Epidemiological Methods I

Jacob Simmering, PhD

February 22, 2018

Objectives

- Describe the major methods and approaches used in observational data analysis
- Describe how these methods differ from both experimental and quasi-experimental approaches
- Describe when to use a cohort study, a case-control study and other designs
- Describe the advantages and disadvantages of each design

Causal Effect Estimation

A causal effect of a factor is typically what we are interested in understanding, regardless of whether we are using experimental or non-experimental methods. The causal effect may be the effect of a risk factor such as an exposure or the effect of a treatment decision. For instance, we may be interested in the effect of taking ibuprofen on a headache 60 minutes after taking the pill. The causal effect of ibuprofen is the difference after 60 minutes between the two states - taking ($T = 1$) and not taking ($T = 0$) the medication. Let's say the intensity of the headache is given by the variable H and then we can express the causal effect as the $E(H|T = 1) - E(H|T = 0)$ (note the $E()$ is the expected value function, you can think of it like an average).

The "fundamental problem of causal inference" is that we cannot observe both $E(H|T = 1)$ and $E(H|T = 0)$ on the same person, at least not for the same headache. A subject must either be in the treatment group or in the non-treatment group for a given headache and so we must rely statistical methods to estimate this effect.

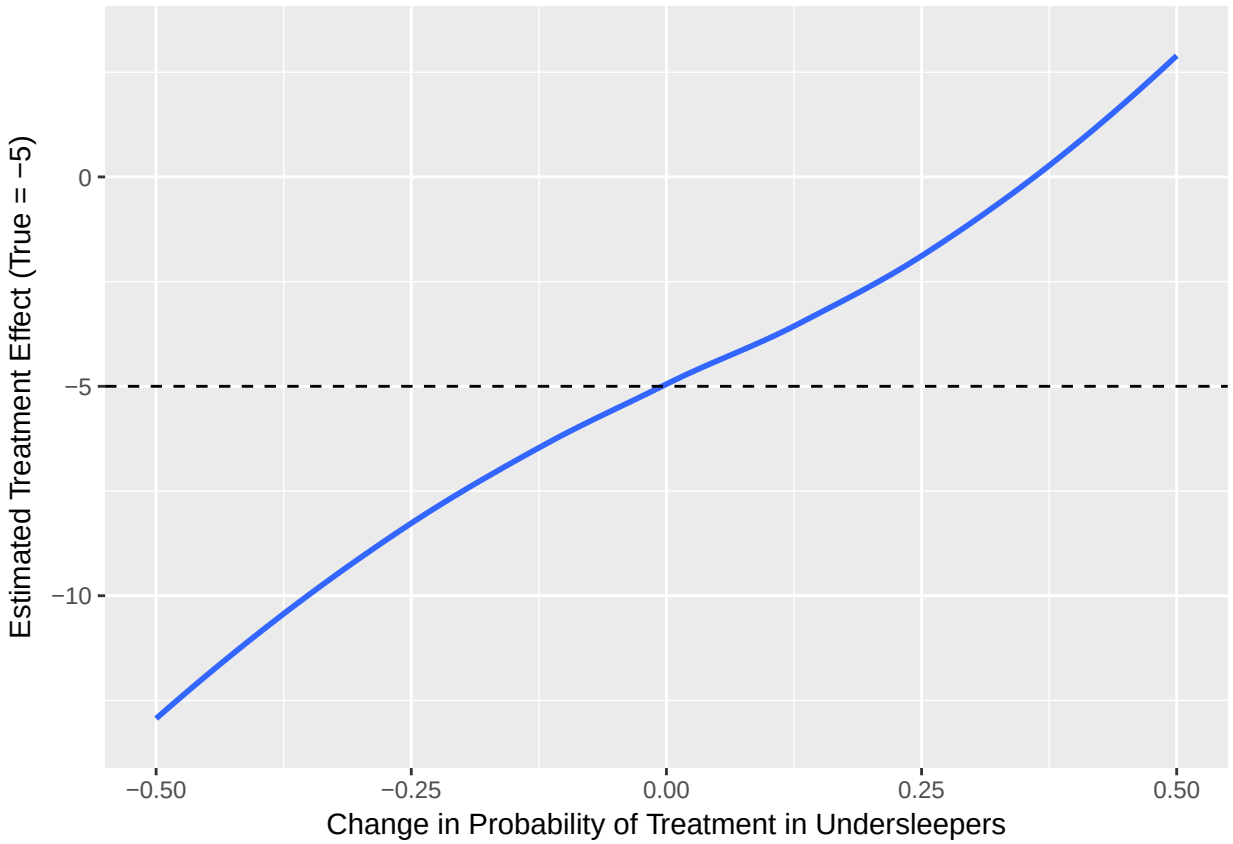
However, when using these methods, we must be concerned with the potential for confounding. There may be another variable, say amount of sleep the prior night denoted as S, that varies between subjects and is related to headache intensity.

In a random assignment framework, such as used in many experimental approaches, or a quasi-random assignment, as in the quasi-experimental methods, we are able to ignore the effect of S on H because it is balanced by virtue of the random assignment. Any potential relationship between S and H is broken by virtue of the random assignment.

You may recruit a sample of people with headaches and assign half to take ibuprofen and the other half to a placebo. Suppose that a typical headache has a "intensity score" of 50 and the score increases or decreases by 5 for each additional hour slept more than/less than 7 hours the prior night. If a patient takes medication for the headache, the intensity decreases by 5. Using these assumptions, we can generate a possible effect for a sample of 500 subjects.

Using this possible sample to calculate the means for each of the groups, you would see an average intensity of 44.8 in those who were treated and 49.9 in those who were not for an estimated causal change in intensity with the meds of -5.05, an effect reasonably close to the true effect of -5.0.

However, suppose that people who did not sleep as much were more likely to take medications to address their headache? Perhaps these people, because they are aware of the lack of sleep from the prior night, are more likely to address the headache immediately instead of waiting to see if it passes. What happens to the estimated treatment effect then?



The estimated difference between those who were treated and those who were not treated rapidly deviates from the true value as the probability of seeking treatment shifts from balanced to extremely unbalanced.

This lack of control over assignment and the potential for the non-random assignment to introduce confounding and bias means that when we study treatment effects using observational data we must explicitly address this feature of the data. This is most typically done with epidemiological and observational methods.

Why

A natural question given the limitation of observational methods relative to randomization-based methods is why/when would we use them? Indeed, the ability of observational methods to support causal inference is not as strong as in a randomized trial; however, frequently we are presented with questions that cannot, either for ethical or logistical, reasons be studied with an randomized design.

Suppose a researcher wanted to study a very rare disease and the effectiveness of a given treatment on the progress of that disease. If there are only 1 case per 100,000, there may be only 3,500 cases in the entire US. Recruitment and followup with such a small potential sample out of 350 million people would prevent a practical randomized study.

Likewise, if the disease has a long latent period. For instance, if a researcher is interested in the effects of an exposure on long-term health over a period of 50 years, the logistical burdens of a randomized design increase rapidly.

More frequently, there are issues with the ethics of conducting a randomized study. For instance, if we have a treatment known to be effective for a disease it is not permitted to give half the patients a placebo.

Often times, either due to these ethical and logistical issues, we are restricted to using observational data where treatment choice is not randomly assigned. However, sometimes the observational data can be useful to provide "real world" evidence of the effect of a treatment or other exposure. While randomized studies

have strong internal validity, often the inclusion and exclusion rules used to create a study likely to produce a significant effect with the smallest sample size means that the study population may not be an accurate reflection of the source population. The average treatment effect estimated in an randomized trial may not be a good reflection of the average treatment effect on a randomly selected patient with the disease.

Major Designs in Epidemiology

All of the study designs in epidemiology attempt to obtain an estimate of the true causal effect of an exposure or treatment in the presence of non-random assignment. The two major designs are cohort and case-control studies while a collection of less common, more purpose drive designs are also used.

Cohort

A cohort study is focused on recruitment by exposure status and ascertaining the disease or outcome status. In this design, enrollees are recruited either by random sampling or by stratification based on exposure status. For instance, subjects may be recruited based on whether they were or were not taking a particular drug and then the outcome is determined.

A cohort study may be done entirely retrospectively - where the outcome is determined and known at the time of the enrollment into the study - or prospectively - where the outcome is not yet realized at the time of enrollment and is discovered during followup, potentially years later.

As a general rule, prospective cohort studies are considered to be a stronger form as the risk of temporal confusion of exposure and disease is lower. Logically, the outcome must occur after the treatment or exposure of interest if the treatment/exposure is to be causally related to the outcome. With retrospective cohort studies, especially if the exposure or date of disease onset occurred significantly in the past, it may be hard to say which confidence that the exposure/treatment occurred before the outcome. In a prospective study, this is less of a threat as the outcome is not yet realized at the time of enrollment.

The extent to which this matters in the context of large secondary claims-based databases is unclear. The data in those databases is recorded prospectively but the outcome is known to the researcher at the time of the analysis. This distinction matters more in the context of primary data collection and similar studies than when using large secondary data sources.

This framework provides for direct estimation of the relative risk (e.g., the multiplicative difference in the probabilities) between the two exposed and unexposed groups. The relative risk is formally given by

$$RR = \frac{P(O = 1|E = 1)}{P(O = 1|E = 0)}$$

where E takes the value of 1 when a subject is exposed and 0 otherwise and O takes the value 1 if the person has the outcome of interest and 0 otherwise. This can be made more concrete as

$$RR = \frac{\frac{N(O=1,E=1)}{N(E=1)}}{\frac{N(O=0,E=0)}{N(E=0)}}$$

where the function $N()$ returns the number of subjects with the described set of values.

If the data is summarized in the "standard table" as shown, then the RR reduces to

	Diseased	Not Diseased
Exposed	a	b
Not Exposed	c	d

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

In contrast to the odds ratio, discussed later, the relative risk is easier to understand and more clearly communicates the association between the exposure and the outcome.

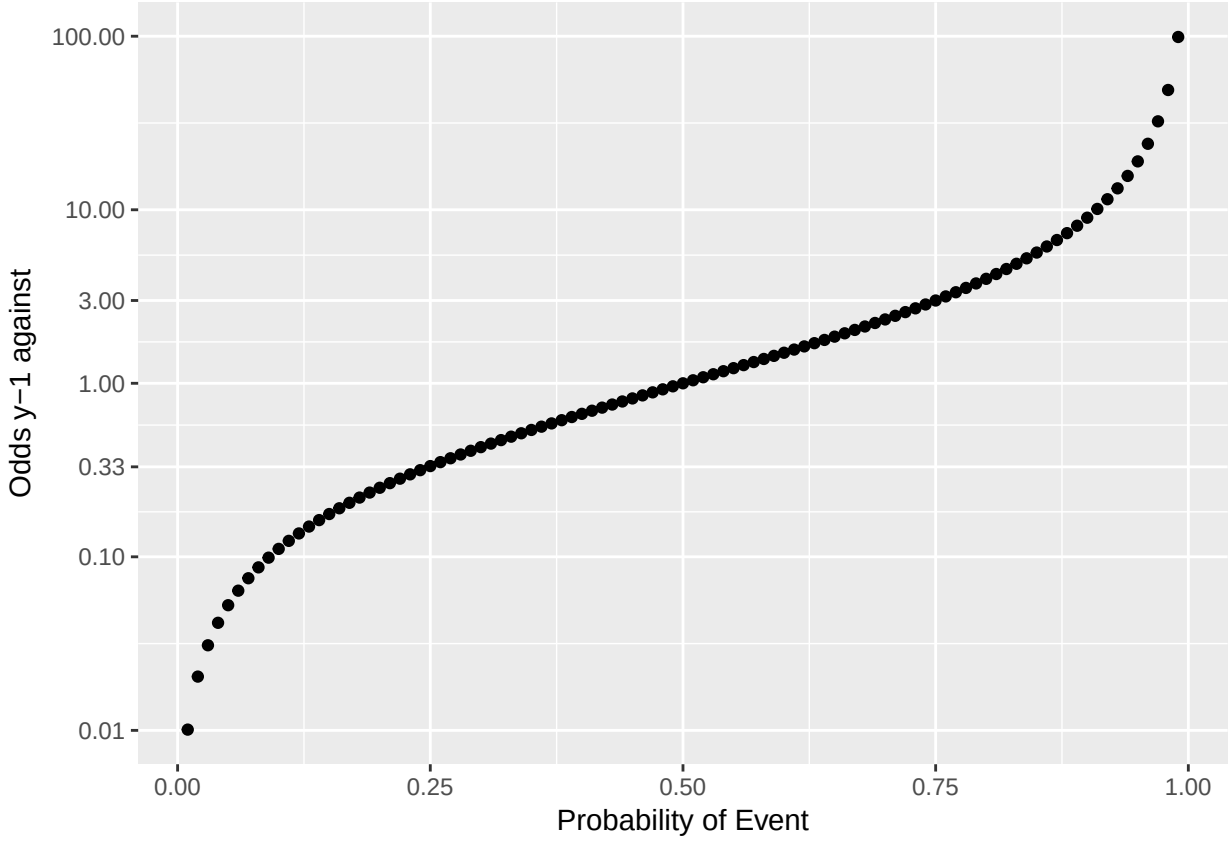
Cohort studies are particularly useful when studying rare exposures. Since the subjects are enrolled on the basis of exposure, it is easy to target potentially rare exposures and build a sufficient sample size. With rare exposures, you get the greatest value for your research budget with cohort studies as you are able to directly focus in on recruiting these subjects.

However, this design is a poor choice for studying rare diseases as a result. As the study recruitment is based on exposure status, it is difficult to recruit a sufficient number of people with the outcome of interest when the outcome is rare. If the outcome only occurs in 1 in 1,000 people in the controls and 5 in 1,000 people in the exposures, it may be difficult to obtain and sufficient sample size to study that outcome within a realistic research plan.

Case-Control

Case-control studies are the second major design used in observational studies. In contrast to cohort studies which group on exposure status, case-control studies group based on the outcome. Cases are subjects who have the disease or outcome of interest while controls are subjects who do not. After being grouped on the basis of disease, the levels of exposure in the two groups is compared.

Since recruitment was based on the outcome and not based on the exposure, we can only compare the multiplicative difference in the odds of disease between the two groups. The odds of an event are given by $\frac{p}{1-p}$.



Odds are generally harder to understand that probability for most people. Unfortunately, in the context of case-control studies the measure of association is even more confusing as it is the ratio of the odds or odds ratio (OR). The odds ratio is computed by calculating the odds of developing the outcome of interest for each the exposed and unexposed groups. These odds are then compared to produce the odds ratio as a measure of association. More formally,

$$OR = \text{Odds}_{E=1}/\text{Odds}_{E=0}$$

or, using the $N()$ function from earlier,

$$OR = \frac{N(E=1,O=1)/N(E=1,O=0)}{N(E=0,O=1)/N(E=0,O=0)}$$

which, after you cross-multiply, reduces to

$$OR = \frac{N(E=1,O=1)N(E=0,O=0)}{N(E=0,O=1)N(E=1,O=0)}$$

Or, in context of the standard table,

	Diseased	Not Diseased
Exposed	a	b
Not Exposed	c	d

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

It is tempting to confuse the relative risk from a cohort study with the odds ratio or discuss the change measured by the odds ratio as a change in risk. However, the value of the odds ratio only relates to the change in odds. This is a subtle, but potentially important, difference.

Whereas cohort studies are well suited to study rare exposure, case-control studies are well suited to study rare diseases and outcomes. Since recruitment is based on outcome, it is easy to start those with a rare condition until a sufficient sample is obtained.

There is another benefit to using case-control studies to investigate rate diseases - when the disease is rare, the value of the odds ratio converges with the value of the relative risk. This is known as the rare disease assumption.

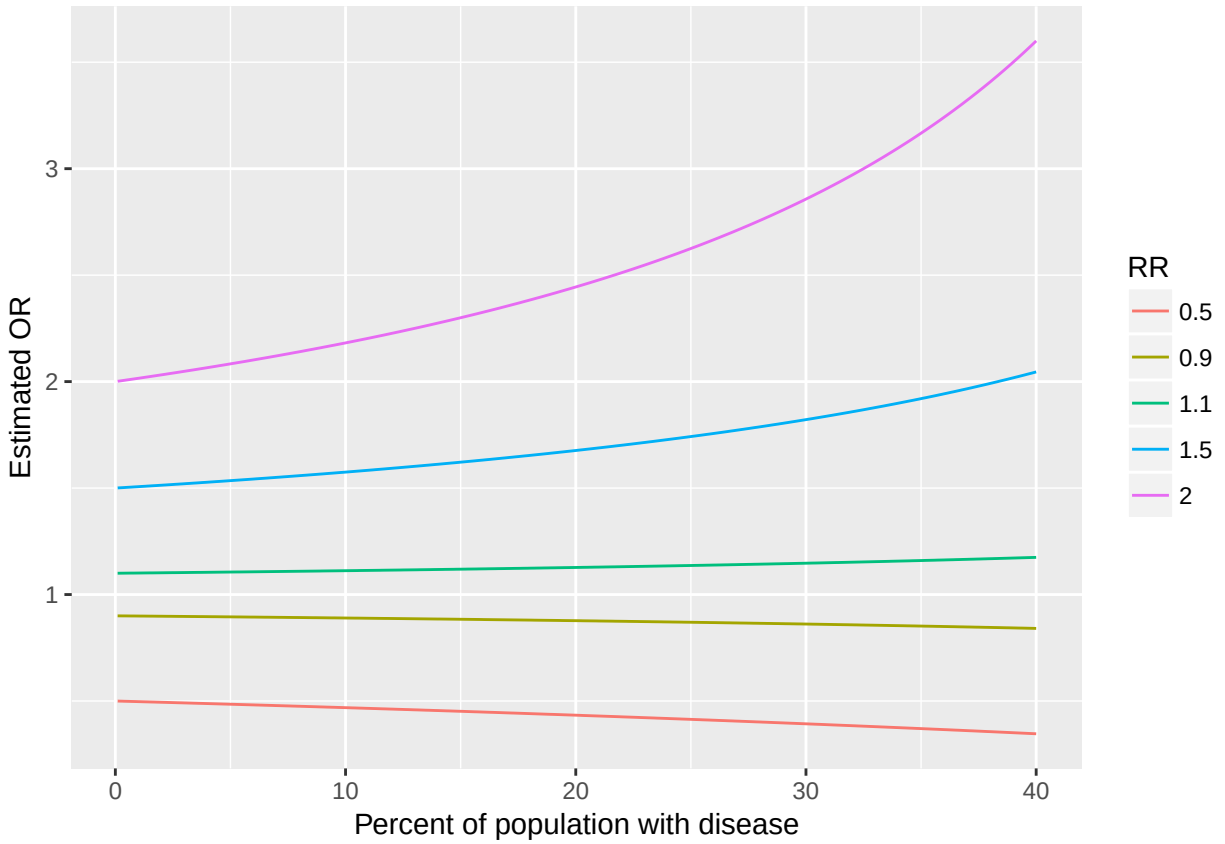
In the case of a rare disease, it is likely that $N(E=1,O=1) << N(E=1,O=0)$ (e.g., the number of exposed but not ill subjects is far larger than the number of exposed and ill subjects). If this is the case, then $N(E=1,O=0) \simeq N(E=1,O=0) + N(E=1,O=1)$. Likewise, it is the case that $N(E=0,O=0) \simeq N(E=0,O=0) + N(E=0,O=1)$

If we rewrite the expression for the relative risk with this knowledge

$$RR = \frac{N(E=1,O=1)/N(E=1)}{N(E=0,O=1)/N(E=0)} \simeq \frac{N(E=1,O=1)/N(E=1,O=0)}{N(E=0,O=1)/N(E=0,D=0)} = OR$$

as the relative proportion of the population with the disease goes to zero.

To make this more concrete, consider a population where 30% of people have an exposure of interest - what would the odds ratio be for a variety of RRs?



As you may note, as the relative number of people with the disease grows the validity of the $RR \approx OR$ assumption decreases. This is especially true for odd ratios further from 1 - the deviation at 1.1 or 0.9 was much smaller, even at 40%, than the examples at 0.5, 1.5 or 2.0. A typical rule of thumb you may see is that when the percent of the population with the disease is under 5 or 10%, it is reasonable to assume that $RR \approx OR$\$ and more dangerous otherwise.

However, this rare disease strength brings the primary limitation to case-control studies: it is a poor design for the study of a rare exposure. As with cohort studies which required an impossibly large sample size to study a rare disease, case-control studies require a very large sample size to have sufficient power to study a rare exposure.

Additional, some may consider case-control designs to have lower generalization and greater risk of bias than cohort studies. In general, the concern is that 1) cases may be more motivated than controls due to their disease status and therefore may provide greater recall (or false recall) and 2) since case status was realized prior to enrollment, there may be greater issues with temporal misordering of exposure and disease onset relative to prospective cohort studies. There is an additional potential threat, especially when studying fatal diseases, with sampling prevalent cases and not only incident or new cases. Sampling of prevalent cases may in these cases include factors related to survival as those who survive longest are most likely to be included in your study under prevalent case sampling. These potential threats will be discussed in greater detail in the next lecture.

Other Common Designs in Epidemiology

Nested Case-Control

Nested case-control designs are case-control studies used within the context of a larger, prior study - most commonly a cohort study. The case-control study is considered nested as it is situated within a well-defined

study sample.

A classic example may be a large pre-existing cohort studying the development of a disease of interest. You may be interested in how the level of certain biomarkers are associated with the disease risk. It would be expensive to collect these biomarkers and, after a certain point, the marginal increase in information from each new control goes to zero. Instead of testing all the controls for the biomarker, you may form a nested case-control study and only test a subset.

Case-Crossover

A case-crossover design is useful when the exposure of interest is transient and the outcome is relatively immediate. This design seems to be used most frequently when studying the effects of weather.

The effect of a heat wave, for example, on human health is an interesting question but difficult to study. A heat wave typically only occurs between May and September and so estimation of the effect may be confounded by other seasonal factors (e.g., school is out for the summer). A case-crossover design would identify people with the outcome of interest and gather information about the time period before the onset of the outcome of interest. Control information would be provided by the same subject but would be sampled at $\pm k$ years. In the papers that I have seen that used this design, control data was collected from years $-3, -2, -1, +1, +2, +3$ years, as possible, from the time on onset.

This is a potentially powerful design as the "temporal" and subject introduced confounding is controlled by the matching implicit in the design. However, it is only suitable for studying transient exposures (e.g., a long duration exposure at a workplace would not be suitable) and for rapidly developing and detecting outcomes (e.g., you would not use this to study a disease with a long latent period).

Case-Time-Control

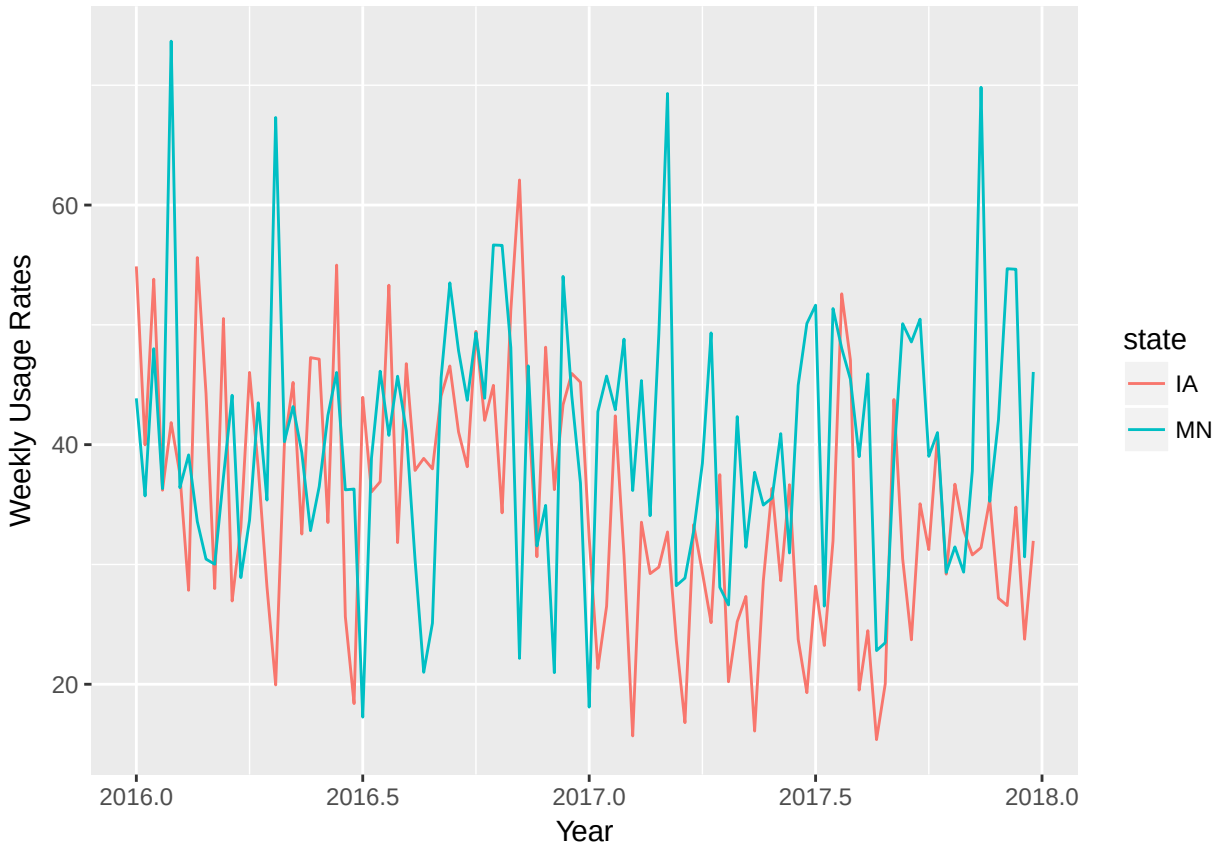
I am pretty sure what this design describes is the same as difference-in-differences (Diff-in-Diff) in econometrics. Basically, you are interested in the effect of an intervention or an exposure between two populations over time. You collect data on the exposed population and also data on a similar unexposed population both before and after the exposure.

For instance, you might assess the effectiveness of an anti-drug effort in one state by comparing the rate of drug use to nearby states over time. You would build a regression model of the form:

$$E(y_{t,i}) = \beta_0 + \beta_1 t + \beta_2 T(i,t) + \beta_3 t T(i,t)$$

where t is the time point of the observation, i is the unit of the observation, $T(i,t)$ takes the value of 1 for units i where the intervention is in effect and where t is later than the start time of the intervention and 0 otherwise.

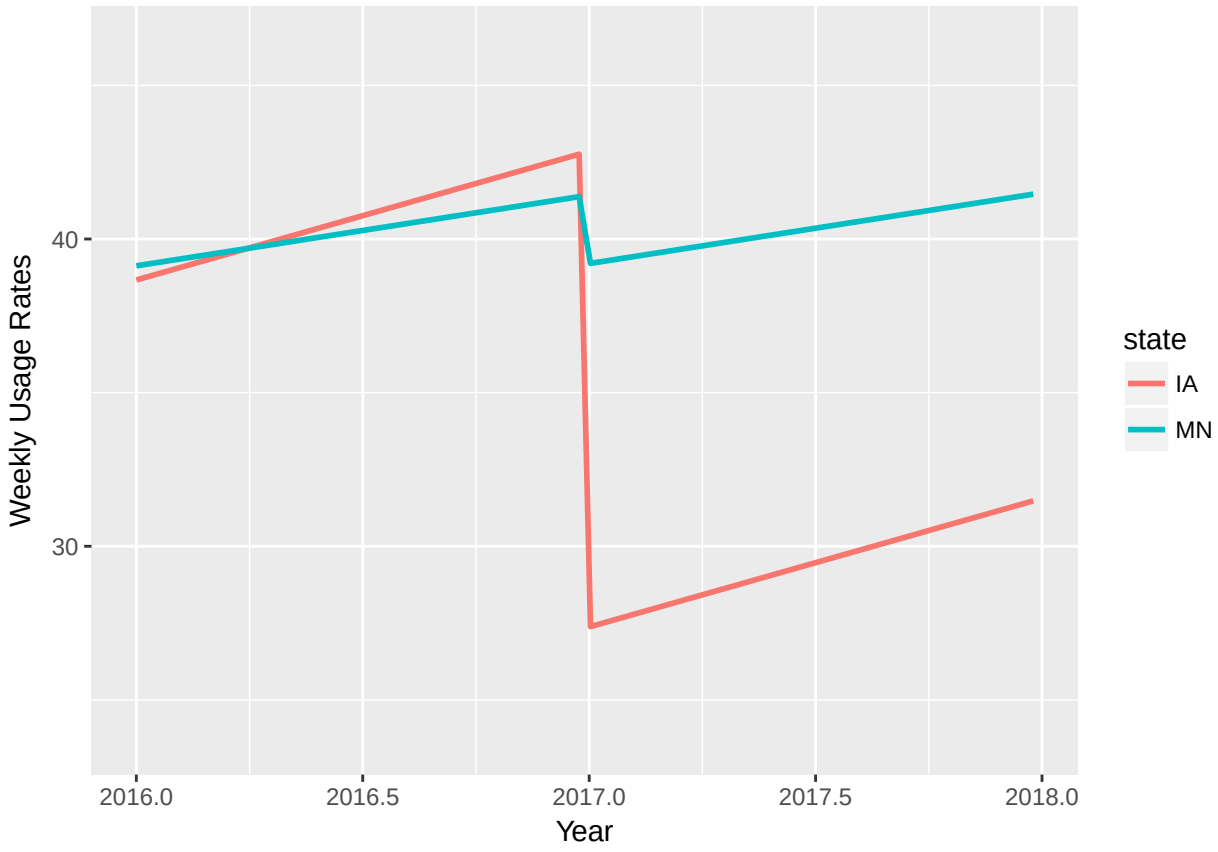
To make this more concrete, suppose you had data from an intervention on drug use in Iowa and you wanted to compare the usage rates in 2016 to those in 2017 when a new intervention was deployed. You selected MN as a control state as the population in MN is similar to that in Iowa.



You would fit a model as described above to the data, you see that both states had a drop when the intervention went into effect but the state with the intervention (Iowa) had a much greater drop than the state without the intervention.

If you look at the regression numbers, you see that after the start of the intervention the rate in Iowa went down by 11.4 more points than the rate in MN.

This method can be useful if you want to determine the effect of some process or exposure but are worried about potential confounding by trends and other natural processes. The biggest concern with this method is making sure that the controls are a reasonable reflection of the counter-factual state that would be assumed by the intervention/exposed group.



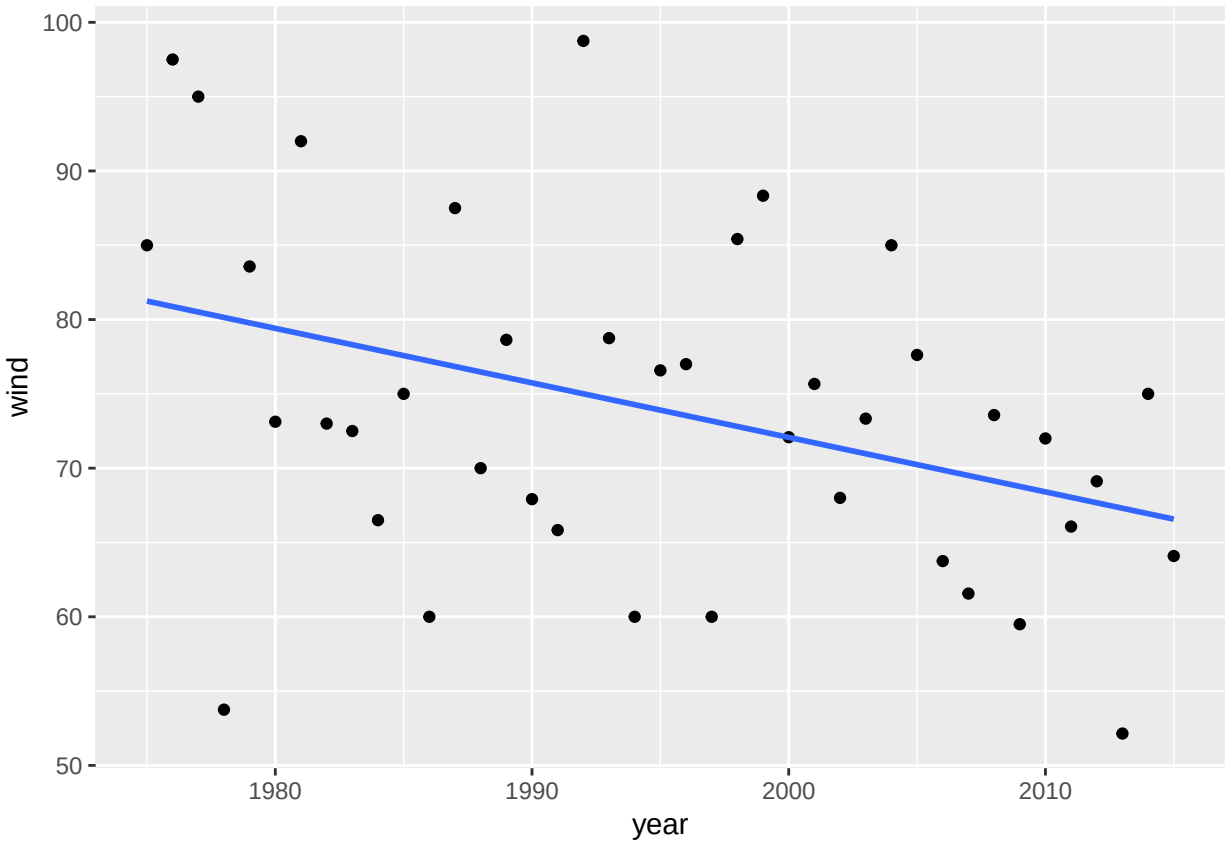
Cross-Sectional

Cross-sectional studies can provide an understanding of current levels of exposure and disease burden; however, they are poor tools for assessing causality as the temporal relationship between the variables is unclear in most cases. Immediate relationships between variables may be suggestive of relationships but other designs are better suited to characterizing that relationship in causal terms.

Ecological

This design is even weaker than a cross-section design. Whereas a cross-sectional design typically features "micro" data, an ecological design features only summary data between places and times.

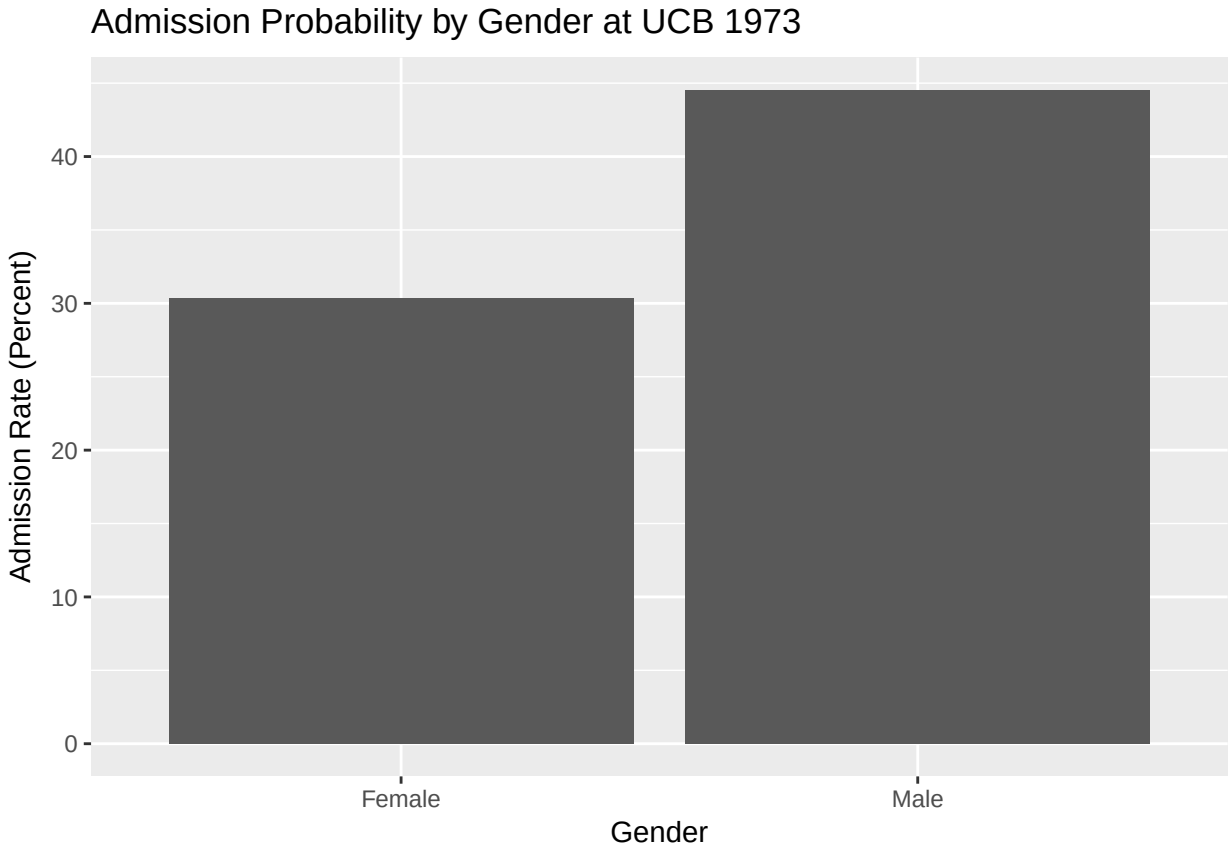
For example, if we look at all named hurricanes in the US from 1975 (a total of 388 storms) and consider the average max wind speed per year, we see a decreasing relationship between wind speed and time.



Indeed, this effect is statistically significant. On average, the max wind speed in a named storm decreases by 0.37 mph per year with a p value of 0.013.

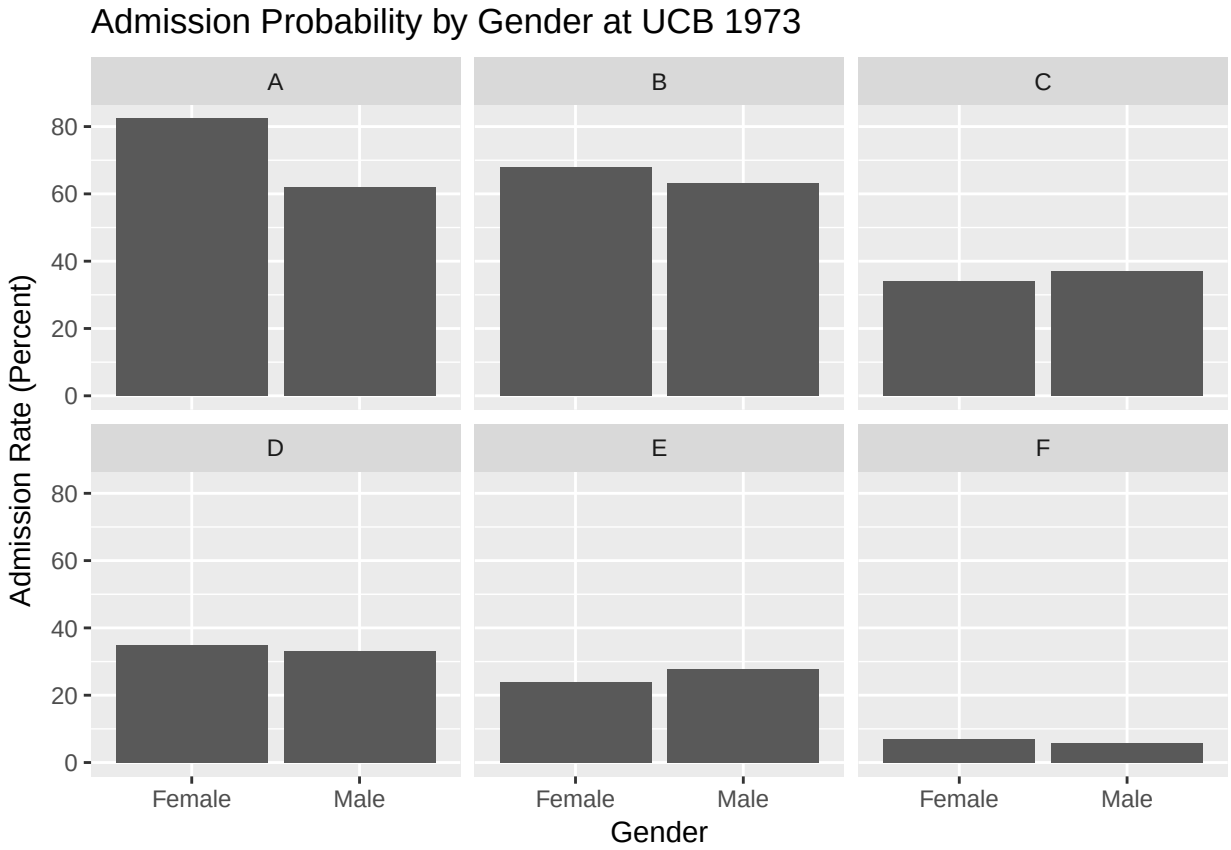
However, the use of the ecological data may be misleading. In this particular example, it is not overly so - using the microdata, there is a decrease by 0.31 mph per year with a p-value of 0.031. However, that is not always the case.

Data at UC Berkeley was collected in 1973 regarding the gender of those who applied and were accepted into the six largest graduate programs. If you aggregate the data to the level of the University, you see that the odds of admission for men were 1.84 times larger than the odds of admission for women.



Pretty compelling evidence of a gender bias in favor of male applicants in admission. Except this does not hold on the micro-level.

If we break down the data by which Department the applicant applied to there is only one case where there appears to be a gender bias - and that bias is actually in favor of female applicants.



What happened here? This is most commonly known as Simpson's paradox. In this particular case, female applicants were more likely to apply to programs with low acceptance rates while male students were more likely to apply to programs with high acceptance rates.

After correcting for the differences in the acceptance rates of the different departments, the evidence for an overall gender preference is eliminated (an OR of 0.90 in favor of admitting women over men, $p = 0.22$).

For these reasons, as well as the reasons that limit cross-sectional work more generally, ecological studies are weak designs for showing causality.

Pre/Post

Pre-post designs involve measuring the variable of interest before and after an exposure or intervention. These are frequently used designs but may suffer from limitations.

First, a pre-post design without a control group provides no protection against any observed change being the result of other non-intervention causes.

Second, recruitment based on eligibility (such has high blood pressure) may pose significant risk of regression to the mean occurring. In a blood pressure intervention (an RCT), there was a double digit drop in the blood pressure of the controls and and slightly larger drop among the intervention arm. If the control was omitted, the drop among the control arm - which may be due to increased treatment or regression to the mean - would have not been known and the estimated change in the blood pressure would have been massively overstated.

Additionally, a simple pre-post design may not sufficient account for pre-existing trends.

In general, an analysis with significant amounts of data on both sides of the intervention and with a control group is a reasonably powerful pre-post design while designs with single measurements or without a control group are much weaker.

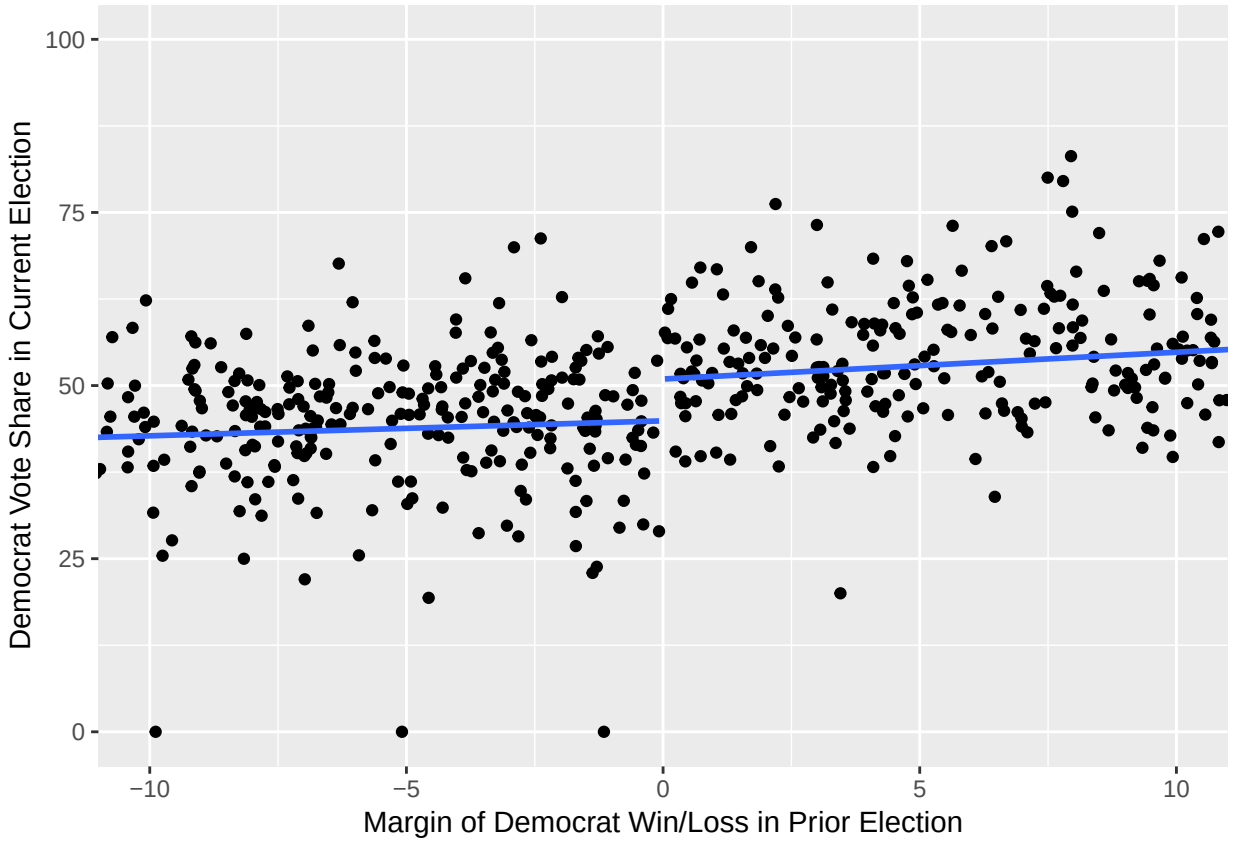
Interrupted Time Series/Regression Discounity

An interrupted time series analysis is functionally the same thing as a pre-post design with with a longer series of points on either side of the "interruption".

Regression discontinuity, which is not covered in the book, is superficially similar and a powerful method.

Suppose you are interested in the effect of crossing some threshold that is somewhat arbitrary (e.g., admitted to a school, having hypertension compared to near hypertension). The threshold is fixed - you either are or are not above the threshold - but observations near the threshold on either side are likely quite similar. Essentially, when we measure or score something, we observe the true value plus some noise. When we apply a discrete filter to those observations with noise, the effect of the noise is to functionally "randomize" marginal people to one side or another.

I'm going to use a relative simple, non-health example here for illustration of the general method. The plot below shows the results of elections for the US Senate from 1914-2010 where the x-axis is margin of victory/loss for the Democrat in the prior election and the y-axis is the share of the vote won by the Democrat in the current election. When the margin (x-axis) is greater than zero, the Democrat is an incumbent. The difference between the vote share at just below a margin of zero and just above a margin of zero is the effect of being an incumbent.



You'll note that the two regression lines fit to the data above/below the margin of zero do have have the same value at a margin of 0. This jump in vote share is the causal benefit of being the incumbent relative to the challenger. Analysis of the data suggests that near incumbents have a 7.4 percentage point increase in their vote share in the next election compared to near non-incumbents (95% CI: 4.1 to 10.9).

The primary considerations in an regression discontinuity design are

1) Is it truly the case that those at $c + \epsilon$ and $c - \epsilon$ are similar?
2) How much data should I use on either side of the cutoff to fit the local models?

The first question is not a statistically one but must be answered with domain knowledge. Statistics may provide some insight by comparing the populations on other factors but human judgement will be required.

The second question can be reduced to a statistical question, but may become rapidly complicated. The `R` and Stata packages `rdrobust` implement methods for selecting these windows, fitting the piece-wise models and then estimating the difference between the models at the cutoff, including adjustments for the standard errors.