Causal and Statistical Inference Kai Su, Qihang Chen

Poster Report: RDD

1 Analysis with 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. Denote such an estimator as $\hat{\tau}(b, k)$.

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

When the bandwidth is sufficiently small,

$$P(Y|W = 0, X = t) \approx P(Y|X = t_-), \tag{4}$$

$$P(Y|W = 1, X = t) \approx P(Y|X = t_+). \tag{5}$$

Then

$$\tau = E[Y|\,\mathrm{do}(W = 1), X = t] - E[Y|\,\mathrm{do}(w = 0), X = t] \tag{6}$$

$$\approx E[Y|X = t_+] - E[Y|X = t_-] = \hat{\tau}, \tag{7}$$

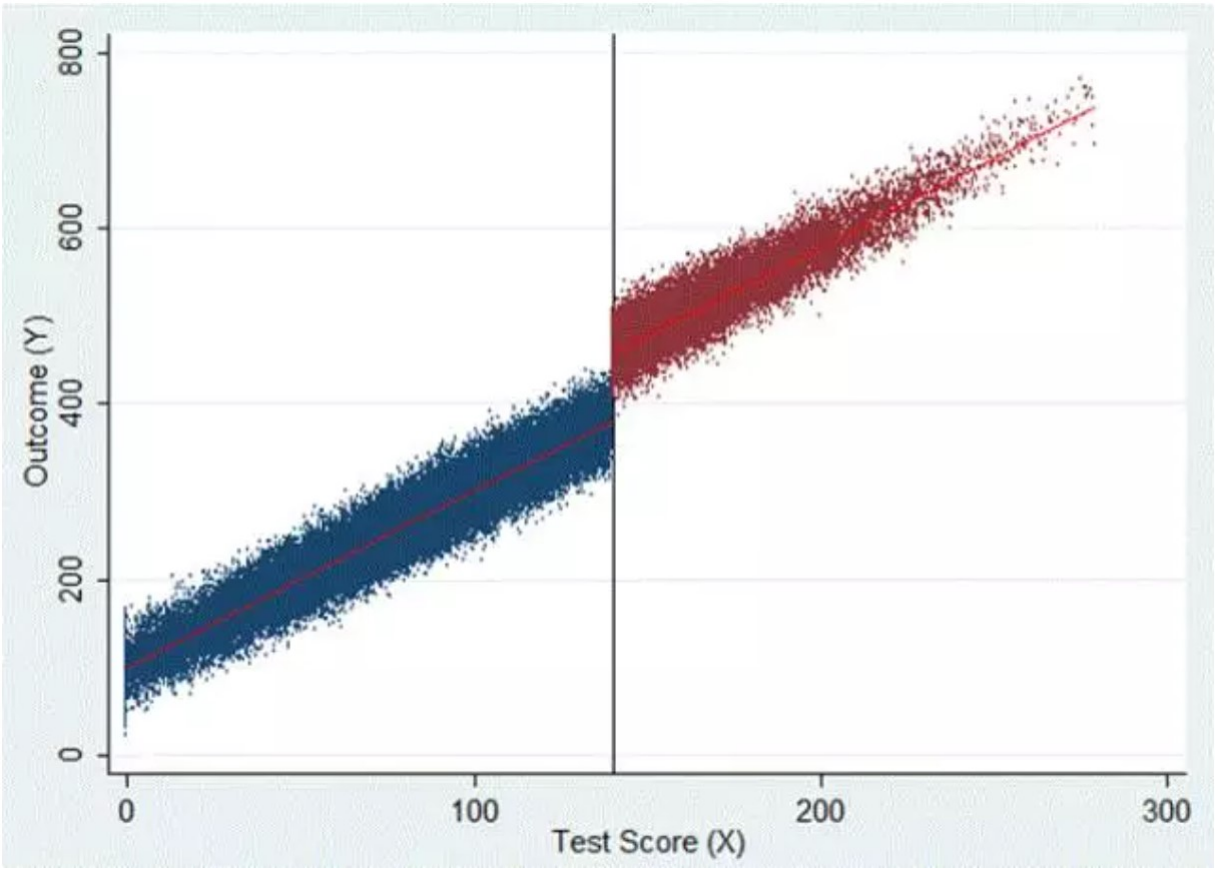


Figure 1: An example of dataset in RDD

proving that the traditional estimation in Eqn.(1) is unbiased.

In contrast, if the bandwidth b is non-negligible, this selection of data specified by b will cause a backdoor path between X and Y. (See Fig.3) Then the see effect observed by regression between X and Y is not the true causal effect, leading to a bias.

covariates

Sometimes there are not only running variables X and Y, but also many other variables Z, called covariates, may have effect on Y. If they are independent with X, the regression still works. However in some cases, Z will affect both X and Y, creating a backdoor path between X and Y. (See Fig.4)

1.3 Obstacles

Under the BN shown in Fig.3, the performance of the estimator $\hat{\tau}(b, k)$ is restricted by two factors:

(a) Large variance due to lack of data;

(b) Bias caused by fitting non-linear $X - Y$ relation with linear regression.

When $b \to 0$, less and less data stay within the bandwidth, so the variance in (a) becomes significant. The bias in (b) converges to 0 under smoothness assumption on the $X - Y$ relation. The conclusion is opposite when $b \to +\infty$.

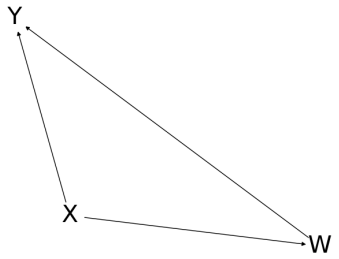


Figure 2: The basic BN for RDD

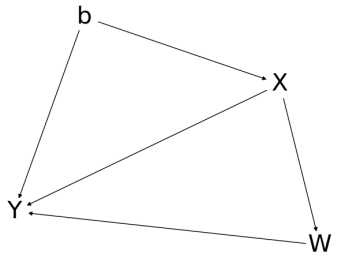


Figure 3: BN with bandwidth

If covariates create a backdoor path as shown in Fig5, the confound effect significantly becomes a source of error in cases with large bandwidth and non-linear $X - Y$ relation.

2 Experiment and Discussion

The bandwidth b and the kernel k together decides the weights of data, yet they are separated into two independent parameters. The bandwidth focuses on the definition of "distance", and thus allows the reuse of kernels across problems of similar properties but with different scales and units.

Moreover, the final probability distribution of samples supplied to the regression model is a composite outcome of the weights of data and the density of actual data. Only then can we improve the accuracy of $\hat{\tau}$ by manipulating b and k. By letting the errors stated in section 1.3 be the loss function, we get a chain of methods to determine b and k:

1. Find a bandwidth b which balances between
   (a) the lack of data;
   (b) and bias caused by non-linearity of $X - Y$ relation.
2. Find a kernel k which
   (a) helps to reduce the bias caused by bandwidth;
   (b) eliminates the confound effect as much as possible.

2.1 Error-minimizing bandwidth selection

In some articles, the method of cross validation is used for selecting an optimal bandwidth. However, this is not always accurate, because the samples near the threshold may perform very differently. Differences in these properties might result in a poor bandwidth for the estimation at the threshold.

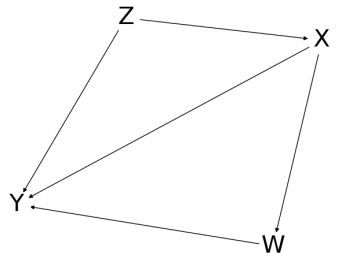


Figure 4: BN with covariates

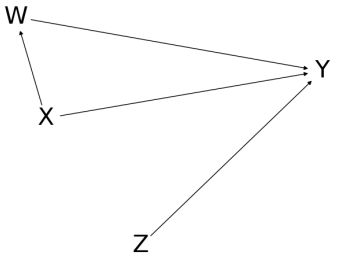


Figure 5: Eliminating $Z \to X$

If $X - Y$ relation is non-linear, the curvature of the function $y = y(x)$ varies. An overestimation of the curvature leads to small bandwidth, which does not take full advantage of the data. An underestimation of the curvature leads to large bandwidth, which might introduce more bias from the non-linear $X - Y$ relation.

Difference of density of data is another issue in RDD particularly. Data near the threshold might be under subjective manipulation. For example, students might work hard not to get below the pass line, so the number of students who is right below 60 is much less than the number of students with other scores. This turbulance in the density of data, again, does not help to select a good bandwidth through cross validation.

Therefore, we propose a method of quantifying the expected error at the cutoff directly, which is an estimation of the two aspects mentioned in section 1.3:

(a) Using 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 to represent the expected error caused by lack of data near threshold.

(b) First use a quadratic hypothesis function on regression to find out a curve showing approximate quadratic relationship between $X - Y$(do not use bandwidth 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.

Finally, we find a bandwidth who has least sum of those two kind of errors, which is the optimal bandwidth we need. (See Fig.6)

2.2 Using kernels to reduce bias caused by bandwidth

In this section we do not consider covariates. There are a vast types of kernels. Triangle and rectangle kernels are used frequently in previous studies. We extracted some properties of these kernels:

1. smooth;

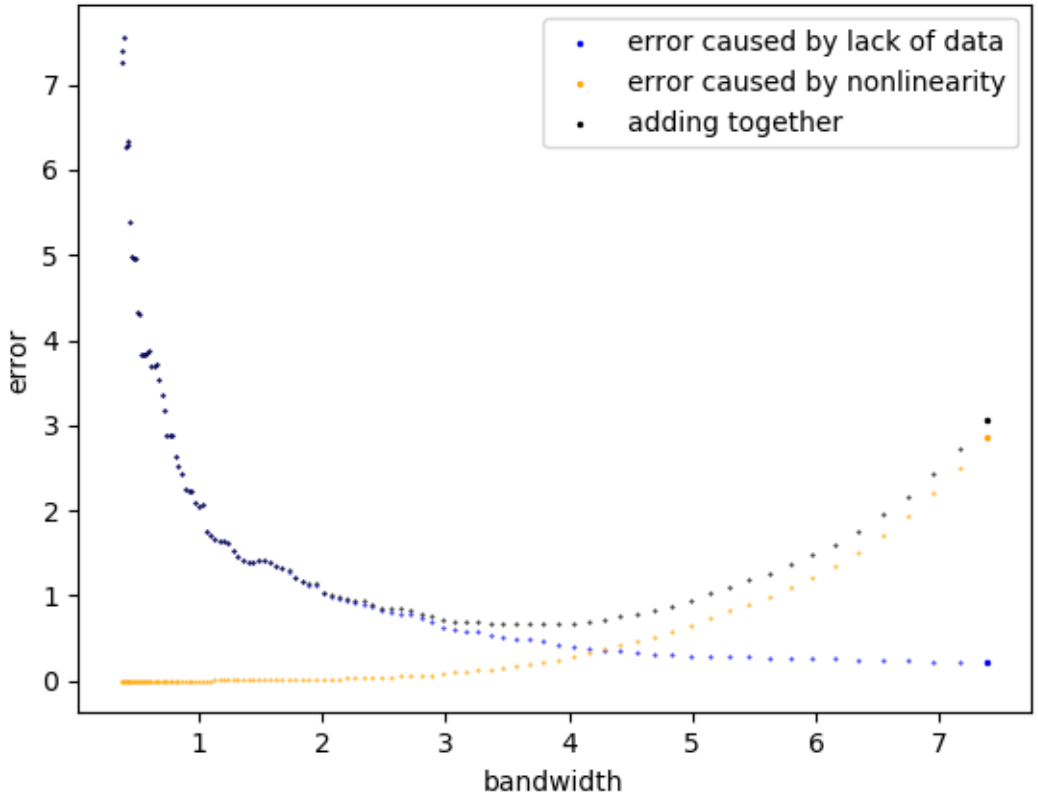


Figure 6: Evaluate error of different bandwidth (on the headstart dataset)

2. monotonic;

3. non-zero within the bandwidth, leaving zeros outside.

In common sense, a good kernel is very likely to contain these properties. So we decided to test against the class of bezier curves, which can mimic most of such function.

$y =$	$0.2(x-59)^2+2$	$-0.2(x-59)^2-0.4(x-59)+2$	$0.1(x-59)^3+2$
Dense middle	case00	case01	case02
Sparse middle	case10	case11	case12
Dense cutoff	case20	case21	case22
Sparse cutoff	case30	case31	case32

Table 1: $X-Y$ relations and data distributions

We designed different $X-Y$ relations and different distributions of samples (See table 1). The cutoff is at $X=59$. Then we randomly generated 500 groups of data and 200 different types of kernels using bezier curves. 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, and use the average as the benchmark for the kernel. Then we get the preferred kernel for a fixed bandwidth.

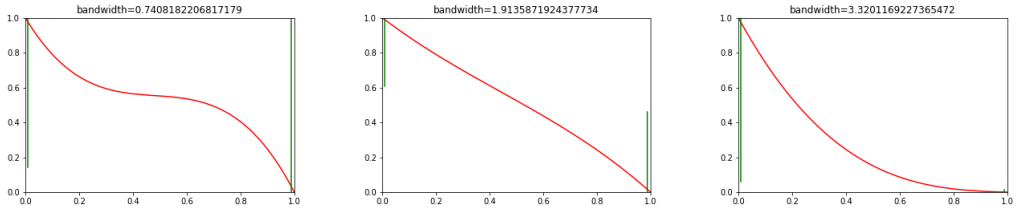


Figure 7: Illustrated best kernels under different bandwidths

Kernels vary with bandwidth

A typical result(case30) is shown in Fig.7. The x-axis is the distance from the cutoff, and the y-axis is the weight of the kernel. At a fixed bandwidth, the kernel does not gather around the cutoff at first, but chooses to spread out the weight to larger distance to the cutoff when bandwidth is small.

This initial behavior of the kernel is related to the error mentioned in section 1.3(a). The lack of samples cause large uncertainty. Assigning more weight to those farther samples basically equivalents to obtaining more samples, and by doing so, the kernel is able to lower the variance of estimation. As the bandwidth b becomes larger, the weights concentrate to the cutoff to lower the bias caused by the nonlinearity in $X - Y$ relation, corresponding to section 1.3(b).

Kernels vary with density of data

Another observation is that, the distribution of samples along X axis really affect the performance of kernels. The reason is the distribution of samples affect the randomness of sampling, thus we need to use the kernel to adjust the sampling method.

Comparing the result in different rows(in html files, can not be shown in pdf), we can find a regular pattern that adding weight on those samples where samples along the X axis are sparse will give better performance.

It is intuitive to explain this observation. If the samples are evenly distributed, it is easier to infer the information from data. But if some areas on X axis has larger density of samples, giving less weight to them can mimic an evenly distributed set of data. Then the kernel will trend to trust samples in the sparse area.

2.3 Differences after adding covariates

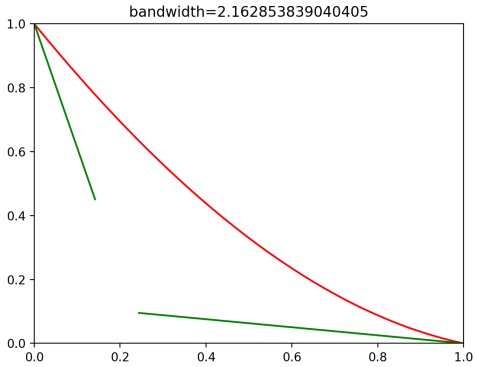
We generate data by different $X - Y$ relation and different ways Z affect X, Y (See table 2) where the samples are uniformly distributed.

After adding covariates, there will be another backdoor path caused by Z, and the perfomance of different kernels becomes different (See Fig.8). This tells us that our selection of bandwidths should concern the effect of covariates.

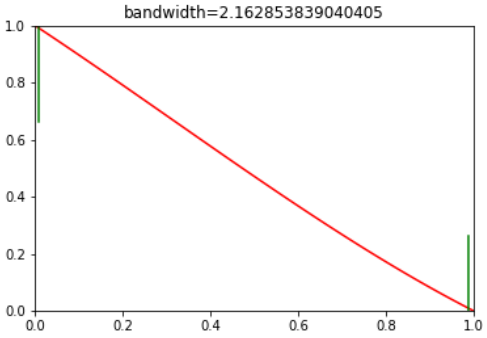
2.4 A step ahead of the kernel

This section discusses the feasibility of eliminating the confound effect using bandwidths and kernels.

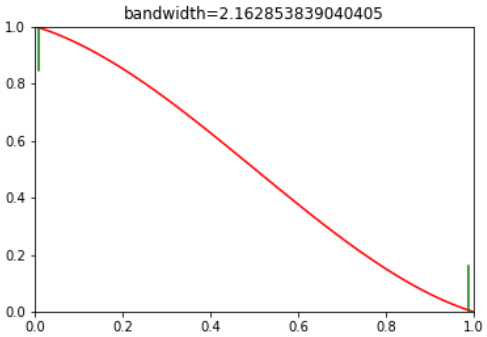
Let $P(*)$ be the density of data, and $P'(*)$ be the probability distribution supplied to the regression model.



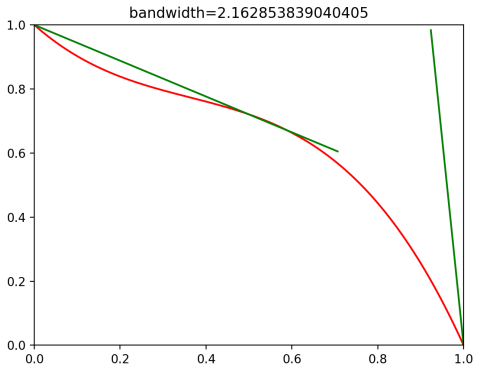
41



z01



z11



z21

Figure 8: Illustrated best kernels under different ways of adding covariates

$f(x) =$	$0.2(x-59)^2+2$	$-0.2(x-59)^2-0.4(x-59)+2$	$0.1(x-59)^3+2$
$x = x_0$ $y = f(x)$	case40	case041	case42
$x = x_0 + z$ $y = f(x) + 0.3z$	casez00	casez01	casez02
$x = x_0 + z^2$ $y = f(x) + 0.3z$	casez10	casez11	casez12
$x = x_0 + z$ $y = f(x) + 0.5Z^2$	casez20	casez21	casez22

Table 2: $X - Y$ relations and ways of adding Z

Let $\omega(x)$ be the random variable representing the nomalized weight of samples located at x. By the definition of bandwidths and kernels, the weight before normalization for x is

$$\omega'(x) = \begin{cases} k\left(\frac{|x-t|}{b}\right) & |x-t| \le b \\ 0 & |x-t| > b \end{cases}. \tag{8}$$

After normalization,

$$\omega(x) = \frac{\omega'(x)}{\sum_{x'} \omega'(x')P(x')}. \tag{9}$$

Then P' is well-defined by $P'(x,*) = \omega(x)P(x,*)$, since

$$\begin{aligned} \sum P'(*) =& \sum_x P'(x) && (10)\\ =& \sum_x \omega(x)P'(x) && (11)\\ =& \sum_x \frac{\omega'(x)P(x)}{\sum_{x'} \omega'(x')P(x')} && (12)\\ =& 1. && (13) \end{aligned}$$

Removing $Z \to X$ from Fig.4 is adding an independence relation between X and Z, which requires

$$\begin{aligned} P'(x,z) =& P'(x)P'(z) && (14)\\ \omega(x)P(x,z) =& \omega(x)P(x)P'(z) && (15)\\ P(x,z) =& P(x)P'(z) && (16)\\ \sum_x P(x,z) =& \sum_x P(x)P'(z) && (17)\\ P(z) =& P'(z) && (18) \end{aligned}$$

Combining Eqn.(16)(18), $P(x,z) = P(x)P(z)$. Therefore, $Z \to X$ can be removed only if $X \perp\!\!\!\perp Z$ are independent at the very beginning.

If we try to remove $Z \to Y$ instead, then $Y \perp\!\!\!\perp Z \mid X$.

$$
\begin{aligned}
P'(y, z \mid x) =& P'(y \mid x)P'(z \mid x) && (19)\\
P'(x, y, z)P'(x) =& P'(x, y)P'(x, z) && (20)\\
(\omega(x)P(x, y, z))(\omega(x)P(x)) =& (\omega(x)P(x, y))(\omega(x)P(x, z)) && (21)\\
P(x, y, z)P(x) =& P(x, y)P(x, z) && (22)\\
P(y, z \mid x) =& P(y \mid x)P(z \mid x), && (23)
\end{aligned}
$$

giving that $Y \perp\!\!\!\perp Z \mid X$ initially.

In conclusion, the confound effect cannot be simply removed by using bandwidths and kernels. To correctly identify $P(y \mid \hat{x})$, one must assign the weights according to Z, too.

$$
\begin{aligned}
P'(y \mid x) =& P(y \mid \hat{x}) && (24)\\
\frac{P'(x, y)}{P'(x)} =& \sum_z P(y \mid \hat{x}, z)P(z \mid \hat{x}) && (25)\\
\frac{\sum_z \omega(x, z)P(x, y, z)}{\sum_z \omega(x, z)P(x, z)} =& \sum_z P(y \mid x, z)P(z) && (26)
\end{aligned}
$$

Thus we only need to find a set of weights that satisfies such equation. A deeper look requires more discussion and will be omitted.

In particular, if Y is linear with respect to X and Z, the parameters in $Y = X\alpha + Z\beta + \varepsilon$ can be obtained by linear regression, thus the do effect can be easily estimated. However, this is not necessary, since the confound effect never shows up in the estimation of treatment in this case.