Estimating the Effect of Internships on Starting Salary

Coding Exercise Module 1

Causal Inference with Linear Regression: A Modern Approach 2025 by CausAI

Introduction

A university seeks to understand whether completing an internship during college has a causal effect on students' starting salaries after graduation. To investigate this, they analyze observational data collected from graduates over the past 5 years.

The dataset contains the following variables:

- **Academic Performance** (`academic_performance`): A categorical variable representing students' academic performance, taking values {Low, Medium, High}.
- **Internship** (`internship`): A binary variable where `internship=1` if the student completed an internship and `internship=0` otherwise.
- **Starting Salary** (`starting_salary`): A continuous variable representing the student's annual starting salary, measured in U.S. dollars.

Your task is to estimate the average treatment effect (ATE) of completing an internship (`internship`) on starting salary (`starting_salary`).

1 Naïve Estimation

1. Download the dataset `internship_salary_data.csv` and load it into Python.
2. Compute the naïve estimate of the ATE by calculating the difference in sample means:

   $\mathbb{E}[\texttt{starting\_salary} \mid \texttt{internship} = 1] - \mathbb{E}[\texttt{starting\_salary} \mid \texttt{internship} = 0]$
3. Suppose we interpret this difference as the true average treatment effect. How would you state the conclusion? What implicit assumption is being made?

2 Identifying Bias

An academic domain expert points out that the naive estimate may be biased because:

- Students with higher academic performance are more likely to complete internships.
- Students with higher academic performance also tend to receive higher salaries, regardless of internship participation.

1. Draw the causal graph that represents the mentioned relationships between the variables.
2. Use the backdoor criterion to determine a valid adjustment set `Z` such that conditional ignorability holds:

   $\texttt{starting\_salary}(1), \texttt{starting\_salary}(0) \perp\!\!\!\perp \texttt{internship} \mid \texttt{Z}$

3 Computing the Adjusted ATE

Now that we have identified an appropriate adjustment set, we will apply the adjustment formula to estimate the true causal effect of internships on starting salary.

1. Compute the sample probabilities `P(Z=z)` for all values of `Z`.
2. Compute the conditional expectations within each subgroup:

   $\mathbb{E}[\texttt{starting\_salary} \mid \texttt{internship} = 1, \texttt{Z} = \texttt{z}] - \mathbb{E}[\texttt{starting\_salary} \mid \texttt{internship} = 0, \texttt{Z} = \texttt{z}]$

   and compute the adjusted ATE as a weighted average:

   $\sum_z \Big(\mathbb{E}[\texttt{starting\_salary} \mid \texttt{internship} = 1, \texttt{Z} = \texttt{z}] - \mathbb{E}[\texttt{starting\_salary} \mid \texttt{internship} = 0, \texttt{Z} = \texttt{z}]\Big)\texttt{P(Z} = \texttt{z)}$
3. Compare the adjusted ATE to the naive estimate. How would you interpret this estimate now?