groups of subjects, out of which one is the *control* group and the others are the *treatment* groups on which the intervention is being performed.
- A *randomization* process takes place, based on which each one of the participants of the study, are assigned to the different groups of control and treatment.
- The number of participants that will participate in the clinical trial *sample size*.The calculation of the sample size plays an important role in the design of the study, as without a proper sample size, the study lacks the statistical power to detect the effects of the intervention (Friedman, et al., 2015).

Introduction to Clinical Trials - Baseline Information

In clinical trials, baseline refers as the status of a participant patient, before the start of the intervention (Friedman, et al., 2015).

The 4 basic uses of baseline data, as described by (Friedman, et al., 2015) are:

1. Description of trial participants: refers to the need to determine to which population the findings of the clinical trial study apply.
2. Baseline Comparability: refers to the need to evaluate whether the study groups were comparable before the start of the intervention.
3. Controlling for imbalances in the Analysis: refers to the need to "balance out" prognostic factors that are not controlled in the randomization process.
4. Subgrouping: refers to the need to analyze data on the basis of a specific subgroup that may benefit more / less from the intervention.

Statistical Framework

We will consider a typical randomized clinical trial which aims to identify the treatment effect, comparing 2 separate groups (control and treatment). To our context, the point of interest is the analysis of a continuous response outcome. As a result, we base our inference on the difference between the means of each group.

- Let Y denote the continuous response variable of the clinical trial.
- Let T denote the treatment indicator where $T = 1$ refers to treatment, with success probability π, and $T = 0$ refers to the control group with probability $1 - \pi$.
- Let $\mathbf{X}$ denote a vector of baseline covariates, including a baseline measurement of Y.
- Note that due to the randomization process T and $\mathbf{X}$ are independent of each other, denoted by $\mathrm{T} \perp \mathbf{X}$.

Statistical Framework

In that respect, for a clinical trial with n participants, we can express the observed data as $(Y_i, T_i, \mathbf{X}_i)$ and $i = 1, ..., n$ independent and identically distributed across i.

Let θ denote the marginal effect measure of $T = 1$ versus $T = 0$, which is the main interest of the analysis.

Under this framework, a common choice for θ is the difference in the group means noted as

$$\theta = \mu_1 - \mu_0$$

Where

$$\mu_t = E(Y|T = t), t = 0, 1$$

Statistical Framework

The main question that arises is how can we estimate θ in a consistent and efficient way.

An initial thought is to estimate the mean difference

$$\theta_0 = E(Y|T = 1) - E(Y|T = 0)$$

or

$$E(Y|T) = \beta_1 + \beta_2 I(T = 1)$$

where

$$\beta_1 = E(Y|T = 0)$$

, and and the comparison of the difference in means of each treatment is represented directly by β_2 and $I(.)$ is the indicator function, and the treatment effect β_2 is defined as the *unconditional* effect of treatment relative to control.

Statistical Framework

A natural choice for an estimator like described above is the simple *unadjusted* estimator:

$$\tilde{\theta} = \bar{Y}_1 - \bar{Y}_0$$

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

Where $\pi = P(T = 1)$ and , which is a consistent estimate of θ_0
The standard error of the unadjusted estimator can be estimated as follows:

$$SE_{\hat{\theta}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}}$$

Covariate Adjustment Methods

The unadjusted estimator method provides an unbiased and consistent estimate of the treatment effect, although it is a widely accepted method among the international bibliography, many authors have proposed other methods in order to obtain a more efficient estimator for θ_0 by *adjusting* the estimator with the use of baseline covariates.

First, we do not make any assumption about the join distribution of $(Y_i, T_i, \mathbf{X}_i), i = 1, ..., n$ such as normality, equal variances etc.

The only assumption which is the basis for further development is, as we mentioned earlier, that T and $\mathbf{X}$ are independent $\mathrm{T} \perp \mathbf{X}$.

Covariate Adjustment Methods

By using the semiparametric theory, (Tsiatis, 2008) showed that all conistent and asymptotically normal estimators $\hat{\theta}$ can by demonstrated by:

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(T_i, Y_i) + o_p(1) \tag{1}$$

where ψ is the influence function, which does not involve any information about the baseline covariates (independent). In addition, for such an estimator, the influence function can be demonstrated by:

$$\psi(T, Y) = \frac{T(Y - \mu_1)}{\pi} - \frac{(1 - T)(Y - \mu_0)}{1 - \pi} \tag{2}$$

Covariate Adjustment Methods

Bases on the semiparametric theory, (Tsiatis, et al., 2008) and (Tian, et al., 2012) showed than an augmented estimator for θ can be obtained by:

$$\hat{\theta}(\alpha) = \tilde{\theta} - \frac{1}{n}\sum_{i=1}^{n}(T_i - \pi)\alpha(\mathbf{X}_i) \tag{3}$$

where $\alpha(\mathbf{X})$ is any arbitrary function of $\mathbf{X}$, for which $E(\alpha(\mathbf{X}))^2 < \infty$

In this context, one can view $\alpha(\mathbf{X})$ as *augmenting* the simple unadjusted estimator. In addition, due to randomization (Tsiatis et at., 2008) showed in the Appendix that the augmentation term converges in probability to zero, and as a result, the augmented estimator $\hat{\theta}(\alpha)$ is consistent for θ for any function of $\alpha(\mathbf{X})$

Covariate Adjustment - Defining the Optimal Estimator

Having defined the form of an augmented estimator, the next step is to find an *optimal* augmented estimator, where the optimality can be expressed as the estimator with the *smaller variance*.

Among all the estimators which are estimators which are equal or asymptotically equivalent to (3), the estimator with the smallest variance can be obtained when $\alpha(\mathbf{X}_i)$ is assumed to be a linear function like:

$$\alpha^{(t)}(\mathbf{X}_i) = E(Y_i | T_i = t, \mathbf{X}_i), t = 0, 1 \tag{4}$$

In this context, the optimal estimator can be derived from the true regression relationship of Y on X for each treatment

Covariate Adjustment - Learning the Optimal Estimator

Having defined the form of the optimal estimator, a logical question that arises is how can we estimate the optimal function α_{opt}.

As (Zhang & Ma, 2019) pointed, in reality, "the optimal function is unknown and must be estimated from the data", and showed that under some mild regularity conditions:

$$\sqrt{n}(\hat{\theta}(\hat{\alpha}) - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\psi(T_i, Y_i) - (T_i - \pi)\alpha^*(\mathbf{X}_i)\} + o_p(1) \qquad (5)$$

Where $\hat{\alpha}$ is a generic estimator with probability limit α^*, althoughα^* is not equal to α_{opt}. Based on this asymptotic result, one can find different ways to estimate $\hat{\alpha}$.

Covariate Adjustment - Learning the Optimal Estimator

Following the previous results (Zhang & Ma, 2019) suggested that we obtain $\hat{\alpha}$, by using the regression expression:

$$\alpha(\mathbf{X}) = \eta(1, \mathbf{x}) - \eta(0, \mathbf{x}) \tag{6}$$

Where

$$\eta(t, \mathbf{x}) = E\{\psi(T, Y)|T = t, \mathbf{X} = \mathbf{x}\} \tag{7}$$

Due to the fact that this regression function is complicated as the $\psi(T, Y)$ is not fully observed and depends on unknown parameters (μ_1 and μ_0). One possible solution to this problem, is to obtain an estimate $\hat{\psi}(T, Y)$ using the empirical estimates of (μ_1 and μ_0).

Covariate Adjustment - Learning the Optimal Estimator

In order to estimate $\hat{\eta}(t, \mathbf{x})$, we can use various machine learning algorithms, treating $(T_i, \mathbf{X}_i)$ as input, and $\hat{\psi}(T, Y)$ as the response (Zhang & Ma, 2019)

Once $\hat{\eta}(t, \mathbf{x})$ is obtained, the corresponding estimate of the augmentation term is

$$\hat{\alpha}(\mathbf{X}) = \hat{\eta}(1, \mathbf{x}) - \hat{\eta}(0, \mathbf{x}) \tag{8}$$

Covariate Adjustment - Asymptotic Variance

Based on the above formulation, the asymptotic variance of the augmented estimator $\hat{\theta}(\alpha)$ can be epressed as:

$$E\{\psi(T, Y) - (T - \pi)\alpha(\mathbf{X})\}^2 \tag{9}$$

Having a sample version which follows:

$$\sum_{i=1}^{N}\{\hat{\psi}(T_i, Y_i) - (T_i - \pi)\alpha(\mathbf{X}_i)\}^2 \tag{10}$$

Clinical Trial Study ACTG175

To demonstrate the covariate augmentation methods we described earlier, we used a dataset which is based on the real clinical trial study **ACTG175** (Hammer et al., 1996).

- The aim of the study was to evaluate 4 different treatments in adults with HIV-1, with CD4 cell counts between 200 and 500 per cubic millimeter
- The primary study end point was a $>$50 percent decline in the CD4 cell count, an event indicating progression to the acquired immunodeficiency syndrome (AIDS), or death.
- The response outcome is CD4 cell count at 20 ± 5 weeks (*cd420*).
- Baseline covariates are: CD4, CD8 cell count , age, weight, Karnofksy Performance Scale Index, hemophilia, homosexual activity (homo), race, history of intravenous drug use, gender, antiretroviral history and symptomatic indicator.

Clinical Trial Study ACTG175

The result of the study were:

- Antiretroviral Therapy can improve survival in patients with CD4 cells below 500 per cubic millimeter
- All three treatments were superior to the control (treatment with zidovudine alone), but there weren't any significant differences between the 3.

Based on the above result, and following the example of (Tsiatis et al., 2008), we will consider 2 different treatment groups: zidovudine monotherapy (ZDV) with $n_0 = 532$ patients, and all the other groups combined $n_1 = 1607$ patients ($\pi = 0.75$).

Exploratory Data Analysis ACTG175

The ACTG175 dataset is freely available via the R package *speff2trial*. The dataset consists of 2139 patients, who $n_0 = 532$ patients are assigned in the control group, and $n_1 = 1607$ patients are assigned to all the other treatments combined.

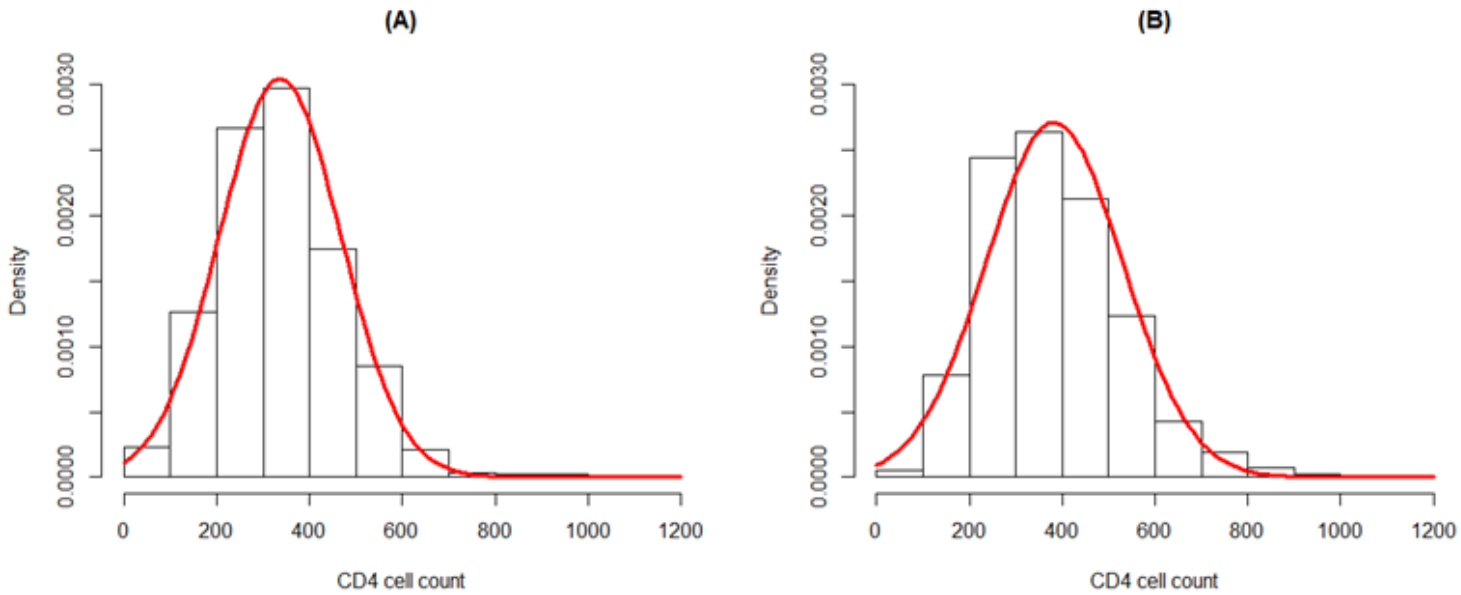
We focus on the analysis of the difference in the means of the outcome variable CD420. Below, we present a table with some summary statistics of the response variable:

Table: Summary Statistics of Response Varialbe

Group	**N**	**Mean**	**SD**	**Median**
All	2139	371.31	144.63	353
Treatment	1607	382.95	147.08	364
Control	532	336.14	130.96	330.5

Exploratory Data Analysis ACTG175

Based on the summary statistics, it is obvious that there exists some skew between the different groups, while, there are evidence of differences between the means.



We can calculate the simple unadjusted estimator as: $\hat{\theta}_{unadjusted} = 46.81$ and $se = 6.76$

Augmentation Algorithms - 1

Below we briefly present the augmentation method settings which we fitted on the ACTG175 dataset.

- LM: Linear Models including all terms
- Backward - 2 : Fit linear models including all terms, squared terms and two-way interactions and then used backward stepwise to select the best models.
- Lasso Regression: Fit lasso regression including all terms using the GLMNET package (10 folds cross-validation).
- Ridge Regression: Fit ridge regression including all terms using the GLMNET package (10 folds cross-validation).

Augmentation Algorithms - 2

Below we briefly present the augmentation method settings which we fitted on the ACTG175 dataset.

- Rpart: Fit recursive partitioning tree models including all terms using the Rpart package.
- Random Forest: Fit random forest models including all terms using the RandomForest package (100 trees, 4 variables to randomly select for each split).
- Super Learner: Fit superlearner models including all terms using the Super Learner package, and include models (LM, Lasso, Ridge, RPart, Random Forest).

Model Fitting

Below we briefly present the model fitting procedure, which is based on (Tsiatis et al 2008), and follows the steps:.

1. Partition the data into the two sets determined by the randomized treatment groups.
2. Based on each of group separately, develop models which try to capture the true relationship of the data.
3. For each separate model, obtain the predicted values using the whole dataset, $\hat{\eta}(t, \mathbf{x})$.
4. Calculate the augmented estimator using equation

$$\hat{\alpha}(\mathbf{X}) = \hat{\eta}(1, \mathbf{x}) - \hat{\eta}(0, \mathbf{x})$$

Numerical Results - Present

Using the various augmentation algorithms we described earlier, we present the table with the numerical results. The Relative Efficiency of each estimator was calculated using:

$$\text{Relative Efficiency} = \frac{(SE\ of\ unadjusted\ estimator)^2}{(SE\ of\ augmented\ estimator)^2}$$

Table: Numerical Results of Augmentation Algorithms on ACTG175 data

Estimator	**Estimate**	**Standard Error**	**Relative Efficiency**
Unadjusted	46.810	6.760	1
LM - All terms	49.818	5.101	1.755
Backward - 2	53.647	4.896	1.906
Lasso – 10 folds – all terms	49.579	5.118	1.744
Ridge – 10 folds – all terms	49.637	5.108	1.751
Rpart – all terms	52.133	4.982	1.84
Random Forest, 4 variables, 100 trees	52.045	3.690	3.354
Super Learner	51.02	4.68	2.07

Numerical Results - Remarks

Based on the above numerical results of the various covariate adjustment techniques, we conclude that:

- All of the proposed methods offer significant efficiency boost over the unadjusted estimator.
- The methods which used linear models (LM, and Backward) despite the being more efficient that the unadjusted estimator, the assumption of the models do not hold, which may result on misleading SE.
- The Random Forest Method, has the higher relative efficiency, but we suspect that the result may be misleading as RF tends to product overfitted models.
- The real world application of such methods means that we can attain the same level of precision, while reducing the sample size, even by reducing it in half ($RE \simeq 2$).

Simulation Study - Setup

In order to empirically evaluate and access the different covariate adjustment methods, we performed a simulation study.

- Our data simulation generating algorithm is based on the bootstrap technique (Efron Tibshirani, 1986).
- We took as the support set of the simulation to be the collection of observed values $(\mathbf{X}, Y)$.
- We consider a population of patients of $N = 2139$ with potential outcomes $Y(0) \equiv Y(1)$ and $\mathbf{X}$ the vector of baseline covariates with a discrete joint probability distribution F as

$$F = \sum_{i=1}^{N} \frac{p_i}{\delta_i}$$

where $p_i \geq 0$ and $\sum_{i=1}^{N} p_i = 1$

Simulation Study - Setup

- We consider a random sample of size $n_{Obs} = 400$ based on the above population.
- We denote T_i as the randomized treatment indicator for the ith subject where

$$T \sim Bernoulli(\pi)$$

and T $\perp$ **X**

- Our analysis will be based on $n_{Sim} = 1000$ simulations.

Performance Measures - 1

In order to best compare the various agmentation techniques, we will estimate some performance measures, as presented below:

$$Bias = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \tilde{\theta}_i - \theta$$

$$Monte\ Calro\ SE\ of\ Bias = \frac{1}{n_{sim}(n_{sim}-1)} \sum_{i=1}^{n_{sim}} \left(\tilde{\theta}_i - \bar{\theta}\right)^2$$

$$Empirical\ SE = \sqrt{\frac{1}{n_{sim}-1} \sum_{i=1}^{n_{sim}} \left(\tilde{\theta}_i - \bar{\theta}\right)^2}$$

Performance Measures - 2

$$Simple\ Bootstrap\ CI\ = \quad \left(\tilde{\theta} - Z_{1-a/2}\, se_B(\tilde{\theta}), \quad \tilde{\theta} - Z_{a/2}\, se_B(\tilde{\theta})\right)$$

$$Coverage\ = \quad \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1\left(\tilde{\theta}_{low,i} \leq \theta \leq \tilde{\theta}_{upp,i}\right)^2$$

Numerical Results - Present

We present the numerical results of the various augmentation algorithms we described earlier, which were fitted in the simulated datasets. To compare the methods, our main points of interest are the **Bias** of each method, the Relative Efficiency **RE**, and the coverage probability **CP** compared to the unadjusted estimator.

Table: Numerical Result of Simulation Data

Estimator	**Bias**	**SE of Bias**	**Emp SE**	**RE**	**CP**
Unadjusted	0.122	0.442	13.997	1.000	0.948
LM - All terms	3.025	0.349	11.046	1.606	0.936
Lasso – 10 folds – all terms	2.651	0.347	10.982	1.624	0.939
Ridge – 10 folds – all terms	2.750	0.349	11.056	1.603	0.933
Rpart – all terms	4.190	0.359	11.364	1.517	0.927
Random Forest, mtry = 4, ntree= 100	3.781	0.347	10.999	1.619	0.938
Super Learner	3.569	0.484	10.822	1.672	0.938

Numerical Results - Remarks

- All of the augmentation techniques are at least 1.5 times more efficient that the simple unadjusted estimator.
- All of the proposed methods produce biased estimates when compared to the unadjusted estimator.
- The Super Learner algorithms produces the most efficient estimator.
- Possible solutions to reduce the bias of the estimates, is to further increase the number of simulations, as well as to incorporate another external cross validation procedure, as showcased by (Zhang & Ma, 2019).

Conclusion - Remarks A

- We reviewed some covariate adjustment techniques which aim to improve the efficiency of a randomized clinical trial, by incorporating auxiliary baseline covariates.
- The various methods include linear models (including stepwise procedures), as well as machine learning algorithms like recursive partitioning, Random Forest and the Super Learner algorithm.
- Based on the application of these techniques on a real example dataset, we saw that that they offered a great efficiency boost.
- We examined the performance of these methods by performing a simulation analysis.

Conclusion - Remarks B

- The two best methods on the real data, were Random Forest and Super Learner, which offered a relative efficiency > 2.
- The simulation analysis that we could achieve at least 1.5 gain in efficiency by leveraging baseline covariate information
- The Simulation analysis suggested that all of these techniques suffer from high bias.

References 1

- Armitage, P., 1982. The Role of Randomization in Clinical Trials. Statistics in Medicine , Volume 1, pp. 345-352.
- Efron, B. Tibshirani, R., 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science, 1(1), pp. 54-77.
- Friedman, L. M. et al., 2015. Fundamentals of Clinical Trials. 5th ed. Switzerland: Springer International Publishing Switzerland.
- Hammer, S. et al., 1996. A Trial Comparing Nucleoside Monotherapy with Combination Therapy in HIV-Infected Adults with CD4 Cell Counts from 200 to 500 per Cubic Millimeter. The New England Journal of Medicine, 335(15), pp. 1081-1090.
- Harrell, F. J. E., 2015. Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. s.l.:Springer, Cham.

References 2

- Hastie, T., Tibshirani, R. Friedman, J., 2009. The Elements of Statistical Learning Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, New York,.
- Lachin, J. M., 1981. Introduction to Sample Size Determination and Power Analysis for Clinical trials. Controlled Clinical Trials, 2(2), pp. 93-113.
- Leon, S., Tsiatis, A. Davidian, M., 2003. Semiparametric estimation of treatment effect in a pretest - posttest study. Biometrics, 59(4), pp. 1046 - 1055.
- Mark, J., van der Laan, Polley, E. C. Hubbard, A. E., 2007. Super Learner. Statistical Applications in Genetics and Molecular Biology, 6(1).
- Moher, D., Dulberg, C. Wells, G., 1994. Statistical power, sample size, and their reporting in randomized controlled trials.. JAMA, 272(2), pp. 122-124.

References 3

- Morris, T. P., White, I. R. Crowther, M. J., 2019. Using simulation studies to evaluate statistical methods. Statistics in Medicine, pp. 2074-2102.
- Pocock, S. J., Assmann, S. E., Enos, L. E. Kasten, L. E., 2002. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statistics in Medicine, 21(19), pp. 2917-2930.
- Robert, C. P. Casella, G., 2010. Introduction Monte Carlo Methods with R. New York: Springer-Verlag New York.
- Tian, L., Cai, T., Zhao, L. Wei, L.-J., 2012. On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial. Biostatistics, 13(2), pp. 256-273.

References 4

- Tsiatis, A. A., Davidian, M., Zhang, M. Lu, X., 2008. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. Statist Med, 27(23), pp. 4658-4677.
- White, I. R., 2010. simsum: Analyses of simulation studies including Monte Carlo error. The Stata Journal, 10(3), pp. 369-385.
- Yang, L. Tsiatis, A. A., 2001. Efficiency Study of Estimators for a Treatment Effect in a Pretest-Posttest Tri. The American Statistician, 55(4), pp. 314-321.
- Zhang, M., Tsiatis, A. A. Davidian, M., 2008. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. Biometrics, 64(3), pp. 707-715.
- Zhang, Z. Ma, S., 2019. Machine learning methods for leveraging baseline covariate information to improve the efficiency of clinical trials. Statistics in Medicine, pp. 1703-1714.

Thank you for your attention !