Epidemiological Methods I

Jacob Simmering, PhD

February 26, 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

Next Week

- Bias
- Confounding
- Threats to validity

Review of Experimental Methods

The causal effect of a treatment or exposure is of primary interest to us in most of our research projects

Does the treatment reduce the intensity of disease? Does the exposure increase risk of a certain outcome?

Review of Experimental Methods

The causal effect of a treatment or exposure is of primary interest to us in most of our research projects

Does the treatment reduce the intensity of disease? Does the exposure increase risk of a certain outcome?

In other words:

$$E(Y|T = 1) - E(Y|T = 0)$$

Fundamental Problem of Causal Inference

With a given subject s, we can only observe one outcome:

Either you are or are not treated

Either you are or are not exposed

If we can only observe one outcome per subject, how can we estimate the effect of a variable?

Randomization

If we randomly assign people to be treated or not treated (exposed or not exposed), then we can easily estimate the causal difference.

On average, the only difference between the groups is the treatment, which was randomly assigned, and so the differences in the outcome in the two groups is the same as the causal effect of the variable that we randomized on.

This is a very powerful method for assessing the effect of some variable but it is not always suitable.

Randomization is Not Always Possible

While randomization may be a "gold standard" for showing causal effects, it is not always possible for both ethical and logistical reason

Note that the The "gold standard" claim of RCTs have been challenged by some, see Deaton A, Cartwright N, "Understanding and misunderstanding randomized controlled trials" in *Social Science & Medicine*, 2017

Randomization is Not Always Ethical

Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials

Gordon C S Smith, Jill P Pell

Abstract

Objectives To determine whether parachutes are effective in preventing major trauma related to gravitational challenge.
Design Systematic review of randomised controlled trials.
Data sources: Medline, Web of Science, Embase, and the Cochrane Library databases; appropriate internet sites and citation lists.
Study selection: Studies showing the effects of using a parachute during free fall.
Main outcome measure Death or major trauma, defined as an injury severity score >15.
Results We were unable to identify any randomised controlled trials of parachute intervention.
Conclusions As with many interventions intended to prevent ill health, the effectiveness of parachutes has not been subjected to rigorous evaluation by using randomised controlled trials. Advocates of evidence

accepted intervention was a fabric device, secured by strings to a harness worn by the participant and released (either automatically or manually) during free fall with the purpose of limiting the rate of descent. We excluded studies that had no control group.

Definition of outcomes
The major outcomes studied were death or major trauma, defined as an injury severity score greater than 15.[6]

Meta-analysis
Our statistical apprach was to assess outcomes in parachute and control groups by odds ratios and quantified the precision of estimates by 95% confidence intervals. We chose the Mantel-Haenszel test to assess heterogeneity, and sensitivity and subgroup analyses and fixed effects weighted regression techniques to explore causes of heterogeneity. We selected a funnel plot to assess publication bias visually and Egger's and Begg's

Department of Obstetrics and Gynaecology, Cambridge University, Cambridge CB2 2QQ
Gordon C S Smith *professor*
Department of Public Health, Greater Glasgow NHS Board, Glasgow G3 8YU
Jill P Pell *consultant*
Correspondence to: G C S Smith gcss2@cam.ac.uk

BMJ 2003;327:1459–61

Figure 1

Randomization is Not Always Practical

Randomization is poorly suited to questions about

- effects with long latent periods
- rare outcomes

due to the massive scale required to follow up and maintain a sample size of sufficient power

Randomization is Not Always Desirable

Randomized trials have (relatively) high costs and so generally the population in the trial is selected to maximize the trials ability to find a significant effect at the lowest possible sample size, time and cost

As a result, complicated patients, old patients and others are frequently excluded from these trials

Internal validity is optimized by randomized, but external validity and applicability is sometimes unclear

Randomization is Not Always Desirable

What is the validity of an estimate about a drug for treating diabetes in a trial that excluded anyone with heart disease given that many people with diabetes also have heart disease?

Real world treatment effects may differ significantly from the treatment effects observed in controlled trials

Inference Without Randomization

Without random assignment, there is concern that the assignment process may be related to the expected outcome or both to the exposure and outcome

- A study of coffee drinking and cancer may be confounded by the fact many people (used to) smoke while drinking coffee
- A study of the effect of a headache medication on headache intensity after 60 minutes may be confounded if people who sleep less are more likely to get headaches and also more likely to medicate them

All Is Not Lost

Without randomization, different study designs are called for to manage and characterize observational research

The two major designs are

- cohort studies
- case-control studies

We will cover these and several less common designs in detail today

Cohort Studies

In a nutshell, cohort studies

- recruit based on exposure status
- determine outcome status
- good for rare exposures
- bad for rare diseases

Cohort Studies

Exposure or treatment is known as the time of enrollment and subjects can be selected specifically for their exposure/treatment status

At the time of the enrollment, the outcome state for a subject may or may not be known

Cohort Studies

Studies where the outcome is not yet realized at the time of enrollment but subjects are followed for the development of the outcome are considered prospective studies

Studies where the outcome is realized at the time of enrollment are considered retrospective studies

Cohort Studies

In general, prospective studies are considered to have greater validity as they avoid potential issues with temporal misordering of exposure and outcome onset but take longer to conduct

Relative Risk

The relative risk is the primary measure of association in cohort studies and compares the risk of developing the disease/outcome of interest in the exposed and unexposed cohorts

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

Relative Risk

The relative risk is the primary measure of association in cohort studies and compares the risk of developing the disease/outcome of interest in the exposed and unexposed cohorts

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

Relative Risk

With the standard 2x2 table:

	Diseased	Not Diseased
Exposed	a	b
Not Exposed	c	d

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

When to Use a Cohort Study

- Rare exposure
- Want to directly estimate risk/relative risk
- Want to collect data prospectively

When Not To Use a Cohort Study

- Rare diseases

Case-Control Studies

In a nutshell, case-control studies

- recruit based on outcome status
- determine exposure status
- good for rare diseases
- bad for rare exposures

Case-Control Studies

Outcome status is known at the time of enrollment and subjects are selected specifically for their outcome status

Exposure status is then determined

Case-Control Studies

All case-control studies are retrospective since the outcome must be realized at the time of enrollment

There is concern about the potential for different rates of recall (or misrecall) of exposures by case/control status

There are greater issues with determining temporal ordering than in prospective cohort studies

Case-Control Studies Studies

In general, case-control studies are considered to be more sensitive to bias and confounding than cohort studies

Awareness of this potential, the nature of your data and the statistical methods you use can manage this risk

Odds Ratio

Since the rate of disease in a case-control study is determined by the researcher, relative risks cannot be calculated

However, the odds ratio is not sensitive is insensitive to the type of sampling and therefore can be used in case-control studies

Odds

The odds of an event are related to the probability of an event E as:

$$\text{Odds}(E) = \frac{P(E)}{1 - P(E)}$$

Typically this is expressed as "X to Y for/against"

Odds

An event with $P(E) = 0.25$ is 1 to 3

An event with $P(E) = 0.75$ to 3 to 1

Odds Ratio

The odds ratio is the ratio of the odds of having the outcome of interest in the exposed and unexposed subjects in the case-control study

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

Odds Ratio

The odds ratio is the ratio of the odds of having the outcome of interest in the exposed and unexposed subjects in the case-control study

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

Odds Ratio

The odds ratio is the ratio of the odds of having the outcome of interest in the exposed and unexposed subjects in the case-control study

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

Odds Ratio

When using the standard 2x2 table:

	Diseased	Not Diseased
Exposed	a	b
Not Exposed	c	d

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

Odds Ratio

Compared to the probability and relative risk discussed in the cohort study, the odds and odds ratios are actually rather unintuitive

It is important to note that an OR does not describe changes in risk (which typically implies probability) but rather changes in the odds of the outcome

Rare Disease Assumption

Under certain circumstances, we note that

$$RR \simeq OR$$

Specifically, when the incidence of the disease is relatively rare in the population

This is known as the rare disease assumption

Rare Disease Assumption

If it the case that $N(E = 1, O = 1) \gg N(E = 1, O = 0)$, then

- $$N(E = 1, O = 0) \simeq N(E = 1, O = 0) + N(E = 1, O = 1)$$
- $$N(E = 0, O = 0) \simeq N(E = 0, O = 0) + N(E = 0, O = 1)$$

Rare Disease Assumption

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

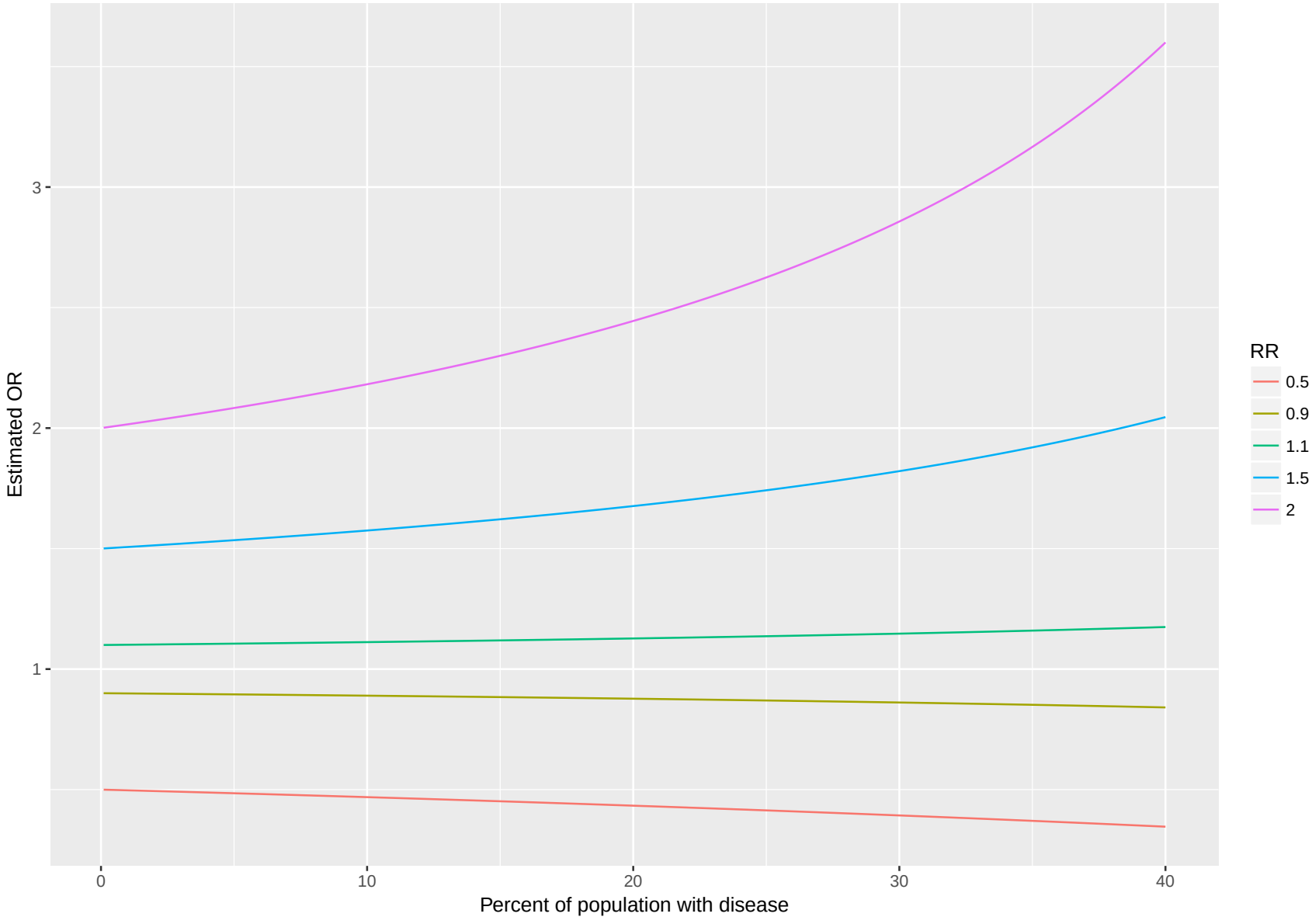
$$\frac{N(E=1, O=1)/N(E=1, O=0)}{N(E=0, O=1)/N(E=0, D=0)} = OR$$

Rare Disease Assumption

When the disease is rare, the $RR \approx OR$

How rare is rare? Generally, under 5-10% of the population results in a reasonably small bias in treating the OR like the RR

Rare Disease Assumption



When to Use a Case-Control Study

- Rare diseases and outcomes
- Costly measurements required (e.g., biomarkers)

When not to Use a Case-Control Study

- Rare exposures

Cohort and Case-Control Summary

	Cohort Studies	Case Control Studies
Grouped by	Exposure Status	Outcome Status
Discovered	Outcome Status	Exposure Status
Measure	Relative Risk	Odds Ratio
Useful for	Rare Exposures	Rare Diseases
Poor for	Rare Diseases	Rare Exposures
Timeline	Both pro and retrospective	Only retrospective

Other Commonly Used Designs

While cohort and case-control studies are used for most observational research, other designs may be more useful depending on the question. Some of these designs are

- Nested case-control
- Case crossover
- Case-time-control
- Cross sectional
- Ecological
- Pre/Post
- Interrupted time series and regression discontinuity

Nested Case-Control

Nested case-control studies are case-control studies done where the cases and controls are sampled from a larger cohort or similar study

The advantages of such a design include easily identifying potential subjects

This type of design is most useful when you are doing a cohort study and are interested in a secondary question, especially if the exposure of interest requires significant investment in time or money to measure

Measuring the exposure on controls exhibits diminishing marginal returns of information

Case-Crossover

Case-crossover is useful when studying transient and short exposures and acute, rapidly detected and short-lived outcomes

Frequently used when studying the effects of weather on health

Controls are generated by using the case by offset by a some time

Case-Crossover

For instance, if you were interested in the effects of weather on the risk of the flu, you might find all cases of the flu and then look at the weather immediately before the subject got the flu and then compare that to the weather immediately before the same date last year, two years ago and so on when the case did not develop the flu

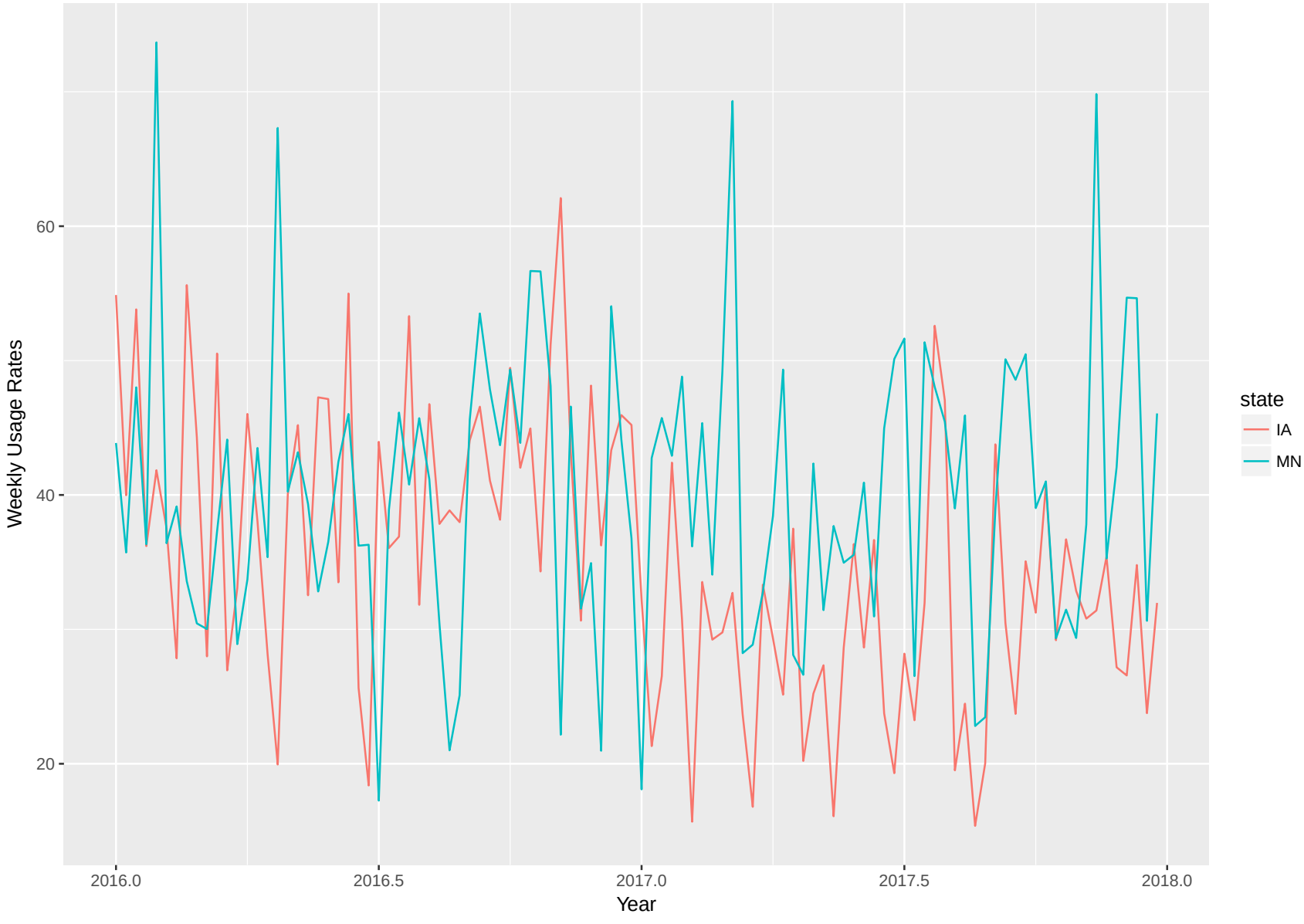
The design implicitly matches on seasonal factors and as the case serves also serves as the control many subject factors are also controlled

Case-Time-Control

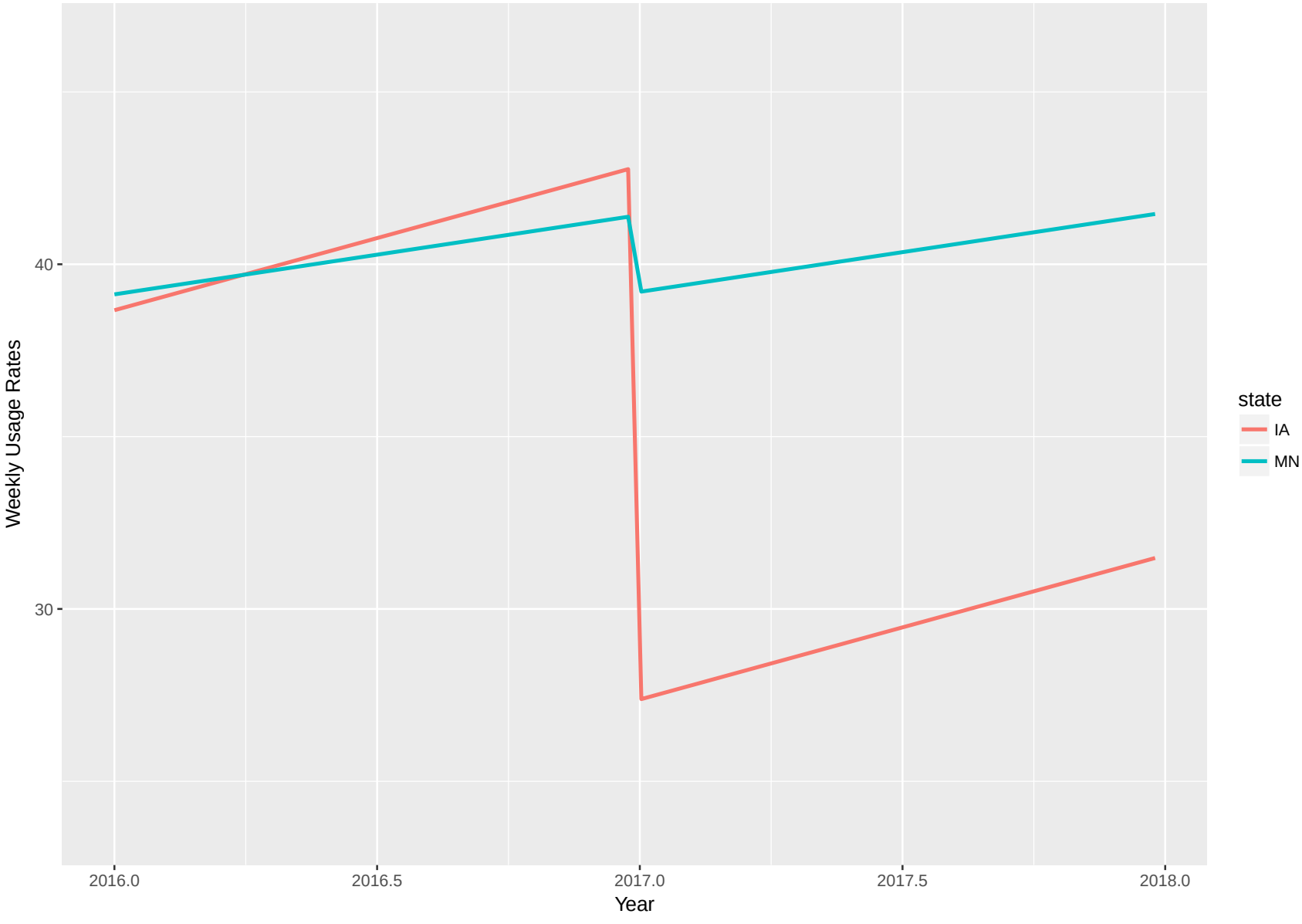
(Diff in diff?)

Describe the differences between a control group and an exposed group over time, especially after an intervention or change in policy

Case-Time-Control (Graphically)



Case-Time-Control (Graphically)



Cross-Sectional

Current and immediate levels of exposure and disease burden

Relatively weak form for causal inference but may suggest future research questions

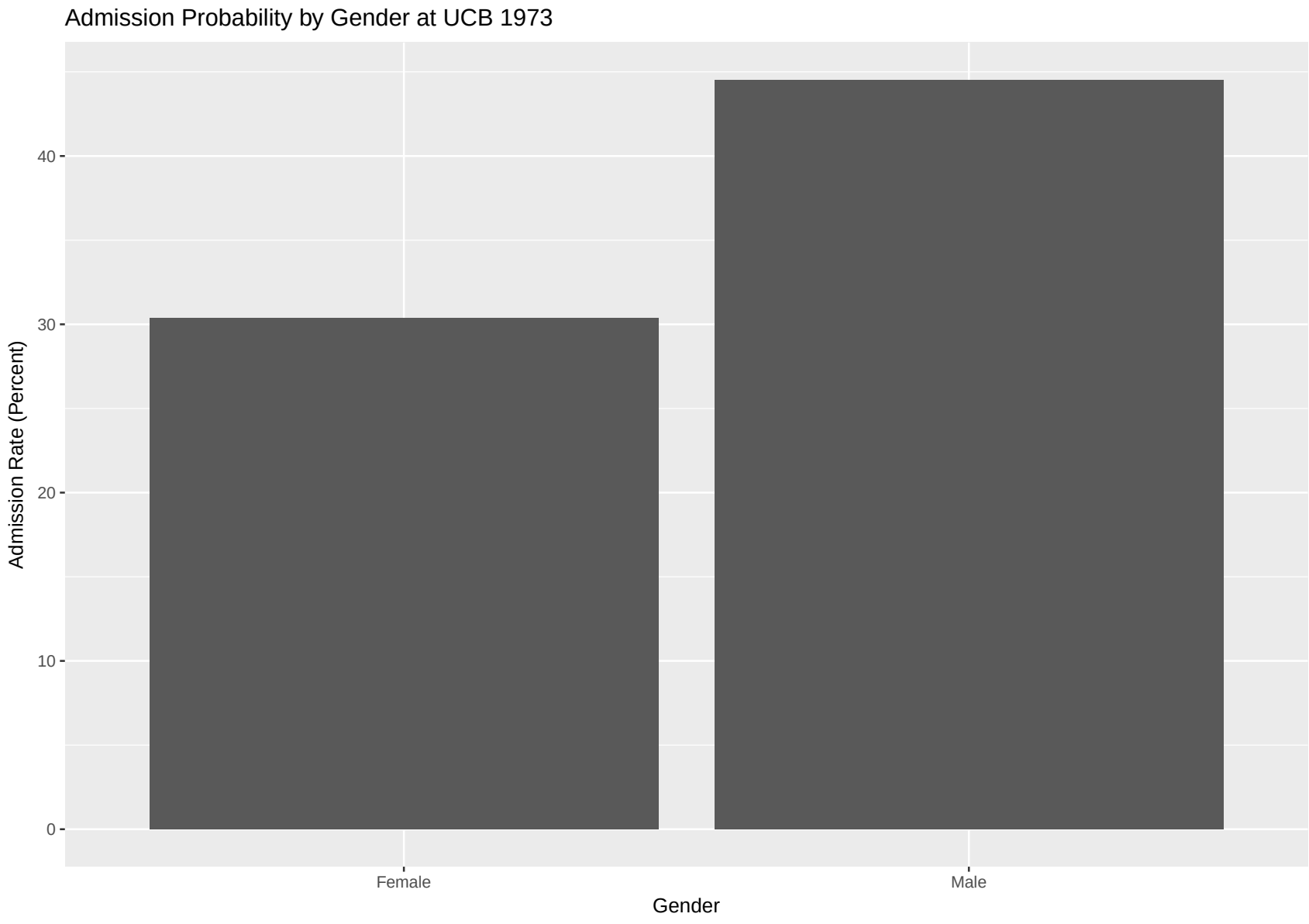
Ecological

Similar to a cross-sectional design but with the data aggregated

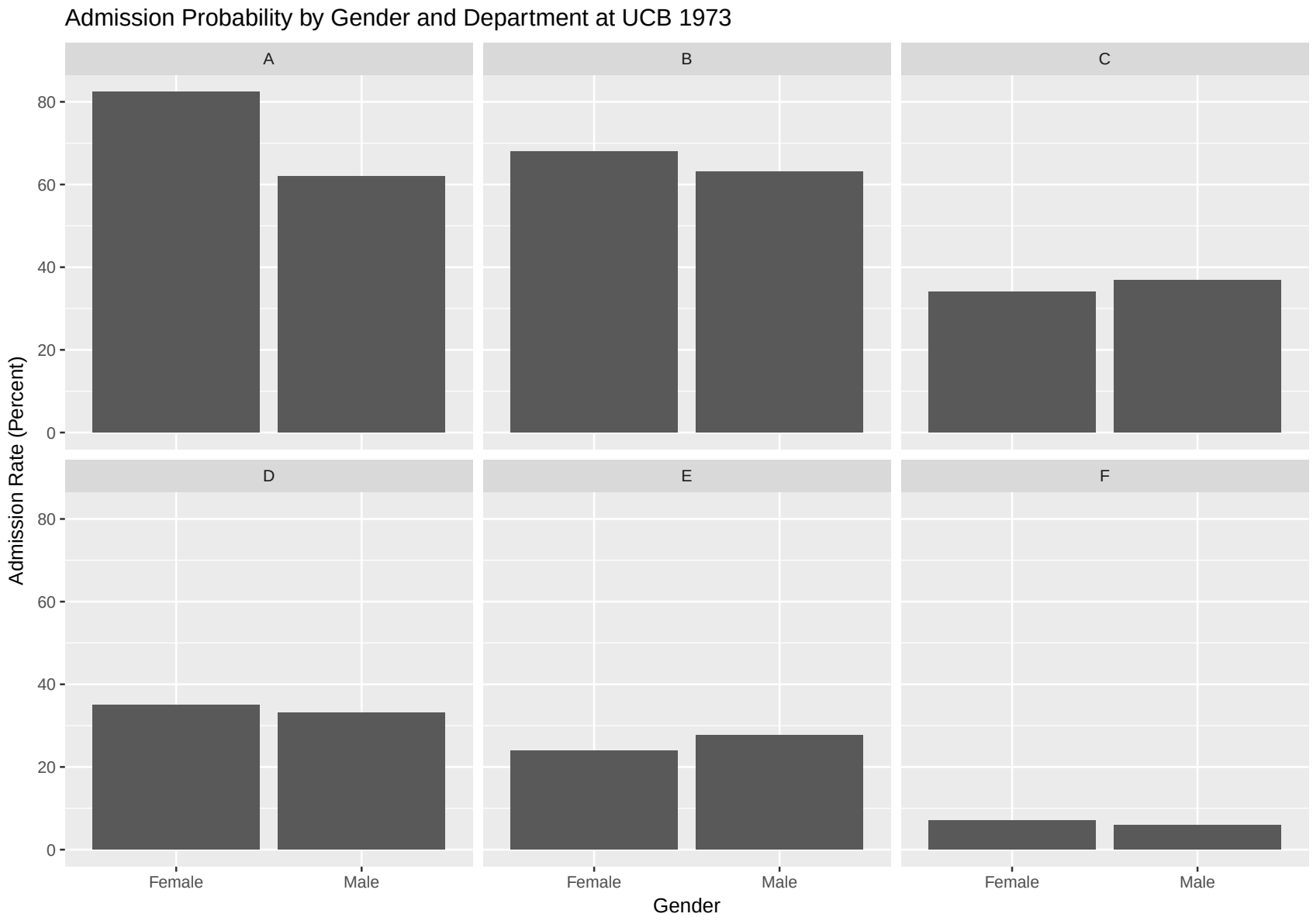
Correlations between two variables summarized at some unit level (state, county, etc)

Potentially misleading summary measures are possible

Gender Preference in Admissions at UCB in 1973



Gender Preferences in Admissions at UCB in 1973



Gender Preferences in Admissions at UCB in 1973

Unadjusted, pooled analysis suggested men had odds of admission that were 1.8 times greater than those of female applicants

On a department-by-department basis, there is no evidence of preference for males - indeed the only unbalanced department actually shows a preference for females

Female applicants were more likely to apply to highly selective and competitive departments whereas male applicants were more likely to apply to departments that were less competitive and tended to take many/most well-qualified applicants

Ecological

Ecological studies that use summary measures may provide misleading results as the summary measures may be misleading due to confounding

Poor form of showing causality, potentially hypothesis generating

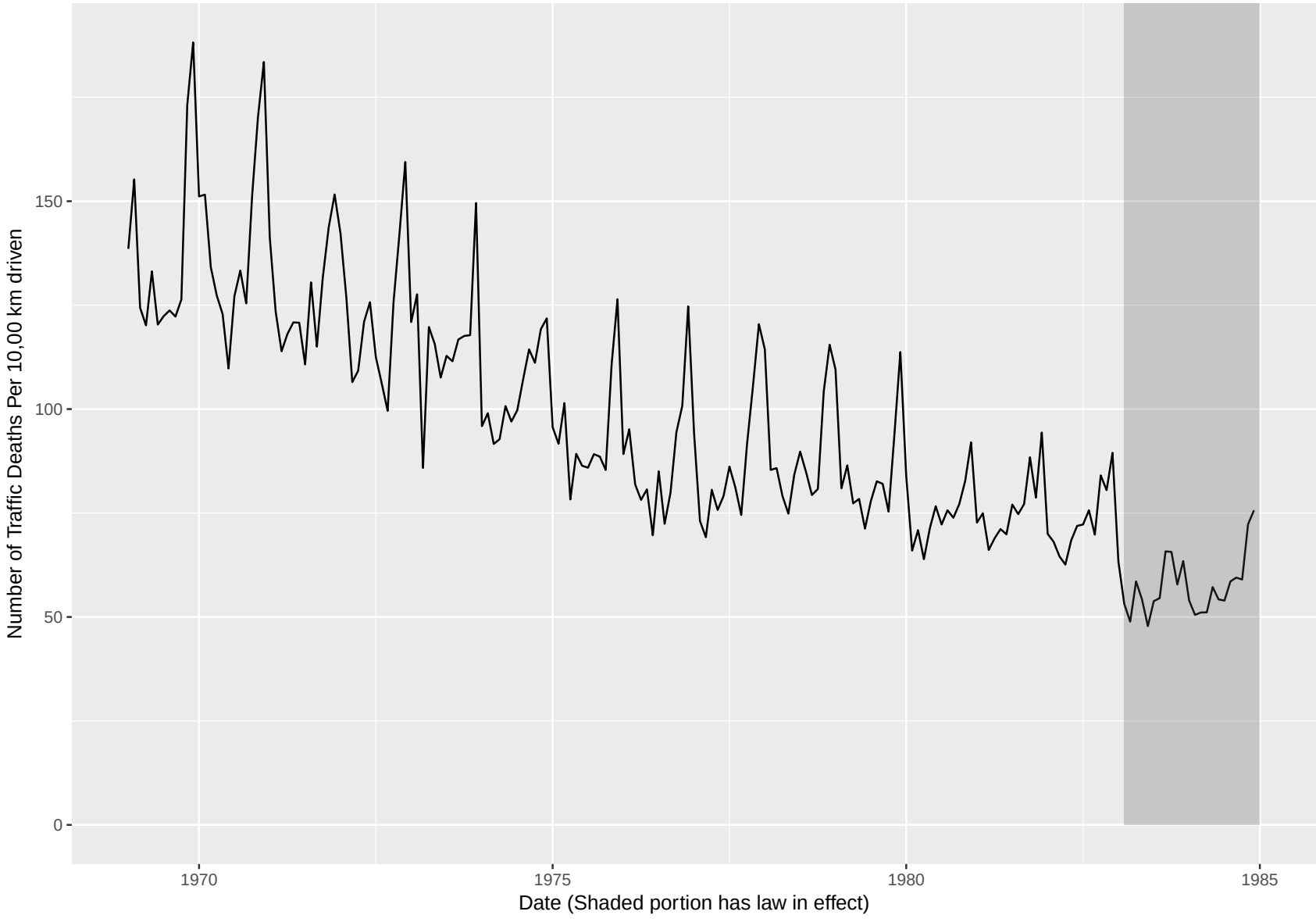
Pre/Post

Measure the outcome of interest before and after an intervention and compare the values

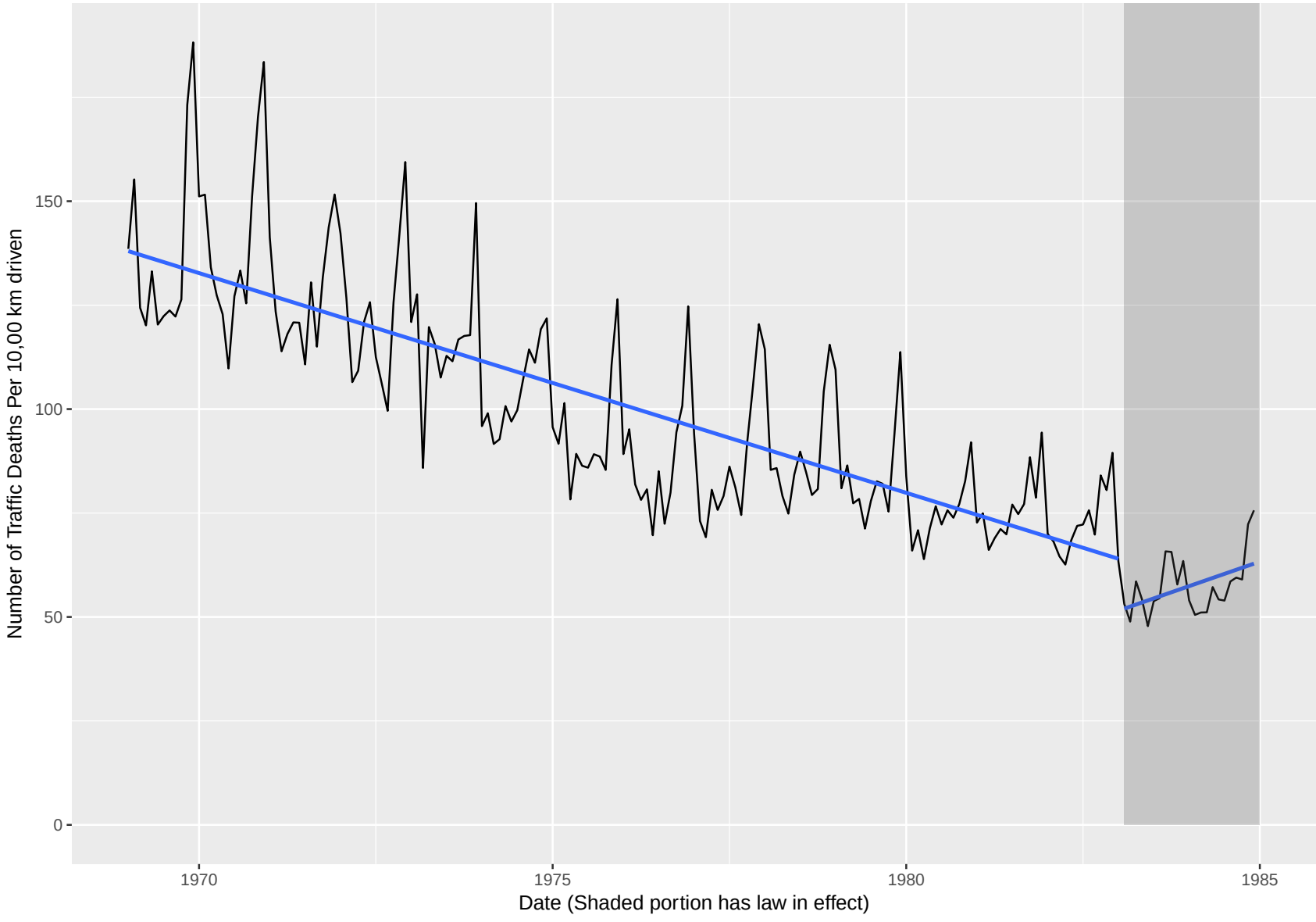
May be done without a control group but at risk to bias due to regression to the mean and failure to control for natural, potentially pre-existing processes

Ideally, you would want a series of several data points on either side of the intervention to have an assessment of how variable the data generating process is and a control group as a measure of any other effects that may be occurring

Interrupted Time Series/Regression Discontinuity



Interupted Time Series



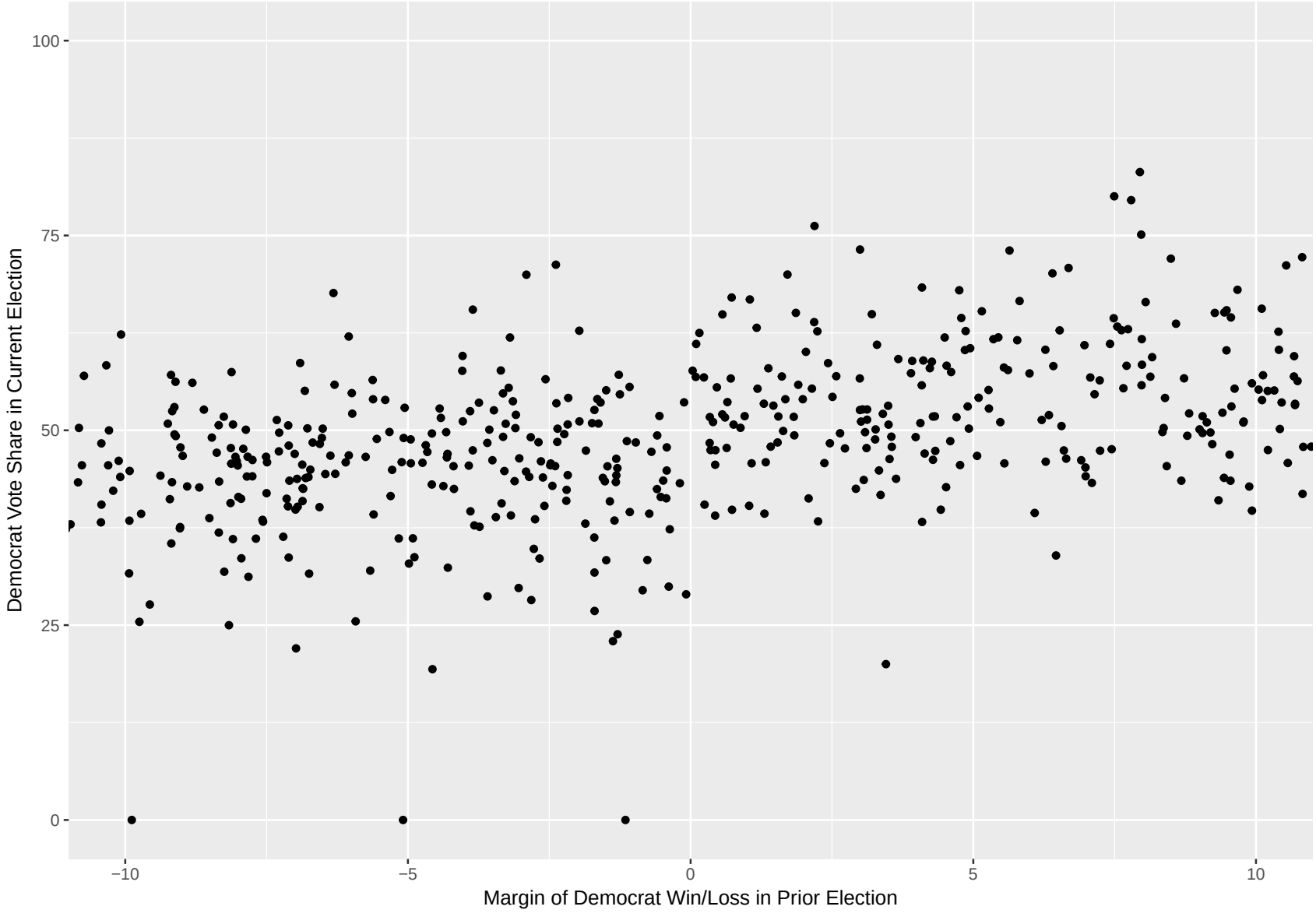
Regression Discontinuity

Imagine there is some variable that has some threshold:

- Blood pressure over 140/90
- Admitted to school
- Treated

And you want to know what is the causal effect of being above the threshold compared to being below

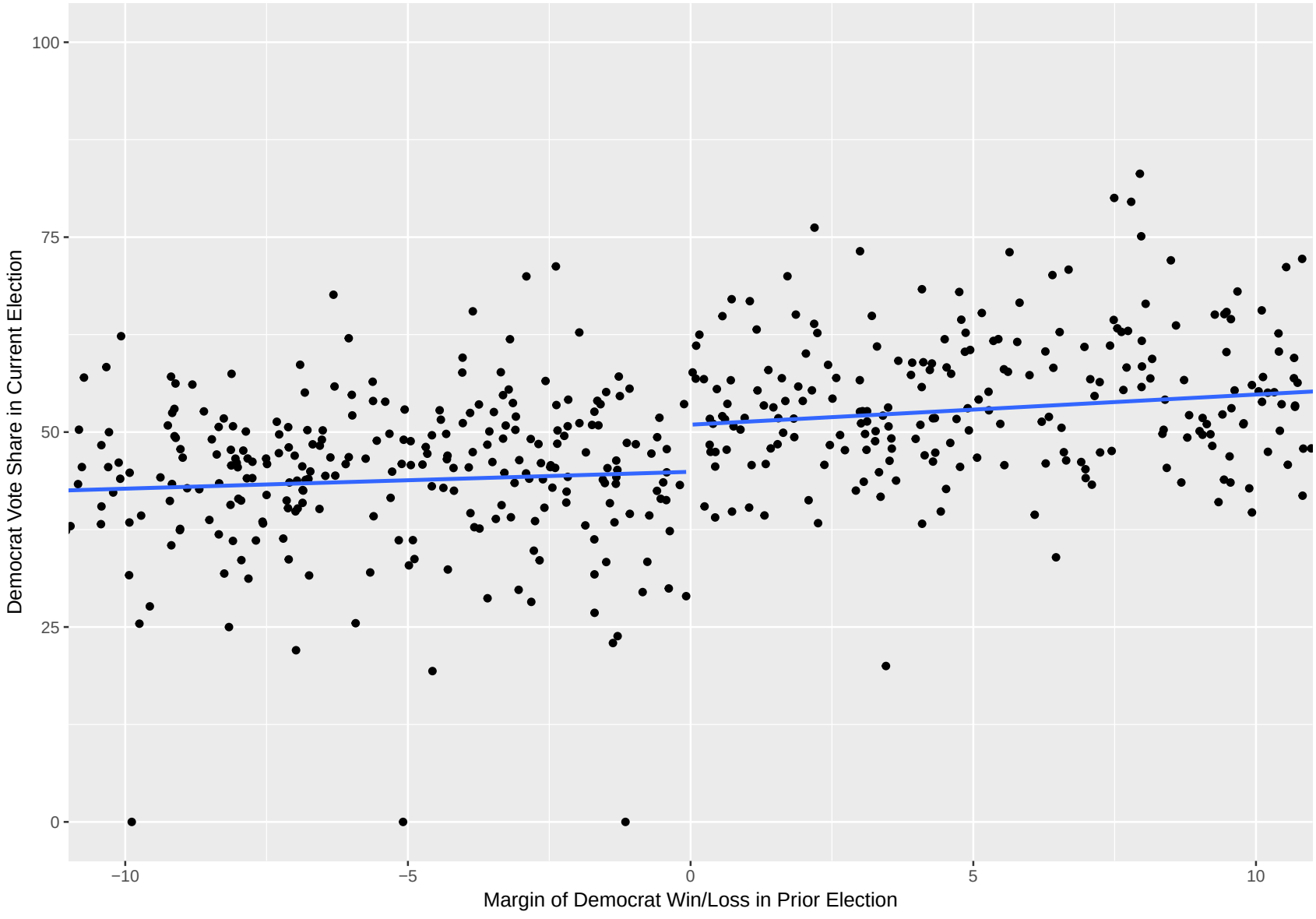
Regression Discontinuity



Regression Discontinuity

Fit a regression model to the data on either side of the cutoff c and compare the difference in the fitted values as the models approach c

Regression Discontinuity



Regression Discontinuity

Two major considerations

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