HUL315: Econometric Methods

Harshit Goyal: 2021MT10143

January 14^{th} 2024

Assignment 1

Problem 1: Write a short note on the importance of causality in data analytics.

Causality indicates a **cause-and-effect** relationship between variables, in context of **Econometrics**, for quantifying the change in y (effect) due to a change in x (the cause). Looking for causality also reveals **insights**, **patterns**, and **inferences** in data. Applications of these inferences and insights lie in **decision-making**, **policy decisions** etc.

For example, deciding "How much investment in advertising should be done in order to produce maximum increase in sales." and "Should I go to hospital for treatment?". Similarly, causal is very useful for **modeling** involving human choices in Economic, political, sociological, marketing, health, and other domains. However, there are special considerations for properly analyzing these, like **Counterfactuals**, **Confounding Variables**, **Instrumental Variables** and **Selection Bias**. By making a proper analysis, we can do much more than predicting.

Correlation, on the other hand, is a weaker relation than Causality and describes a **statistical** association between variables. Hence, Correlation does not imply causation.

Establishing causality requires rigorous analysis and evidence. Without a focus on causality, **analysts** might misinterpret data patterns that appear correlated but lack a meaningful cause-and-effect connection. Causality forces studying that are the **true causes** of an effect, in other words, that are **relevant** features, which makes better **Predictive Modeling**. Studying causality involves **Controlled experiments**, that needs techniques like **Randomization**. In data analytics, **Experimental Design** helps isolate variables to determine their true impact.

Problem 2: What are the ways in which causal framework can be incorporated into Machine Learning? Discuss.

A Machine Learning model just takes in the **features** and **data**. The **Econometrician** needs to design the right features to be fed into the model, in order to incorporate the causal framework into Machine Learning. Measuring **casual inference** needs important considerations, explained here.

Counterfactuals refer to thinking about what could have happened but did not happen. For example, to measure the effectiveness of **healthcare**, we need to compare the difference in health of people who went to the hospital and that if the same people had stayed home. It may not be possible to do both experiments on the same set of people. Solving this **Selection Bias** needs techniques like **Randomization**.

Confounding variables refer to **unobserved** variables that correlate with both y and x. Solving this requires **domain knowledge**, practice, and experiments. Knowing which features are relevant, they can be fed into the model.

The **right question** depends on the task. In order to make a decision, causal impact (**Ceteris paribus**) is relevant, for example, "Change in sales associated with a change in advertising expenditure everything else held constant?". On the other hand, to make a prediction, one should ask (**Mutatis mutandis**) "Change in sales you would expect to observe when advertising expenditure changes?"

There are techniques for **Data Collection** to measure casual inference. The best way is true randomized **treatment-control** experiment (like Google, Bing etc) but one might need to settle for **Natural experiments** (like draft lottery) which may or may not be randomized.