CAUSAL INFERENCE

MILTON LIN

Contents

Date: March 20, 2024.

1. Introduction to this file

References: A very nice notes by Andrew Gelman, and python causality handbook.

In this article, we focus on task of estimating the average causal effect on a basic scenario. The basic problem of causal inference is that it is impossible to know individual treatment effect

$$\tau := Y(1) - Y(0)$$

hence one computes the *average treatment effect*,

$$\mathbb{E}_I[Y_I(1) - Y_I(0)]$$

where I is a distribution over the candidates. This is still hard to estimate, and one can also not compute

$$\mathbb{E}[Y_I|T = 1] - \mathbb{E}[Y_I|T = 0]$$

due to *bias*. In this article, we address this problem using regressional analysis.

In Section 2 we discuss examples of causal relations and draw their causal diagram. This short exercise reflects how simplified our situation is. In Section 3 and Section 4 we will estimate the treatment effect in a semi-synthetic dataset based on the 20 Newsgroup dataset by various forms of regressional analysis. The 20 Newsgroup dataset is a collection of approximately 20K newsgroup documents corresponding to different topics; some of the topics are highly related and some are not, so we relabel the topics according to subject matter in the code.

In the provided `get_data` function, we have the following variables.

- U is defined as whether or not the news document belongs to a specific category.
- Z is the observed variable dependent on U, thought of as the treatment. It is determined by normally distributed random number.

  ```
  Z = int(1.0*U + np.random.normal(loc=0, scale=1) > 0)
  ```

- Y is the outcome. This setup models a scenario where the number of lines read (or any other outcome of interest) is influenced by the medium of consumption and the topic of the document, with the topic serving as a confounder that affects both the medium chosen and the likelihood of reading.

Now we discuss how the change of `Z_bias` across various forms of regression.

- In Section 3 the initial regression without controlling for any confounders, the estimate for Z reflects the raw, unadjusted relation.
- In Section 3.1 we explicitly including the confounder.
- In Section 4 we discuss the situation when there

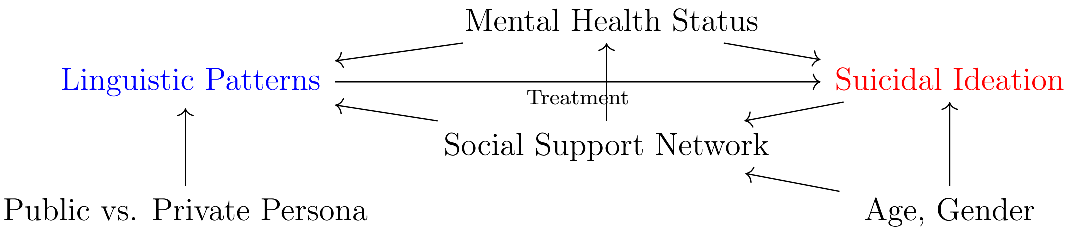
2. Examples of Causal Relations

Text labelled blue are treatments and red are outcomes. Black ones are potential cofounders.

(1) From the population of users tweeting about mental health, what linguistic structures or linguistic patterns differentiates those who proceed to discuss suicidal ideation in the future, from those who do not?

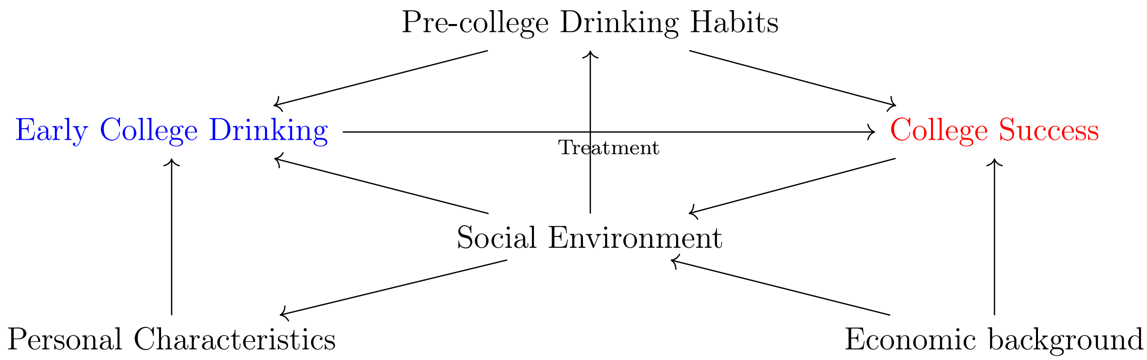
- Treatment: specific linguistic structures or rate of posting.
- Outcome: suicidal ideation.
- Cofounder:
  - Mental health status of the tweeter. Preexisting conditions can affect the likelihood of discussing suicidal ideation.
  - Social support network. The size and quality of an individual's social support network might affect both their communication patterns.
  - The media of choice: the fac that twitter is more of a "social" medium, can affect the rate of posting, or use of language - in general the post here could be more emotional and to elicit public approvals.

The diagram from this



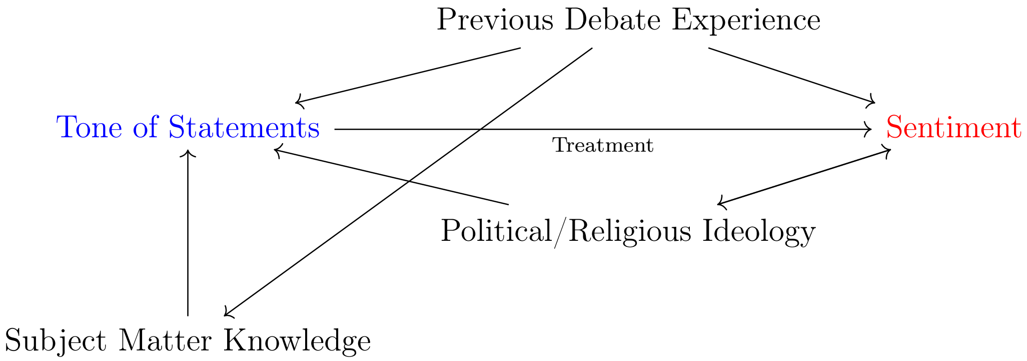
(2) By collecting Reddit timelines from students entering college, we study what effect does drinking early in college have on college success, including habits, social relationships, and even criminal activity, of those who mention drinking during their first semester versus those that do not.

- Treatment: Drinking during the first semester of college.
- College success, which includes habits, social relationships, but most realistic and quantifiable aspect is their grades.
- Cofounders:
  - personality traits and experiences, such as social environment, pre-college drinking habits, and friend groups.
  - A further cofounder is wealth and background - the "drinking habits" can vary drastically depending on background. Do these come from a dining setting or a party setting?



(3) We collect politically charged debates and investigate how the tones of the statements made during the debates changes the linguistic and sentiment characteristics in subsequent responses.

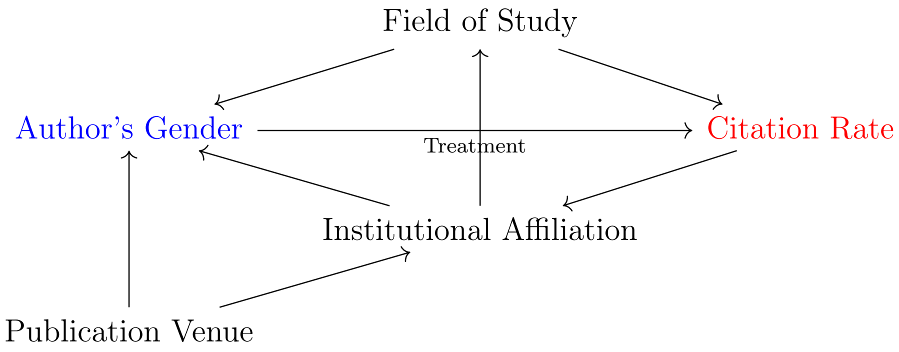
- Treatment: Tone of the statements made during the debates.
- Outcome: Sentiment characteristics in subsequent responses. Can use various possible emotion frameworks for the text analysis.
- Cofounders:
  - Prior experience in debates: can influence both the tone used and how one's statements are interpreted or responded to. Knowledge about the debate topic can also affect.
  - Belief: For example, the underlying political or religious ideology might shape the responses elicited.



(4) If an AI article published under a woman's name and were instead published in the same venue under the name of a man with the same scholarly credentials, would it be cited more?

- Treatment: Gender of the name under which an AI article is published.
- Outcome: Number of citations received by the article.
- Cofounders:
  - Field of study : Some AI subfields might have different citation norms or biases.

- Institution: The prestige of the author's institution could influence both the perceived gender norms and the citation rate.
- Publication venue: The prestige and expertise of audience of the publication might also have its own gender bias.



3. Regressional analysis without confounder

Below we do a simple regressional analysis using ordinary least squares to find the `Z_bias` coefficient. In Section 3.1, we include the cofounder factor U.

Table 1. Regressional analysis without cofounder

	Coef.	Std. Err.	t	$P > \|t\|$	[0.025	0.975]
Constant	-0.2841	0.025	-11.475	0.000	[-0.333	-0.236]
Z	0.8674	0.033	25.945	0.000	[0.802	0.933]
Dep. Variable:	Y					
Model:	OLS					
Method:	Least Squares					
No. Observations:	9816					
Df Residuals:	9814					
Df Model:	1					
R-squared:	0.064					

3.1. Estimating the Treatment Effect by Regressing Y on Z and the Confounder U.

In this model, we see that The U coefficient of 4.9962 shows the justifies a positive influence of U on Y.

Table 2. Coefficients

	coef	std err	t	$P > \|t\|$	[0.0250.975]
const	-0.4993	0.001	-333.148	0.000	[-0.502 -0.496]
Z	0.0512	0.002	24.673	0.000	[0.047 0.055]
U	4.9962	0.003	1639.658	0.000	[4.990 5.002]

4. Controlling for confounders

In the previous section, we had access to the confounder, but in real-life scenarios, this is unobserved. We will use topic modeling. [1]

4.1. Non-negative matrix factorization.

We use NMF (Non-Negative Matrix Factorization) from `sklearn.decomposition` to find topics in the documents. NMF decomposes the TF-IDF matrix into two matrices: (W, H), one representing the documents as combinations of topics and the other representing topics as combinations of words. The goal is to minimize

$$\approx |X - WH|_{\text{loss}}$$

where this loss function is often taken as Frobenius norm. Our results can be summarized below.

[1] Note that before applying topic modeling, you need to convert the text data into a suitable numeric format. TfidfVectorizer from `sklearn.feature_extraction.text` will be used for this purpose, as it transforms the text into a matrix of TF-IDF features.

Table 3. Coefficients

	Coef.	Std. Err.	t	$P > \|t\|$
Constant	0.0639	0.038	1.674	0.094
Z	0.4908	0.026	19.173	0.000
topic_0	2.6269	1.260	2.085	0.037
topic_1	-6.9509	0.925	-7.516	0.000
⋮	⋮	⋮	⋮	⋮
topic_18	7.6653	0.760	10.092	0.000
topic_19	-5.2319	0.578	-9.046	0.000

4.2. **Methods for improvement.** We explore: LDA: a popular method for topic modeling that assumes documents are generated by a mixture of topics, where each topic is a distribution over words. The `get_topics_NMF` and `get_topics_LDA` functions transform the documents into a lower-dimensional topic space using NMF and LDA, respectively. These serve as additional parameter we can put into our model.

A quick experiment shows that LDA with 50 topics does not really get `Z_bias` score of 0.05. Below is table using LDA, with 50 topics. As we see, the score can only get to 0.4569 which is still far from the desired bias coefficient.

Table 4. OLS Regression Results with LDA

Variable	Coef.	Std. Err.	t	$P > \|t\|$	$[0.025, 0.975]$
Constant	-0.1503	0.025	-6.099	0.000	$[-0.199, -0.102]$
Z	0.4569	0.024	18.822	0.000	$[0.409, 0.504]$
Topic_1	-0.8250	0.134	-6.168	0.000	$[-1.087, -0.563]$
Topic_2	0.3076	0.100	3.068	0.002	$[0.111, 0.504]$
⋮	⋮	⋮	⋮	⋮	⋮
Topic_50	-0.9199	0.159	-5.786	0.000	$[-1.232, -0.608]$

4.3. **Inverse Probability Weighting.**

- Data Splitting: The data and documents are split into two equal-sized halves for training and testing to avoid overfitting the propensity score model.
- Logistic Regression for Propensity Scores: A logistic regression model is trained on the TF-IDF features from the training set to estimate the probability (propensity score) of receiving the treatment based on the text data.
- ATE Estimation:

In this section we apply inverse probability weighting to compute (un) adjusted ATE. The unadjusted ATE is simply the difference in mean outcomes between treated and control groups in the test set. The adjusted ATE uses the calculated weights to account for confounders represented by the text data, providing a more accurate estimate of the treatment effect Our results are:

```
DATA: Y = -0.5 + 0.05*Z + 5.0*U
Adjusted 0.4837261658118349
Unadjusted 0.8459253354974212
```

The result is slightly off from the suggested adjusted and unadjusted score - but it is close enough!