Estimation of average treatment effect for ACIC Data Challenge

Wanwan Xu

Univeristy of Connecticut, Department of Statistics

Abstract

This is a data-driven project based on ACIC 2019 data challenge, the main objective is to estimate the population average treatment effect (ATE). The performance of regression estimator, inverse propensity weighted estimator, augmented inverse propensity weighted estimator and calibration estimator is compared on test datasets as well as some theoretical properties. The final result for all 3,000 sets of training data is also presented.

Keywords: average treatment effect, counterfactual, causal inference

1 Introduction

Treatment effect has always been of interest, as it quantitatively show the causal effect of a binary variable on an outcome variable. This 0-1 variable, also called treatment, is now used much more generally than a medical procedure or experimental drug. In economics, the causal effect of a subsidized training program [1] is widely analyzed, so are the effects of active labor market programs such as job search assistance or classroom teaching program [8]. Since treatment effect is defined as the difference between two treatment groups, while in reality, one unit or one observation is only related to one of the treatment assignment, it's impossible to measure causal effects at the individual level, therefore we focus on the expectation, i.e. average treatment effect(ATE). In experimental study, randomized trial is a gold standard to identify average treatment effects. But in order to achieve randomization, the balance distributions of subject characteristics across groups is required, so that groups are similar except for the treatments, it may be infeasible, or even unethical, to conduct in practice. On the other hand, observational studies are common in economy, social science, and public health, where the participation of intervention is only observed rather than controlled by designers. A typical concern for inferring causality in an observational study is confounding, treatment exposure may e

associated with covariates that are also associated with potential response. For example, experiment individual characteristics such as demographic factors can be related to both the treatment selection and the outcome of interest.

In order to perform unbiased comparison despite the confounding factors, the inferences with causal interpretation need to be adjusted. There are three broad classes of strategies, outcome regression, propensity score estimation and nonparametric estimation. For outcome regression, the model of outcome given covariates need to be specified, then it can be used for predicting unobserved potential outcomes and estimating ATE [10]. It follows similar procedure of calculating counter factuals. For propensity score estimation, Rosenbaum and Rubin [13] proposed several properties that facilitate causal inferences. A popular method for estimating the causal difference of two treatment means is to stratifying individuals based on estimated propensity scores, then use the average of within stratum effects as ATE [14]. An alternative approach is to construct weights for individual observations via using estimated propensity scores. individual observations [12]. Above two classes require specification of propensity score model or outcome model or both, some nonparametric estimators are then developed to provide valid inference in large samples without relying on parametric assumptions in the intermediate steps of estimation [4]. Chan [3] applied the class of survey calibration estimators on the estimation of average treatment effects, the authors also showed that a globally efficient estimator can be adapted from it.

This report focuses on applying several estimation methods described above to ACIC data challage low dimension track datasets. The report is organized as follows. In Section 2, we introduce the data structure and explanation. The methods applied is briefly introduced in Section 3, the comparison based on test dataset is presented in Section 4 as well as the results for training dataset. Some final remarks and future directions are given in Section 5.

2 Dataset

2.1 Introduction

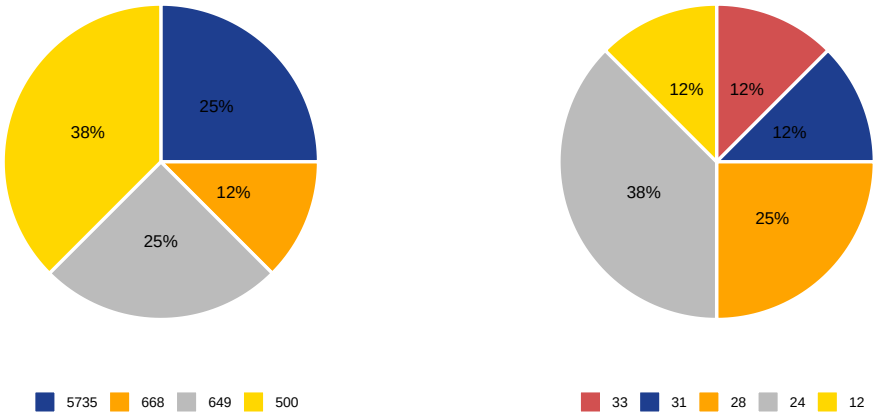
Two tracks of data (low dimension and high dimension) are included in the ACIC data challenge, each track have 3200 csv files drawn from 32 unique data generating processes (DPG). In this report, we focus only on low dimensional track. Within each csv file, there are:

- response variable $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$, half are continuous while the other half is discrete;
- binary treatment indicator $\mathbf{A} = \{A_1, A_2, \ldots, A_n\}$;
- pre-treatment covariates $\mathbf{V} = \{V_1, \ldots, V_p\}$, which includes the mixture of continuous, binary, and categorical variables.

In order to understand the data structure better, we made an example of the first dataset into clinical trial framework. Suppose a trial is designed to analyze whether a

specific drug can lower blood pressure or not. Data from $n = 668$ patients are collected, some of them are assigned to take the drug ($A_i = 1$) and others took placebo ($A_i = 0$). Their pre-treatment vital is recorded as well. For example race and gender are saved as catogorical, certain disease history is saved as binary, and patients' height, weight, temperature are saved as continuous. After the treatment, their blood pressure change is recorded in $\mathbf{Y}$. Each of the dataset can be explained by such similar "stories", the goal is to find an appropriate model that can perform good on all the datasets.

Aside from the mixture of data type, number of covariates and number of observations may also be different for each worksheet as shown in Figure 1.



(a) number of observations (b) number of covaraibles

Figure 1: Frequency plot low track dataset

The description site of data challenge also provided some more information. For instance, there are no mediators among covariates, that is we only need to consider direct effect from covariates. Also, all the covaraites also include all the cofounders, we don't need to consider hidden variables. However, selection variables may still exists, given some of the covaraites may be related to the treatment factor $\mathbf{A}$.

2.2 Objectives

The measure of estimation is the population average treatment effect (ATE) and 95% confidence interval for each dataset. For individual $i \in \{1, 2, \ldots, n\}$, denote Y_{0i} as the outcome variable for individual i if he/she is not treated, y_{1i} as the value of the outcome variable for individual i if he/she is treated. The treatment effect for individual i is given by $Y_{1i} - Y_{0i} = \beta_i$. In the general case, there is no reason to expect this effect to be constant across individuals. The average treatment effects is given by

$$\text{ATE} = E[Y_{1i} - Y_{0i}].$$

where the expectation is with respect to the distribution of covariates. However, in observational data, we can only observe one of the two for each individual, thus estimation techniques are needed. Under this circumstances, one of the y_{0i}, y_{1i} will be replaced by counterfactual output. Since the causal assumptions of consistency and strong ignorability are guaranteed, the target statistical estimand can be rewritten as:

$$\text{ATE} = E[E(\mathbf{Y}|\mathbf{A}=1,\mathbf{V}) - E(\mathbf{Y}|\mathbf{A}=0,\mathbf{V})]$$

2.3 Difficulties

Apart from the complicated structure of the data and requirement for generality of applied model, there are a few more challenges for this specific objective.

- Non-linearity of the response surface. Not only does the distribution of covaraites $\mathbf{V}$ is unknown, the model may also be non-linear.
- Treatment effect heterogeneity. There may exist differences across subjects, going back to the clinical trail example, different patients may have different response level for the same treatment.
- Varying proportion of true confounders among the observed covariates. Even though we are provided with the fact that all the confounders are measured and included in the covariste set, the proportion or location of these confounders still remain unknown.

3 Method

3.1 Notation

Following the notation of dataset, use $i \in \{1, 2, \ldots, n\}$ to denote ith observation, Y_i as observed outcome, $A_i \in \{0, 1\}$ as treatment assignment, $\mathbf{V} = \{V_1, V_2, \ldots, V_p\}$ as the set of covaraites, $Y_i(1)$ as potential outcome if $A_i = 0$ and $Y_i(0)$ as potential outcome if $A_i = 1$. Furthermore, we use $\pi(\mathbf{V}) = P(\mathbf{A} = 1|\mathbf{V})$ to denote propensity score.

There are a few common assumptions for the methods discussed below: (1)all units Y_i are random sampled; (2)the treatment value is stable such that potential outcomes $Y_i(1), Y_i(0)$ are completely determined and the observed outcome will be equal to the potential outcome corresponding to the assigned treatment: $Y_i = A_iY_i(1)+(1-A_i)Y_i(0)$; (3) strong ignorability $\{\mathbf{Y}(0), \mathbf{Y}(1)\} \perp\!\!\!\perp \mathbf{A}|\mathbf{V}$; (4)plausible range of propensity score function $0 < \pi(\mathbf{V}) < 1$.

3.2 Regression Estimator (REG)

Lots of traditional causal estimation relies on the formulation of a regression model for the outcome variable $\mathbf{Y}$. In other words, the focus is on the estimation of $E[\mathbf{Y}|\mathbf{A}, \mathbf{V}]$.

Given ATE is defined with respect to the distribution of covariates, the empirical distribution will be an easy estimate of $F_{\mathbf{V}}$. Thus the regression estimate can be written as:

$$\widehat{\text{ATE}}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\{E(\mathbf{Y}|\mathbf{A}=1,V_i) - E(\mathbf{Y}|\mathbf{A}=0,V_i)\} \tag{1}$$

where the conditional expectation functions can be estimated using any consistent estimator. Options include ordinary least squares, generalized linear models, generalized additive models (GAMs), local regression, kernel regression, etc [11].

Even though the regression estimator for ATE is easy to understand, the shortcoming is also obvious. It depend heavily on the estimation method and the choice of regression model, especially with high dimensional $\mathbf{V}$, it may be difficult to estimate both regression functions over the full range of Z. It has been shown when the observed values of $\mathbf{V}$ are not similar for the treatment and the control groups, then one of the conditional expectation functions will often be poorly estimated [7]. However, in our dataset, this problem doesn't seem to appear, so the regression estimator is also included in the comparison.

3.3 Inverse Propensity Weighted Estimator (IPW)

Another widely applied method for estimating ATE relies on a model for treatment assignment instead a regression model for the outcome. Suppose the true probability of assigning treatment were known, this could be used to define propensity scores, and further be used for matching or weighting covariates. A well-known weighting estimator is the IPW estimator:

$$\widehat{\text{ATE}}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_iY_i}{\hat{\pi}(\mathbf{V}_i)} - \frac{(1-A_i)Y_i}{1-\hat{\pi}(\mathbf{V}_i)}\right\} \tag{2}$$

where $\hat{\pi}(\mathbf{V}_i)$ is the estimated propensity score, that is the estimated conditional probability of treatment given $\mathbf{V}_i$. As stated by Glynn in [6], if the propensity score estimate is consistent, the IPW estimator is consistent as well. However, as can be seen in the formula, observations with $A_i = 1$ and $\hat{\pi}(\mathbf{V}_i)$ close to 0 may contribute extremely to the estimate, similarly to the observations with the untreated group of $A_i = 0$ and $\hat{\pi}(\mathbf{V}_i)$ close to 1. In some cases, these extreme contributions can produce implausible estimate of ATE (greater than 1). Therefore here we apply one of the improvements to the basic IPW estimator proposed by [6]. The authors re-normalize the weights so that they sum to one:

$$\widehat{\text{ATE}}_{IPW^*} = \left\{\sum_{i=1}^{n}\frac{A_i}{\hat{\pi}(\mathbf{V}_i)}\right\}^{-1}\sum_{i=1}^{n}\frac{A_iY_i}{\hat{\pi}(\mathbf{V}_i)} - \left\{\sum_{i=1}^{n}\frac{1-A_i}{1-\hat{\pi}(\mathbf{V}_i)}\right\}^{-1}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-\hat{\pi}(\mathbf{V}_i)} \tag{3}$$

3.4 Augmented Inverse Propensity Weighted Estimator (AIPW)

Based on the previous work of IPW estimation, one improvement including fully utilizing the information in the conditioning set $\mathbf{V}$ is AIPW [5]. Aside from the information about treatment assigning probability, $\mathbf{V}$ also carry the predictive information about the response variable $\mathbf{Y}$:

$$\widehat{\mathrm{ATE}}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\left[\frac{A_iY_i}{\hat{\pi}(\mathbf{V}_i)} - \frac{(1-A_i)Y_i}{1-\hat{\pi}(\mathbf{V}_i)}\right] - \frac{(A_i-\hat{\pi}(\mathbf{V}_i))}{\hat{\pi}(\mathbf{V}_i)(1-\hat{\pi}(\mathbf{V}_i))} \\ [(1-\hat{\pi}(\mathbf{V}_i))\hat{E}(Y_i|A_i=1,\mathbf{V}_i) + \hat{\pi}(\mathbf{V}_i)\hat{E}(Y_i|A_i=0,\mathbf{V}_i)]\Bigg\} \tag{4}$$

Similar to IPW estimand, AIPW also needs to specify a binary regression model for the propensity score, and specify a regression model for the outcome variable. However, as summarized by the authors, AIPW has so called "double robustness" property. Simply speaking, the estimator remains consistent for the ATE if either the propensity score model or the outcome regression is misspecified but the other is properly specified. AIPW estimator are also shown to be asymptotically normally distributed and valid large-sample standard errors can be derived through the theory of M-estimation. One advantage of AIPW method in our data application is it provides multiple ways of calculating sample variance, including an empirical sandwich estimator proposed by Lunceford [9], alternative large-sample results as well as the bootstrap.

3.5 Calibration Estimator (CAL)

Previous few estimate methods both require estimation of a propensity score function, and an outcome regression function, IPW and AIPW is globally semiparametric efficient when a sieve maximum likelihood propensity score estimator is used. Chan [3] proposed a wide class calibration weights to attain the moments of observed covariates among the treated, the control, and the combined group. Furthermore, these empirical weights can be applied for globally efficient non-parametric inference of ATE. Chan and Yam reformulated the problem into optimization framework:

$$\begin{aligned} &\min \sum_{i=1}^{n} A_i D(np_i, 1) \quad s.t. \quad \sum_{i=1}^{n} A_i p_i u_K(\mathbf{V}_i) = \frac{1}{n}\sum_{i=1}^{n} u_K(\mathbf{V}_i), \\ &\min \sum_{i=1}^{n} (1-A_i) D(nq_i, 1) \quad s.t. \quad \sum_{i=1}^{n} (1-A_i) q_i u_K(\mathbf{V}_i) = \frac{1}{n}\sum_{i=1}^{n} u_K(\mathbf{V}_i) \end{aligned} \tag{5}$$

where u_K is a $K(N)$-dimensional function of $\mathbf{V}$, D is chosen distance measure. The proposed empirical balancing estimator for ATE is:

$$\widehat{ATE}_{CAL} = \sum_{i=1}^{n}\{A_i\hat{p}_K(\mathbf{V}_i)Y_i - (1-A_i)\hat{q}_K(\mathbf{V}_i)Y_i\} \tag{6}$$

where $\hat{p}_K, \hat{q}_K$ are the dual solutions for above equation. One thing worth noticing here is the usage of concave function $\rho(v)$ when solving the dual optimization problem. In [3], four different functions are used, including exponential tilting (ET) $\rho(v) = -e^{-v}$; empirical likelihood (EL) $\rho(v) = log(1+v)$; quardratic (Q) $\rho(v) = -(1-v^2)/2$ and inverse logistic (IL) $\rho(v) = v - e^{-v}$. In our data application, we apply the ET version, which is also shown to be the "best" selected by simulation part in the paper.

4 Result

5 Discussion

References

[1] Ashenfelter, Orley. "Estimating the effect of training programs on earnings." The Review of Economics and Statistics (1978): 47-57.

[2] Chambers, John M., and Trevor J. Hastie, eds. Statistical models in S. Vol. 251. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1992.

[3] Chan, Kwun Chuen Gary, Sheung Chi Phillip Yam, and Zheng Zhang. "Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78.3 (2016): 673-700.

[4] Chen, Xiaohong, Han Hong, and Alessandro Tarozzi. "Semiparametric efficiency in GMM models with auxiliary data." The Annals of Statistics 36.2 (2008): 808-843.

[5] Glynn, Adam N., and Kevin M. Quinn. "An introduction to the augmented inverse propensity weighted estimator." Political analysis 18.1 (2010): 36-56.

[6] Imbens, Guido W. "Nonparametric estimation of average treatment effects under exogeneity: A review." Review of Economics and statistics 86.1 (2004): 4-29.

[7] King, Gary, and Langche Zeng. "The dangers of extreme counterfactuals." Political Analysis 14.2 (2006): 131-159.

[8] LaLonde, Robert J. "Evaluating the econometric evaluations of training programs with experimental data." The American economic review (1986): 604-620.

[9] Lunceford, Jared K., and Marie Davidian. "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study." Statistics in medicine 23.19 (2004): 2937-2960.

[10] Oaxaca, Ronald. "Male-female wage differentials in urban labor markets." International economic review (1973): 693-709.

[11] Pearl, Judea. "Causality: models, reasoning, and inference." IIE Transactions 34.6 (2002): 583-589.

[12] Robins, James M., Miguel Angel Hernan, and Babette Brumback. "Marginal structural models and causal inference in epidemiology." (2000): 550-560.

[13] Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." Biometrika 70.1 (1983): 41-55.

[14] Rosenbaum, Paul R., and Donald B. Rubin. "Reducing bias in observational studies using subclassification on the propensity score." Journal of the American statistical Association 79.387 (1984): 516-524.