Causal and Statistical Inference Kai Su, Qihang Chen

Poster Report: RDD

1 Analyzing Baysian Network

1.1 Introduction to RDD

When the running variable X takes value at different side of a threshold value t, there will be a treatment(marked $w = 1$) having an effect on the output Y on one side of $X = t$ and no treatment(marked $w = 0$) on the other side. Our goal is to measure the average effect τ solely caused by the treatment. That is, $\tau = E[Y|\,\mathrm{do}(W) = 1] - E[Y|\,\mathrm{do}(W) = 0]$.

Such effect appears as a sudden increase or decrease of y in the neighborhood of $X = t$(See Fig.1). Therefore, τ is traditionally estimated using the see effect of W. Formally put,

$$\hat{\tau} = \lim_{x \to t^+} E[Y|X = x] - \lim_{x \to t^-} E[Y|X = x]. \tag{1}$$

For simplicity, denote

$$E[Y|X = t_+] = \lim_{x \to t^+} E[Y|X = x], \tag{2}$$

$$E[Y|X = t_-] = \lim_{x \to t^-} E[Y|X = x]. \tag{3}$$

However, we are lack of data near the threshold in most cases, difficult for us to calculate these values on the threshold directly. Thus we need some regression to infer the relation between X and Y so as to predict $E[Y|X = t_+]$ and $E[Y|X = t_-]$.

A common practice is to apply linear regression on both sides. The data near the threshold is more valuable, thus we can set a bandwidth b, which is the largest distance where data are taken into account; and a kernel k, which assigns weights to data.

1.2 Baysian Network

Basically, running variable X will decide W and affect Y; W will have effect on Y, as in Fig.2.

bandwidth

After selecting a bandwidth b, it will cause a selection of data, thus there will be a backdoor path between X and Y caused by b. Then the see effect observed by regression between X and Y is not the true causal effect, leading to bias. To eliminate this bias, we need to adjust the way of sampling by changing kernel. See Fig.3.

covariates

Sometimes there are not only running variable X and Y, but also many other variables Z, called covariates, may have effect on Y. If they are independent with X, the regression

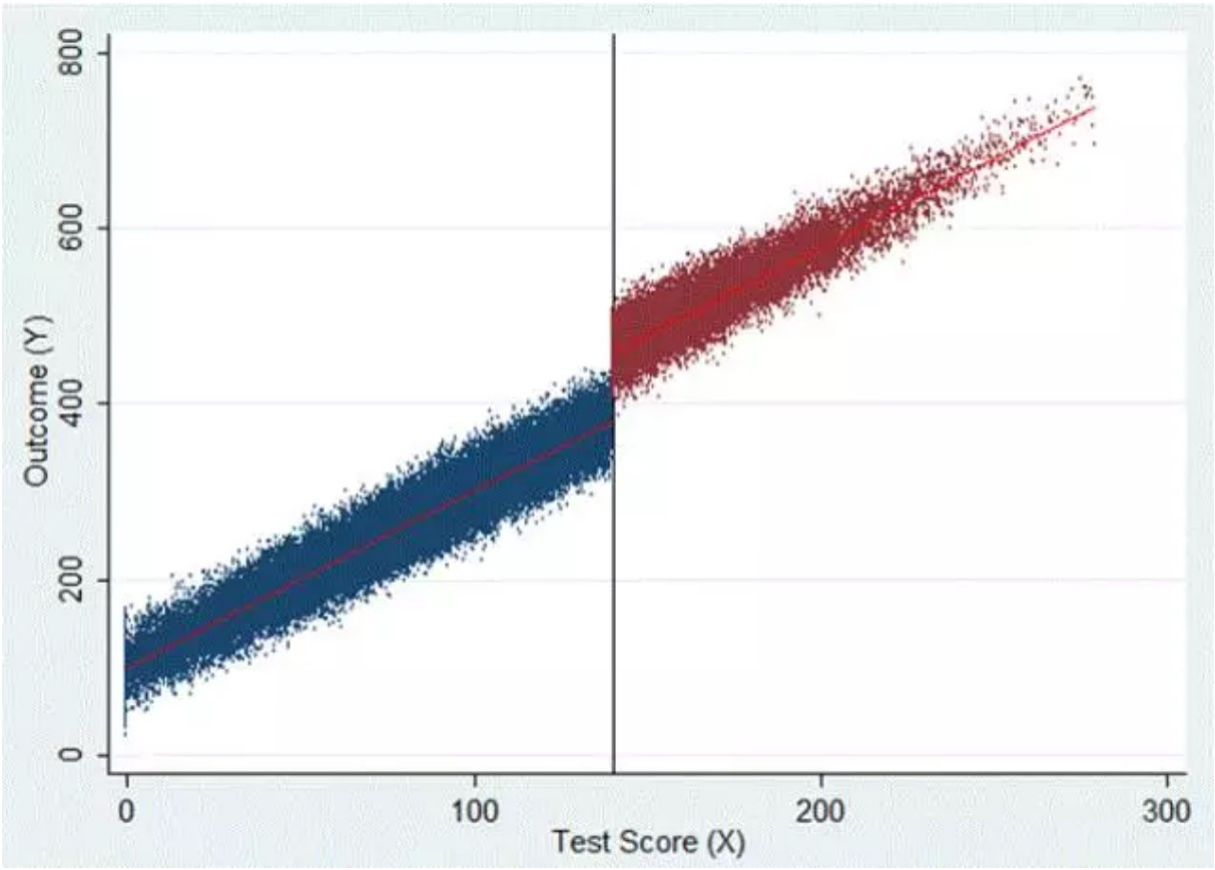


Figure 1: An example of dataset in RDD

still works. However in some cases, Z will affect both X and Y, creating a backdoor path between X and Y. (See Fig.4)

In order to eliminate this bias, we need to find a bandwidth, to make sure among the included samples, Z is independent with X.

2 Experiment and Discussion

2.1 Using kernel to elimimate bias caused by bandwidth

In this section we do not consider covariates. We generate data by different $X - Y$ relation and different distribution of samples.

We randomly generated 500 groups of data and 200 different types of kernels. For each kernel, we calculate the difference of Y at threshold compared with $Y(t)$ by True $X - Y$ relation in those 500 groups of data, getting the average to show the performance of the kernel. Thus we get the best kernel for different bandwidth.

2.2 Differences after adding covariates

After adding covariates, there will be another backdoor path caused by Z, and the perfomance of different kernels may become different.

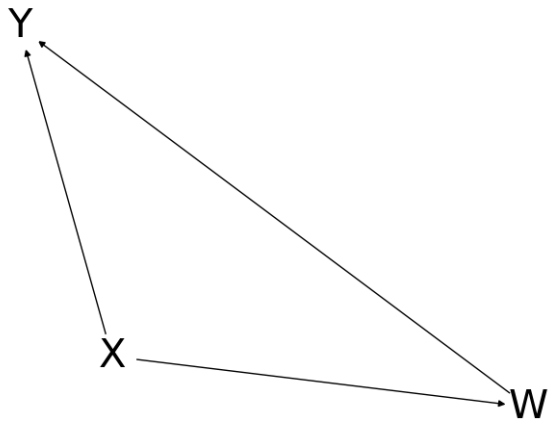


Figure 2: The basic BN for RDD

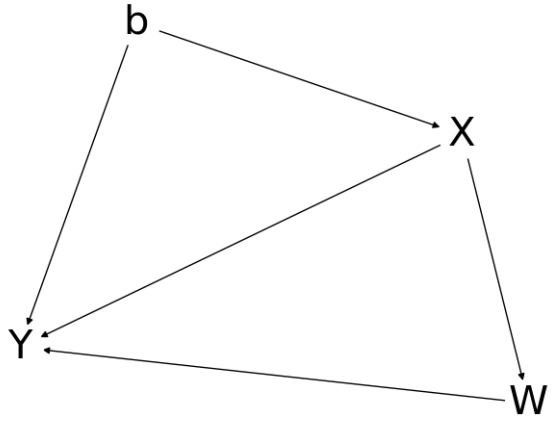


Figure 3: BN with bandwidth

3 Other Attempts

A better way to find best bandwidth

In some articles, cross validation method is used for selecting a optimal bandwidth. However this has not always accurate because samples near threshold may not have similar distribution and $X - Y$ relation.

When bandwidth is large, the $X - Y$ relation's nonlinearity may cause bias if using linear regression. When bandwidth is small, there may be too few samples, leading to a high variation.

We have these ways to quantify the expected error for different bandwidth caused by two aspects above:

(1) Use bandwidth b, suppose the linear regression has the result $Y = a(X - t) + b$, we use a formula to calculate the standard derivation of b two represent the expected error caused by lack of data near threshold.

(2) First use a quadratic hypothesis function on regression to find out a curve showing approximate quadratic relationship between $X - Y$(do not use bandwitdh b). Then project all the samples onto the curve, use linear regression with bandwidth b and find the difference of value between the two ways of regression on the threshold.

Then we find a bandwidth who has least sum of those two kind of errors, just the optimal

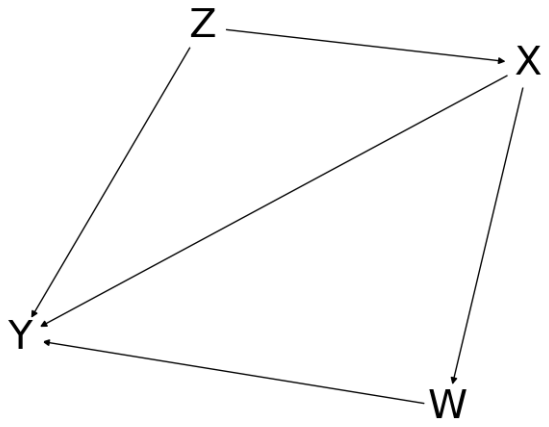
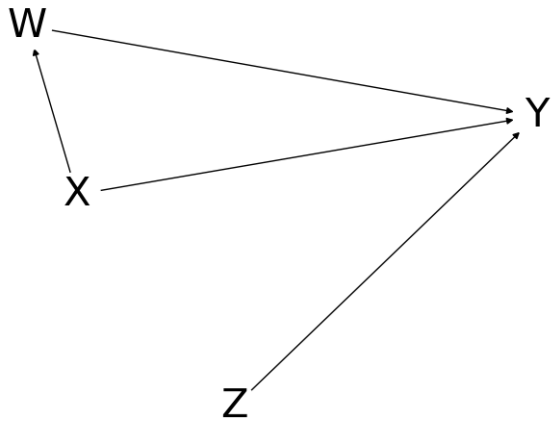
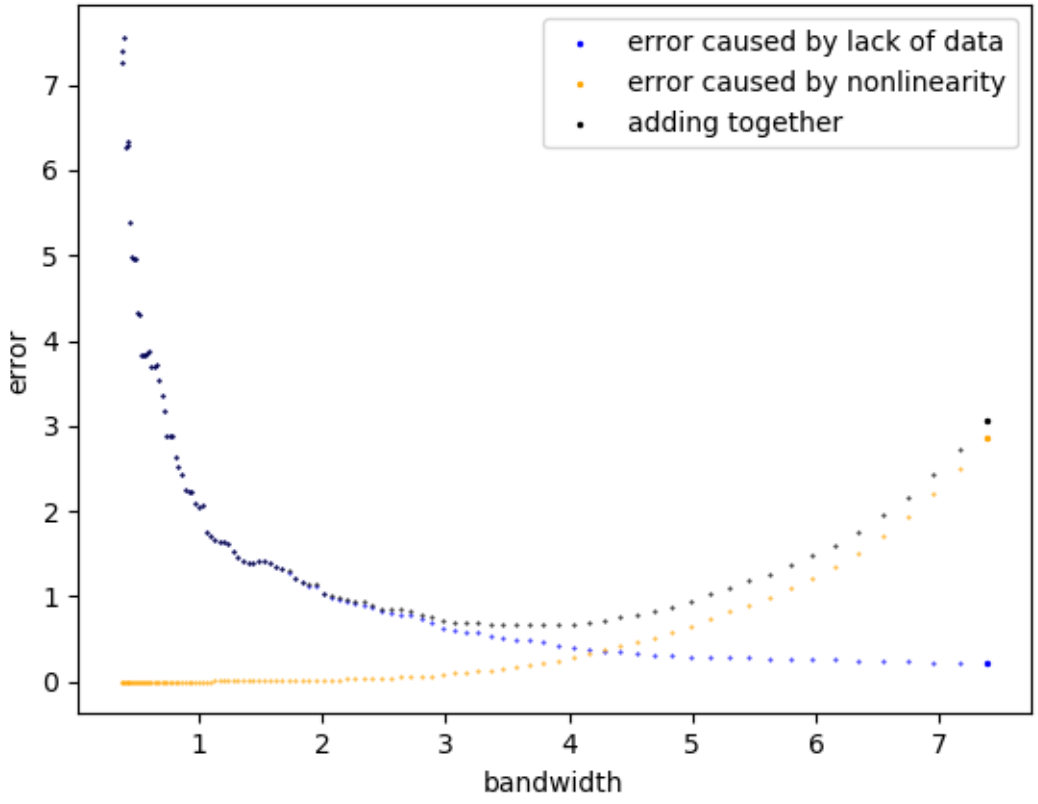


Figure 4: BN with covariates



bandwidth we need.


error caused by lack of data
error caused by nonlinearity
adding together
error
bandwidth
0
1
2
3
4
5
6
7