Ching Fung Chow 3303229 Computational Statistics term paper

Topic: Regression Discontinuity Design versus Causal Forest: religiousness and cash transfer payment as an example

1. Introduction

Regression discontinuity design (RDD) is employed by social scientists to estimate the effect of a treatment in a nonexperimental setting. It is a classical tool first invented by Thistlethwaite and Campbell (1960). It became widely adopted by economists. As time goes by, economists invent new tools in their toolbox for hunting the treatment effect. Wager and Athey (2018) developed the method of causal forest (CF) based on the idea of tree methods. This paper will compare RRD with CF, showing their strengths and weaknesses. Building on their relative performance, we can provide practical guidelines for the usage of CF versus RDD when the assumptions of the methods are violated.

We will support our analysis by simulating Buser (2015) study, in which the effect of income on the level of religiousness is estimated. Our conclusion, supported by simulation results, is that RDD works better when the treatment is deterministic, the sample size is small or omitted variable. On the other hand, CF works better when there are unclear polynomial interactions between the treatment and covariates.

2. Research question: The Effect of Income on Religiousness

Buser (2015) studies the effect of a cash transfer program on religiousness in Ecuador in 2009, by exploiting the eligibility cutoff for the cash transfer. It is found that income is positively related to religiousness. The cash transfer income with one's attendance to church services. Also, families with higher income are more probed to be members in the evangelical community rather than of the mainstream Catholic Church.

We will replicate the dataset for our simulation study. Unfortunately, we do not have access to the original raw dataset. So, we have to modify the economics model in order to run simulations. Despite the inability to directly compare our results to the author's, our simulations and conclusions still remain logically valid. In our simulations, the outcome of interest is church attendance and the treatment is a binary variable indicating treatment of receiving cash transfer. They are generated using our models. Covariates such as expenditure, schooling level and household size are simulated from the dataset the author uploaded. We simulate the original dataset by computing the mean and the covariance of the covariates and then produce samples through the multivariate normal distribution. Treatment then will be assigned based on the simulated values of expenditure.

3. Set up of the research problem

Our research problem can be expressed in a simple framework proposed by Lee and Lemieux (2010):

$$Y = W\delta_1 + D\tau + \mu$$
$$D = I[X \geq c]$$
$$X = W\delta_2 + \nu$$

When $\delta_2 = 0$, meaning that covariates do not affect the probability of taking treatment D. The assignment of treatment would be only motivated by ν and is random.

When $\delta_2 \neq 0$, meaning that covariates will affect the probability of taking treatment D, RDD estimates the treatment effect by making use of the cutoff, C. In a *sharp* RDD, the treatment is deterministically switched on when $X \geq c$. X, according to which the treatment will be switched on if its value passed the cutoff, is our running variable.

If $\nu \neq 0$, then there will be some randomness deciding the treatment. We can think of it as a lucky draw process deciding the recipient of the transfer. A fuzzy RDD will be used. Treatment is assumed to be randomly assigned when ***X is around the cutoff***. *So, observations around the cutoff received treatment randomly.*

There can be two ways of doing RDD: nonparametric and parametric. The nonparametric simply limit the observations to the ones around the cutoff in the local linear regression, sacrificing a larger sample size. The parametric one includes all observations while assuming a parametric functional form of the running variable, facing the risk of estimation being sensitive to the assumed order of the polynomial[1]. In our paper, we focus on the parametric RDD.

4. Causal forest

CF is developed based on Classification And Regression Tree (CART). A tree is grown with the purpose of **minimizing** mean squared error **(MSE)** of the outcome variable by each partition within the covariate space, where mutually exclusive regions are split. The minimization problem can be expressed as

$$\min_{j,s} = \sum_{x_i \in R_1(j,s)} (y_i - \bar{y}_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - \bar{y}_2)^2$$

[1] As we will reveal in section 5.2

In each final partition, data points within are supposed to have similar values for Y. Data will be partitioned will an objective to minimize the sum MSE by choosing the mutually exclusive regions *j* and *s*. Data points end up region R will have an expected value of Y = $\overline{\mathrm{Y}}$. A forest averages out B number of trees, each of which is grown by randomly subsetting the dataset. The prediction of a new data point will be the average of predictions given by B trees. However, a normal tree/ forest does not help us to estimate the treatment effect. A tree is grown with the purpose of minimizing the square difference between Ys and $\overline{\mathrm{Y}}$, while treatment effect is calculated by E[Y | W=1] - E[Y | W=0] Instead, when estimating the treatment effect, we estimated the conditional average treatment effect in each final partition (or "leaf") L as:

$$\widehat{\tau_{CF}}(\mathrm{x}) = \frac{1}{i\colon \mathrm{W}_i = 1, \mathrm{x}_i \in \mathrm{L}} \sum_{\mathrm{i}\colon \mathrm{W}_i=1, \mathrm{x}_i \in \mathrm{L}} \mathrm{y}_i \quad - \quad \frac{1}{i\colon \mathrm{W}_i = 0, \mathrm{x}_i \in \mathrm{L}} \sum_{i\colon \mathrm{W}_i=0, \mathrm{x}_i \in \mathrm{L}} \mathrm{y}_i$$

Athey and Imbens (2016) prove that maximizing the variance of $\widehat{\tau_{CF}}(\mathrm{x})$ is similar with minimizing MSE. Therefore, if a tree is grown for predicting treatment effects, its splitting rule is to choose mutually exclusive regions that maximizes the variance of $\widehat{\tau_{CF}}(\mathrm{x})$ of the regions.

Grounding on CARTs, Wager and Athey (2018) proposed causal forest as a way to estimate the treatment effect. A causal forest is featured with the property of *honesty*, which means that for each tree, we split the randomly subsample data into two separate subset: training data S^{tr} and validation data $\mathrm{S}^{\mathrm{est}}$. S^{tr} is used to estimate the model. Then we put $\mathrm{S}^{\mathrm{est}}$ data into the model and estimate the treatment effect. Finally we average out the treatment effect obtained in each tree. Honesty estimation "prevents overfitting and bias, but comes at a cost of increased variance, as estimates are produced with a smaller sample" (Gulen, Jens and Page, 2020, p.15) .

Two assumptions are needed to assert causality in CF: *unconfoundedness* and *overlapness*. Unconfoundedness means that in the final partition, covariates are controlled for and treatment is randomly assigned. The only variable that will affect differences in the outcome is the treatment. Overlapness means that a sufficient amount of treated and control samples in the final partitions when calculating the treatment effect. In our setup, only observations around the cutoff will have a propensity score not close to 0 or 1, meaning that data far away from the cutoff will end up in partitions where there is only a very small amount or no "counterfactual" observations. In order to correct this problem, we will estimate the average treatment effect with an overlap correction (ATO) = $\widehat{\mathrm{ATO}} = \frac{\sum_{i=1}^{n}(W_i - \widehat{W}_i)(Y_i - \hat{Y}_i)}{\sum_{i=1}^{n}(W_i - \widehat{W}_i)^2}$, where more weighting has been put on regions with more overlapping.

5. The comparison between RDD and CF in simulation

We use Monte Carlo experiments to compare the bias and precision of the treatment effect estimated by RDD and CF. Simulations are run to show the performance of the two methods in different scenarios. The better relative performances imply the more appropriate method to use in practice. Three scenarios will be demonstrated: 1) deterministic versus probabilistic assignment of treatment; 2) treatment heterogeneity and nonlinearity; 3) omitted variables. For each scenario, we estimate 100 Monte Carlo iterations[2] for simulated samples of sizes 500, 1000, 5000, and 10000. For each iteration *I*, we calculate the bias of an estimator as a percentage of the true treatment effect = $\frac{\widehat{ATE}_I - ATE_I}{ATE_I}$. Under each sample size, there will be a mean bias = $\frac{1}{100}\sum_{I=1}^{100}\left|\frac{\widehat{ATE}_I - ATE_I}{ATE_I}\right|$, which is the average of the absolute value of bias in the 100 iterations. We will also present the root mean square of error (RMSE) of the estimators which is the standard deviation of the difference between the estimation and the true value. RMSE measures the precision of an estimator. RMSE = $\sum_{I=1}^{100}\frac{(\widehat{ATE}_I - ATE_I)^2}{100}$. If the RMSE is small, it means that $\hat{\tau}$ is distributed more centered around the average ATE. In other words, it will be more probable that a particular $\widehat{\text{ATE}}$ is close to ATE if the distribution is denser.

5.1 Deterministic versus probabilistic assignment of treatment

We first consider a simplified set up of our empirical problem:

$$\text{Y}_i = \beta_0 + \tau \text{D}_i + \beta_e \text{expenditure}_i + \beta_s \text{schooling}_i + \gamma \text{X}_i + \mu_i$$
$$\text{D}_i = I(\text{P}_i \le 200)$$
$$\text{P}_i = \text{expenditure}_i + \nu_i$$

The outcome Y_i, is the church attendance of an individual *i*. D_i is the treatment status i.e. whether *i* receive the transfer income. expenditure_i is the monthly expenditure. schooling_i is the years of schooling. X_i is a vector of other control variables including age and community participation etc. τ will be the treatment effect. In this scenario, the true ATE, τ is set to be 7.

We study the situations where d ={deterministic, probabilistic} when the $\nu_{i,d}$= 0 or

[2] We are well noticed that literatures usually have 10000 Monte Carlo iterations for each sample size. However due to our limited computational power, 100 iterations is our optimal number.

$\nu_{i,d}$ ~ N(0, 30). When $\nu_{i,d} = 0$, there is no randomness while deciding the treatment. All observations having expenditure equal to or less than 200 will have treatment. Graphically speaking, there will be a sharp "jump" for data points on the left panel (Figure 1a). When $\nu_{i,d}$ ~ N(0, 30), data points far away from the cutoff will still have probability receiving the treatment close to 1, but the points around the cutoff will have a decreasing probability between [0,1] (Figure 1b). $D_{i,d}$, $Y_{i,d}$ are different treatments and outcomes generate in a deterministic or probabilistic situation.

Figure 1a. Deterministic treatment assignment

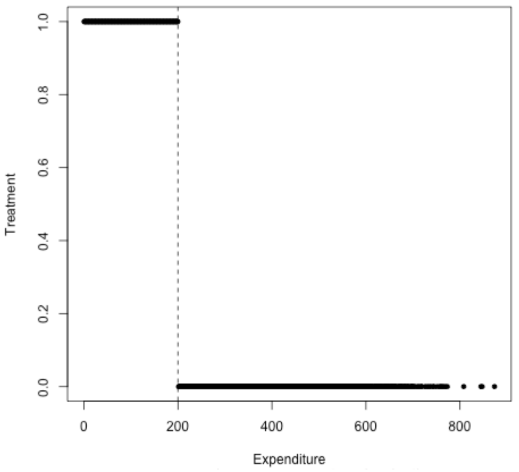
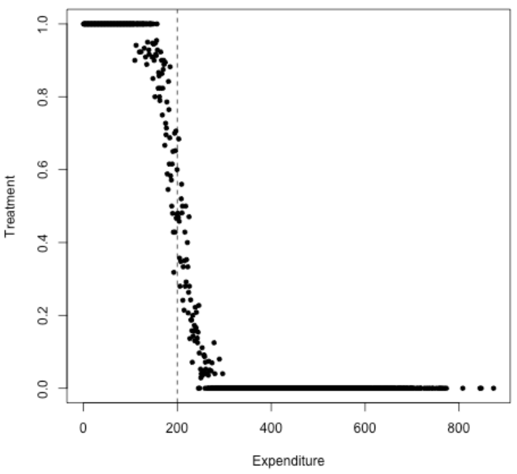


Figure 1b. Probabilistic treatment assignment



RDD estimated is given by

$$Y_{i,d} = \beta_0 + \tau W_i + \beta_e \text{expenditure}_i + \beta_s \text{schooling}_i + \gamma X_i + \mu_i$$

, where W_i is a binary variable indicating the treatment.

For CF, apart from raw ATE, we also compute the ATO.

Table 1. Monte Carlo results of RDD and CF performance of deterministic versus probabilistic assignment of treatment

			Sample size = 500	Sample size = 1000	Sample size = 5000	Sample size = 10000
Deterministic	RDD	Bias	11.22%	8.16%	4.06%	2.64%
		RMSE	0.864	0.517	0.125	0.054
	CF (ATE)	Bias	46.95%	85.48%	907.23%	3518.18%
		RMSE	17.522	57.184	7369.926	84156.945
	CF (ATO)	Bias	42.90%	67.61%	833.25%	3349.92%
		RMSE	14.856	41.194	6460.254	78249.826
Probabilistic	RDD	Bias	11.47%	7.73%	3.92%	2.77%
		RMSE	1.031	0.433	0.130	0.057
	CF (ATE)	Bias	24.53%	20.17%	10.60%	10.21%
		RMSE	4.317	3.037	0.918	1.214
	CF (ATO)	Bias	14.28%	10.17%	4.13%	2.93%
		RMSE	1.541	0.729	0.139	0.072

We can see that RDD performs much better in a deterministic assignment of treatment than CF, regardless of sample size. This is because we do not have overlap regions for CF in a deterministic case. Most data points are either partitioned into a region wholly consists of treated observations or of control observations. CF will mistakenly attribute errorness to the estimation of ATE. It is worth noticing that increasing sample size does not help but rather severely worsen the estimation.

If the treatment is assigned with a probability < 1, the overlapness assumption holds when sample size is big enough. CF estimation is still a more biased than RDD when sample size is only 500 or 1000. By comparing the RMSE, we can see that the CF less precise than RDD. However, starting from 5000 sample size, CF gives ATO estimations that are almost the accurate as RDD. We can see that when the sample size is small, usually less than 1000, RDD wins. When the sample size is big enough and we model correctly, both RDD and CF (ATO) perform well.

While using CF, it is evidential that ATO performs better than ATE in different sample size, the rationale explained in Section 4. Therefore, we will report ATO instead of the raw ATE in the following simulations.

5.2 Treatment heterogeneity and nonlinearity

Now we consider a more complicated set up: what if the treatment effect also depends on the running variable, expenditure? It could be expressed as follows.

$$\begin{aligned}\mathrm{Y}_i \;=\; \beta_0 + \tau \mathrm{D}_i + \; \beta_1 * \mathrm{D}_i * \text{expenditure}_i + \beta_2 * \mathrm{D}_i * (\text{expenditure}_i)^2 + \; \beta_e \text{expenditure}_i \\ + \beta_s \text{schooling}_i + \gamma \mathrm{X}_i + \mu_i\end{aligned}$$

For an observation *i*, if the treatment is switched on, $\mathrm{D}_i = 1$, the additional increase in the outcome is given by $\frac{\partial Y}{\partial D} = \tau + \; \beta_1 * \; \text{expenditure}_i + \beta_2 * (\text{expenditure}_i)^2$. So the treatment effect will depend on the treatment status as well as the treated subject's expenditure and the square of the expenditure. The quadratic polynomial implies that the effect of treatment is nonlinear along expenditure. As we assumed $\beta_1 = 0.05$ and $\beta_2 = -0.00007$, there will be a moderate effect of treatment. Under this setup, the true ATE will be given by $\tau + \; \beta_1 * mean(\text{expenditure}) + \beta_2 * mean(\text{expenditure})^2$.

We run regression on three RDD models and estimate ATE respectively.

$$\mathrm{Y}_i \;=\; \beta_0 + \tau \mathrm{D}_i + \; \beta_1 * \mathrm{W}_i * \text{expenditure}_i + \beta_e \text{expenditure}_i + \beta_s \text{schooling}_i + \gamma \mathrm{X}_i + \mu_i$$

$$Y_i = \beta_0 + \tau D_i + \beta_1 * W_i * \text{expenditure}_i + \beta_2 * W_i * (\text{expenditure}_i)^2 + \beta_e \text{expenditure}_i + \beta_s \text{schooling}_i + \gamma X_i + \mu_i$$

$$Y_i = \beta_0 + \tau D_i + \beta_1 * W_i * \text{expenditure}_i + \beta_2 * W_i * (\text{expenditure}_i)^2 + \beta_3 * W_i * (\text{expenditure}_i)^3 + \beta_e \text{expenditure}_i + \beta_s \text{schooling}_i + \gamma X_i + \mu_i$$

One pitfall of parametric[3] RDD is that the estimation we got could be very sensitive to the order of polynomial in the econometrics model. If there are misspecification errors, the estimation will be biased. In a simulation study, we *design our DGP* and we *know the true ATE*. We can pick the best order of polynomial simply by looking for the model that gives us the least average bias. However, in a real-life application of (parametric) RDD, it is "impossible to know which case has a smaller bias without knowing something about the true function" (Lee and Lemieux, 2010, p.284). We will have to rely on the graphical presentation of the RDD, as Figure 2 to identify the true order. But it is not a precise method of investigating the true function, as it can be blurred by errorness and personal prior prejudices. Very differential $\widehat{\text{ATE}}$ could result we choose different polynomial order, as illustrated in Figure 3a, 3b and 3c.

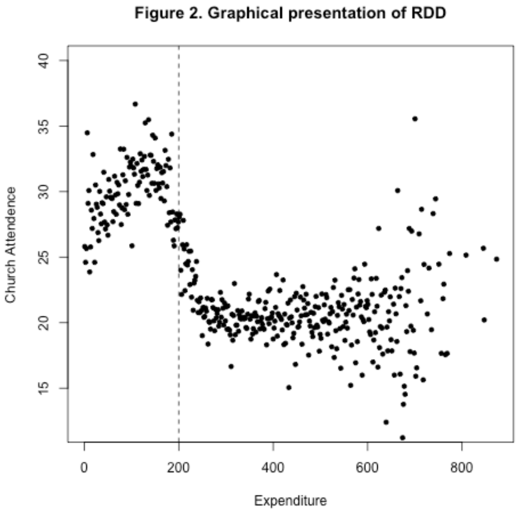


Figure 2. Graphical presentation of RDD

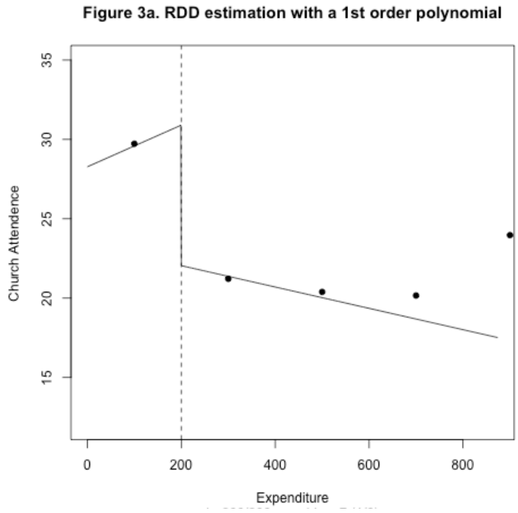


Figure 3a. RDD estimation with a 1st order polynomial

[3] As mentioned in previous section, nonparametric RDD avoids this problem by limiting the dataset to observations close to the cutoff, but it comes at the cost of having less samples.

Figure 3b. RDD estimation with a 2nd order polynomial

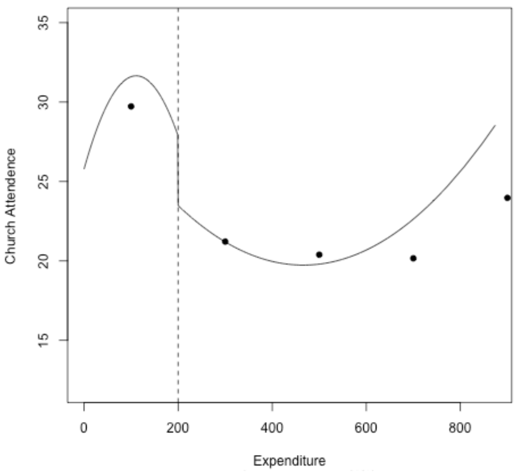
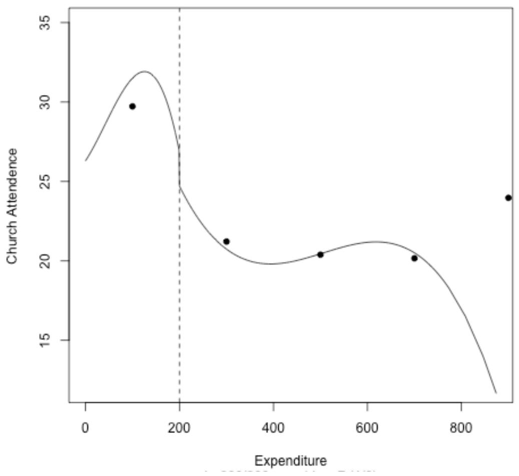


Figure 3c. RDD estimation with a 3rd order polynomial



CF performs better than RDD due to CF's nonparametric nature. There is no need to assume what the true model looks like in a CF. Simply by including a covariate in the CF model, we can include the nonlinearity and interaction terms of the covariate. In the final partitions, data points have similar covariates. ATE will be calculated based on difference in outcome between treated and control observations in the estimating subset. Therefore, without specifying the true function, CF will calculate a less biased $\widehat{\text{ATE}}$.

Table 2. Monte Carlo results of RDD and CF performance of heterogeneous treatment effect

	Polynomial		Sample size = 500	Sample size = 1000	Sample size = 5000	Sample size = 10000
RDD	1st	Bias	25.25%	25.05%	25.54%	25.35%
		RMSE	14.26	13.12	13.30	12.98
	2nd	Bias	27.60%	20.42%	8.40%	6.10%
		RMSE	23.18	13.35	2.23	1.28
	3rd	Bias	125.61%	85.31%	34.33%	22.96%
		RMSE	533.13	244.13	38.59	18.02
CF		Bias	7.08%	4.83%	1.92%	1.49%
		RMSE	1.52	0.72	0.13	0.07

Our results are consistent with the above analysis. The true model is with a 2nd order polynomial function. For RDD, if we model correctly, with a 2nd order polynomial, our estimate is with less bias, starting from a sample size =5000. But if we model incorrectly, our estimate will suffer from misspecification error, and increasing smaple size will not help. Huge bias and imprecision are produced. Also, RDD estimates are imprecise (big RMSE) in all specification and sample sizes. For CF, the estimation is nearly perfect, as both the mean bias and RMSE are very small. The implication is that if we have no very sound reason to argue what the true function is like, CF will avoid the misspecification problem.

5.3 Omitted variables

This time, we test the performance of the two methods against omitted variables. Suppose we cannot observe the schooling level of the respondents, how would our RDD and CF estimate behave? The true model is:

$$\mathrm{Y}_i \;=\; \beta_0 + \tau \mathrm{D}_i + \beta_e \mathrm{expenditure}_i + \beta_s \mathrm{schooling}_i + \gamma \mathrm{X}_i + \mu_i$$

The coefficient β_s is set to be 1.4 as before. We obtain treatment effect estimation $\widehat{\mathrm{ATE}}$ by running RDD and CF respectively *without including* schooling level as a covariate.

The RDD specification is

$$\mathrm{Y}_i \;=\; \beta_0 + \tau \mathrm{D}_i + \beta_e \mathrm{expenditure}_i + \gamma \mathrm{X}_i + \mu_i.$$

In CF, we do not include schooling in the covariate matrix.

Both methods will give a biased estimation. For RDD, since schooling is correlated with other covariates, our estimation of $\hat{\tau}$ will be biased. For CF, the unconfoundedness assumption is also violated

Table 3. Monte Carlo results of RDD and CF performance against omitted variable

			Sample size =500	Sample size =1000	Sample size =5000	Sample size =10000
Variable omitted	RDD	Bias	11.92%	7.34%	3.47%	2.60%
		RMSE	1.067	0.433	0.094	0.049
	CF	Bias	18.36%	11.67%	5.38%	3.72%
		RMSE	2.523	1.113	0.226	0.111
Variable included	RDD	Bias	11.47%	7.73%	3.92%	2.77%
		RMSE	1.031	0.433	0.130	0.057
	CF	Bias	14.28%	10.17%	4.13%	2.93%
		RMSE	1.541	0.729	0.139	0.072

We can see that RDD works relatively better across different sample sizes. Actually, RDD estimation is similar with or without including schooling. On the other hand, we see that the mean bias of CF increases to 18% from 11% when the sample size =500, and increases to 12% from 8% when the sample size = 1000. So CF is more fragile to omitted variable bias when the sample size is small. RDD will be a better candidate when there is an omitted variable.

6. Concluding remarks: Applications in empirical work for using CF over RDD

Being the latest weapons for economists, CARTs have some advantages over the traditional method. But they are not The Solution to every empirical question. Traditional tools still have their place in our toolbox. Through our simulation, we can see that in a probabilistic assignment, CF work under certain conditions: 1) there is a huge sample size, 2) there is no omitted variable. If the conditions are fulfilled, employing CF enjoys the benefits of avoiding misspecification errors compared to RDD. These conditions are consistent with the fact that CF, or CARTs in general, is associate with *big data analytics*. In a big data scenario, data are of high volume (big sample size), and high dimension (a lot of variables are observed). When the two conditions will be held ,we should be comfortable to add CF estimation of treatment effect in our result tables. However, in situations where the insufficient sample size or omitted variables problems are too serious such that economists cannot pretend to ignore them or assume them out, the good way might be the old way.

References:

Athey, Susan, & Imbens, Guido. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences - PNAS*, 113(27), 7353-7360.

Buser, T. (2015). The Effect of Income on Religiousness. *American Economic Journal. Applied Economics*, 7(3), 178-195.

Gulen, H., Jens, C., & Page, T. B. (2020). An application of causal forest in corporate finance: How does financing affect investment?. *Available at SSRN*.

Lee, David S, & Lemieux, Thomas. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48(2), 281-355.

Thistlethwaite, Donald L, & Campbell, Donald T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6), 309-317.

Wager, Stefan, & Athey, Susan. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.