POL332: Using Data to Understand Politics

Shamel Bhimani

Fall 2025

Contents

1 Introduction to Causality

1.1 Chapter I – Kosuke Imai. Quantitative Social Science: An Introduction. Princeton: Princeton University Press, 2017.

1.1.1 Introduction to Causality

Experimental Data: examines how a treatment causally affects and outcome by assigning varying values of the treatment variable to different observations, and measuring their corresponding values of the outcome.

Contingency Table: Summarizes the relationship between the treatment variables and the outcome variable.

Binary Variable/Dummy Variable: Takes the value of 1 if a condition is true and 0 if the condition is false. The sample of a binary variable equals the sample proportion of 1s. This means that the true observations can be conveniently calculated as the *sample mean*, or *sample average*.

To calculate the sample mean:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

Where:

x_i represents each individual value or data point in the sample;
n represents the total number of observations or data points in the sample.

1.1.2 Causal Effects and the Counterfactual

Causal inference is the comparison between the factual and the counterfactual, i.e., what actually happened and what would have happened if a key condition were different. Unfortunately, we would never observe this counterfactual outcome, because changing one key variable and keeping the rest the same may, in some cases, affect internal validity.

For each observation i, we can define the **casual effect** of a binary treatment T_i as the difference between two potential outcomes, $Y_i(1) - Y_i(0)$, where $Y_i(1)$ represents the outcome that would be realized under the treatment condition ($T_i = 1$) and $Y_i(0)$ deontes the outcome that would be realized under the control condition ($T_i = 0$).

The **fundamental problem of causal inference** is that we observe only one of the two potential outcomes, and which potential outcome is observed depends on the treatment status. Formally, the observed outcome Y_i is equal to $Y_i(T_i)$.

This simple framework of causal inference also clarifies what is and is not an appropriate causal question. Characteristics like gender and race, for example, are called *immutable characteristics*, and many scholars believe that causal questions about these characteristics are not answerable. In fact, there exists a mantra which states, "No causation without manipulation". However, immutable characteristics *can* and have been studied. Instead of tackling the task of directly estimating the causal effect of race, researchers use *perception scores* of the unit of analysis.

1.1.3 Randomized Controlled Trials

In a **randomized controlled trial (RCT)**, each unit is randomly assigned either to the treatment or control group. This randomization of treatment assignment guarantees that the average difference in outcome between the treatment and control groups can be attriobuted solely to the treatment, because the two groups

are on average identical to each other in all pretreatment characteristics.

Sample Average Treatment Effect: is defined as the sample-average of individual-level causal effects (i.e., $Y_i(1) - Y_i(0)$). Formally, in the potential outcomes framework:

Let $Y_i(1)$ = potential outcome for unit i if treated;
Let $Y_i(0)$ = potential outcome for unit i if untreated;
The individual treatment effect is:

$$\tau_i = Y_i(1) - Y_i(0)$$

The Sample Average Treatment Effect (SATE) is then:

$$SATE = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$$

where n is the sample size.

The SATE is not directly observable. For the treatment group that received the treatment, we observe the average outcome under the treatment but do not know what their average outcome would have been in the absence of treatment for the same unit (the fundamental problem of causal inference). The same problem exists for the *control group* because this group does not receive the treatment and as a result, we do not observe the average outcome that would occur under the treatment condition.

In order to estimate the average counterfactual outcome for the treatment group, we may use the observed average outcome of the control group. Similarly, we can use the observed average outcome of the treatment group as an estimate of the average counterfactual outcome for the control group. This suggests that SATE can be estimated by calculating the difference in the average outcome between the treatment and control groups, or the *difference-in-means estimator*.

2 Natural Experiments

2.1 Chapter II – Kosuke Imai. Quantitative Social Science: An Introduction. Princeton: Princeton University Press, 2017.

2.1.1 Observational Studies

Although RCTs can provide an internally valid estimate of causal effects, in many cases social scientists are unable to randomize treatment assignment in the real world for ethical and logistical reasons. Here, we consider Observational Studies.

Observational Studies: Researchers simply observe naturally occurring events and collect and analyze the data, without direct intervention.

- In such studies internal validity is likely to be compromised because of possible selection bias.
- External validity is often stronger than that of RCTs.
- Findings are more generalizable.

Cross-Section Comparison Design: More commonly known as a **cross-sectional study**, is a type of observational research that analyzes data from a population, or a representative subset, at a single point in time. It is used to measure the prevalence of an outcome and its associated factors in a specific population.

The important assumption of observational studies is that the treatment and control groups must be comparable with respect to everything related to the outcome other than the treatment.

Confounding Variables: A pretreatment variable that is associated with both the treatment and the outcome variables is called a **confounder** and is a source of **confounding bias** in the estimation of the treatment effect.

Self-selection Bias: Confounding bias due to self-selection into the treatment group is called *selection bias*. Selection bias often arises in observational studies because researchers have no control over who receives the treatment.

- The lack of control over treatment assignment means that those who self-select themselves into the treatment group may differ significantly from those who do not in terms of observed and unobserved characteristics.
- This makes it difficult to determine whether the observed difference in outcome between the treatment and control groups is due to the difference in the treatment condition or the differences in confounders.

In observational studies, the possibility of confounding bias can never be ruled out. However, researchers can try to address it by means of *statistical control*.

Statistical Control: Confounding bias can be reduced through statistical control whereby the researcher adjusts for confounders using statistical procedures. Some methods of statistical control are:

- **Subclassification:** The idea is to make the treatment and control groups as similar to each other as possible by comparing them within a subset of observations defined by shared values in pretreatment variables or a subclass.

In observational studies, the data collected over time are a valuable source of information. Multiple measurements taken over time on the same units are called *longitudinal data* or *panel data*.

- Longitudinal data often yield a more credible comparison of the treatment and control groups than *cross-section data* because the former contain additional information about changes over time.

Before-and-after Design: Examines how the outcome variable changed from the pretreatment period to the post-treatment period for the same set of units. The design is able to adjust for any confounding factor that is specific to each unit but does not change over time. However, the design does not address possible bias due to time-varying confounders.

Difference-in-Differences Design: Extends the before-and-after design to address the confounding bias due to time treends. The key assumption behind the DiD design is that the outcome variable follows a parallel trend in the absence of treatment.

- Under the DiD design, the sample average causal effect estimate is the difference between the observed outcome after the treatment and the counterfactual outcome derived under the parallel time-trend assumption.
- The quantity of interest under the DiD design is called the *sample average treatment efect for the treated (SATT)*.

2.1.2 Sample Average Treatment Effect for the Treated (SATT)

The SATT is the difference between the average outcome of the treated group with the treatment and the average outcome the sample group *would have had* if they had not been treated.

The formula can be expressed as:

$$SATT = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]$$

Where:
$Y_i(1)$ is the potential outcome for individual i if they receive the treatment;

$Y_i(0)$ is the potential outcome for individual i if they do not receive the treatment (the counterfactual);
$D_i - 1$ indicates that individual i is in the treatment group;
$E[.]$ is the expectation operator, which in this context means we are taking the average over the individuals in the specified group.

A key challenge in calculating the SATT is that we can never observe both potential outcomes for the same individual at the same time. We only observe the outcome for the treated group with the treatment. The counterfactual – what would have happened to the treated group without the treatment – is unobserved.

Therefore, to estimate the SATT, we need to find a suitable comparison group of untreated individuals that can serve as a proxy for the counterfactual outcome of the treated group. A common way to express the estimation of SATT using observed data is:

$$SATT = \frac{1}{n_1} \sum_{i=1}^{n} T_i[Y_i(1) - Y_i(0)]$$

Where:
T_i is the binary treatment indicator variable;
$n_1 = \sum_{i=1}^{n} T_i$ is the size of the treatment group.

3 Measurement Bias

3.1 Chapter III

3.1.1 Introduction to Survey Sampling

Survey sampling is one of the main data collection methods in quantitative social science. It is often used to study public opinion and behaviour when such information is not available from other sources. Survey sampling is a process in which researchers select a subset of the population, called a sample, to understand the features of a target population.

It should be distinguished from a *census*, for which the goal is to enumerate all members of the population.

3.1.2 The Role of Randomization

As in the Randomized Control Trials, randomization plays an essential role in survey sampling. We focus on a class of sampling procedures called *probability sampling* in which every unit of a target population has a known nonzero probability of being selected.

- **Simple Random Sampling (SRS):** the most basic form of probability sampling which avoids **sample selection bias** by randomly choosing units from a population. Under SRS, the predetermined number of units is randomly selected from a target population without replacement, where each unit has an equal probability of being selected. The resulting sample is representative of the population in terms of any observed and unobserved characteristics.
  - The sampling is done *without replacement* rather than with replacement so that once individuals are selected for interview they are taken out of the *sampling frame*, which represents the complete list of potential respondents. Therefore, sampling without replacement assigns at most one interview per individual.
- **Multistage Cluster Sampling:** a method used to collect data from a large, and geographically dispersed population. It's a more complex version of cluster sampling and involves selecting a sample in multiple steps or stages.

  - **Primary Stage:** First, the entire population is divided into large groups, called clusters or primary sampling units (PSUs). A random sample of these PSUs is then selected. These clusters are often based on geography, like states or cities.
  - **Secondary Stage:** Next, each of the selected primary clusters is further divided into smaller groups, known as secondary sampling units (SSUs). From these, a random sample is chosen. For example, if the primary clusters were cities, the secondary clusters might be neighbourhoods or school districts within those cities.
  - **Subsequent Stages:** This process can continue for more stages. You can break down the selected secondary units into even smaller groups until you reach a manageable saple size. The final individuals or units selected for the survey are called the ultimate sampling units (USUs).

3.1.3 Nonresponse and Other Sources of Bias

While probability sampling has attractive theoretical properties, in practice conducting a survey faces many obstacles. A sampling frame, which enumerates all members of a target population, is difficult to obtain. In many cases, we end up sampling from a list that may systematically diverge from the target population in terms of some important characteristics. Even if a representative sampling frame is available, interviewing randomly selected individuals may not be straightforward, and result in *nonresponse*.

There are two types of nonresponse:

- **Unit Nonresponse:** A case in which a potential respondent refuses to participate in a survey.
- **Item Nonresponse:** Occurs when a respondent who agreed to participate refuses to answer a particular question.

Both nonresponses can result in biased inferences if those who respond to a question are systematically different from those who do not.
Beyond item and unit nonresponse, another potential source of bias is **misreporting**. Respondents may simply lie because they may not want interviewers to find out their true answers.

- **Social Desirability Bias:** Refers to the problem where respondents choose an answer that is seen as socially desirable regardless of what their truthful answer is. For example, it is well known that in advanced democracies voters tend to report that they participated in an election even when they actually did not, because abstention is socially undesirable.

One way that researchers can address social desirability bias when researching sensitive topics is to use *list experiments*:

- **List Experiments:** Instead of asking a direct question that might lead someone to give a socially acceptable but untrue answer, a list experiment indirectly gathers this information. The core idea is to present two different lists of statements to two randomly assigned groups of respondents:
  - **Control Group:** This group receives a list of non-sensitive statements and is asked to report how many of the statements are true for them, not which ones.
  - **Treatment Group:** This group receives the same list of non-sensitive statements as the control group, plus one additional sensitive statement. They are also asked to report how many of the statements are true for them.
- By comparing the average number of 'true' statements between the two groups, researchers can estimate the percentage of the population for which the sensitive statement is true. The difference in the averages between the two groups is attributed to the sensitive item.
- **Ethical Advantages:**
  - **Plausible Deniability:** Because respondents only report the number of items that are true, they are not directly revealing their stance on the sensitive issue. This provides 'cover' and reduces the potential for embarrassment or negative consequences.

  - **Reduced Psychological Harm:** By not having to directly admit to something sensitive, respondents may experience less stress or anxiety.
  - **Increased Honesty:** This method can lead to more accurate data on topics like illegal activity, prejudice, or stigmatized behaviours, as people are more likely to be truthful when their individual responses are not directly known.

- **Potential Ethical Considerations:**
  - **Complexity and Confusion:** The indirect nature of the questioning can be confusing for some respondents, potentially leading to inaccurate data. Researchers have an ethical responsibility to ensure their instructions are as clear as possible.
  - **Informed Consent:** As with all research, it is crucial that participants provide informed consent. They should understand the general nature of the study, even if the specific sensitive item is not explicitly highlighted beforehand.
  - **Data Interpretation:** Researchers must be careful in how they interpret and present the results. Since the data is an aggregate estimate, it cannot be used to identify individuals with the sensitive characteristic.

- **Other Problems:**
  - **Floor Effects/Ceiling Effects:** Happens when most or all of the respondents in the *treatment group* report that none or all of the sensitive statements are true for them. Similarly, this also happens when most or all of the respondents in the *control group* report that none or all of the non-sensitive statements are true for them.

Another way researchers overcome the problem of social desirability bias, while addressing the floor/ceiling effect is a popular methodology called the *randomized response technique*.

- **Randomized Response Technique:** Researchers use randomization to provide anonymity to respondents. It works by introducing a random element that shields the respondent, making it impossible for the interviewer to know whether the answer reflects the individual's true status regarding the sensitive topic.
  - For example, respondents are asked to roll a six-sided fair die in private without revealing the outcome. They are then asked to answer yes if the outcome of rolling the die was 1, no if 6, and give an honest answer if they outcome was between 2 and 5. Since the probability of each outcome is known, the researchers can estimate the aggregate proportion of honest responses out of those who responded with a yes answer even though they have no way of knowing the truthfulness of individual answers with certainty.

4 Linear Regression

4.1 Linear Regression

A relationship between two variables is best characterized by the following linear model:

$$Y = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} X + \underbrace{\epsilon}_{\text{error term}}$$

Where
Y is the outcome or response variable
X is the predictor variable or the independent variable
ϵ is the error term
α is the intercept – the average value of Y when X is zero

β is the coefficient or slope of X representing the increase in the average outcome associated with a one-unit increase in X.

We use the linear model under the assumption that it approximates *normally distributed data*. Since the value of the intercept and coefficient of the variable are unknown, they must be estimated from the data itself. Once we obtain the estimated values of the coefficient and slope, then we have the *regression line*. We can use this line to predict the value of the outcome variable.

Specifically, given a particular value of a predictor, $X = x$, we compute the *predicted value* of the outcome value using the regression function:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}x$$

Most likely, the predicted value will not equal the observed value. The difference between the outcome/predicted value and the observed outcome is called the *residual error*. We can write the residual as:

$$\hat{\epsilon} = Y - \hat{Y}$$

4.2 Least Squares

The regression line is the 'line of best fit' becausse it minimizes the magnitude of prediction error. To estimate the line's intercept and slope parameters, a commonly used method is that of the *least squares method*. The idea is to choose $\hat{alpha}$ and $\hat{\beta}$ such that together they minimize the ***sum of squared residuals (SSR)***, defined as:

$$SSR = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$$

Where
Y_i is the observed outcome variable for the ith observation
X_i is the predictor or independent variable for the ith observation
$\hat{epsilon_i}$ is the residual for the ith observation
n is the sample size.

The third inequality follows from the second, where we substitute the linear model $\hat{Y}_i$.

The value of the SSR is difficult to interpret. However, we can use the idea of *root mean square* (RMS). Specifically, we can compute the *root mean squared error* as:

$$RMSE = \sqrt{\frac{1}{n} SSR} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i^2}$$

Therefore, RMSE represents the average magnitude of the prediction error for the regression, and this is what the method of least squares minimizes.

The mean of residuals is always zero, and the regression line always goes through the center of data $(\bar{X}, \bar{Y})$ where $\bar{x}$ and $\bar{Y}$ are the sample means of X and Y, respectively.
The least squares estimates of intercept and slope parameters are given by

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X},$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Recall that the sample means of Y and X are given by $\hat{Y} - \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, resprectively. The results imply that the regression line always goes through the ceter of the data $(\bar{X}, \bar{Y})$. This is so because substituting $x = \bar{X}$ into the model equation and using the expression for $\hat{\alpha}$ yields $\hat{Y} = \bar{Y}$:

$$\hat{Y} = \underbrace{(\bar{Y} - \hat{\beta}\bar{X})}_{\hat{\alpha}} + \hat{\beta}\bar{X} = \bar{Y}$$

In addition, when the method of least squares is used to estimate the coefficients, the predictions based on the fitted regression line are accurate on average. More precisely, the mean of residual $\hat{\epsilon}$ is zero, as the following algebraic manipulation shows:

$$mean\ of\ \hat{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}(Y_i = \hat{\alpha} - \hat{\beta}X_i) = \bar{Y} - \hat{\alpha} - \hat{\beta}\bar{X} = 0$$

It is also important to understand the relationship between the estimated slope of the regression and the correlation coefficient:

$$correlation\ of\ X\ and\ Y \cdot \frac{standard\ deviation\ of\ Y}{standard\ deviation\ of\ X}$$

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}} \cdot \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

The estimated slope coefficient from a linear regression model equals the ρ standard deviation unit increase in the outcome variable that is associated with an increase of 1 standard deviation in the predictor, where ρ is the correlation between the two variables.

Regression towards the mean represents an empirical phenomenon where an observation with a value of the predictor further away from the distribution's mean tends to have a value of an outcome variable closer to that mean. This tendency can be explained by chance alone.

4.3 Model Fit

Model fit measures how well the model fits the data, i.e., how accruately the model predicts the observations. We can assess model fit by looking at the *coefficient of determination* or R^2, which represents the proportion of total variation in the outcome variable explained by the model. To define R^2, we first introduce the *total sum of squares* or TSS, which is defined as:

$$TSS = \sum_{i=1}^{n}(Y_i - \hat{Y})^2$$

The TSS represents the total variation of the outcome variable based on the square distance from its mean. Now we can define R^2 as the proportion of TSS explained by the predictor X:

$$R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS}$$

The value of R^2 ranges from 0 (when the correlation between the outcome and the predictor is 0) to 1 (when the correlation is 1), indicating how well the linear model fits the data at hand.