UNIVERSITY of WASHINGTON

Causal Inference in the Record of the SAT Scores

Hadi Nazirool Bin Yusri
Mentor: Steven Wilkins-Reeves

Main reading material

Causal Inference in Statistics: A Primer

Book by Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell

Citation:

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. , 2016. Internet resource.

Background

- SAT becomes among the most powerful indicators of college admissions to assess academic performance of the students applying
- Main broad question: What affects a student's performance in the SAT scores?
  - eg: *The problem with America's college entrance exam* (2019)* from Vox's youtube media revealed that the SAT also reflects income status of students' families
  - Are there more factors that could affect how students outperform their peers in the SAT exam?

*link: https://youtu.be/WjVVwMGJ9S8

Dataset Details

- Dataset by the CORGIS Dataset Project, recorded in October 2016
- Data was recorded by state from 2005-2015
- Contains a total of 99 distinct variables (income, math/ reading scores, state, year, high school gpa, years of learning foreign languages etc)

- school_scores.csv

Key Descriptions

Key	List of...	Comment	Example Value
Year	Integer	The year of this report.	2005
State.Code	String	The two-letter abbreviation of the state for thsi report.	"AL"
State.Name	String	The full name of the state for this report.	"Alabama"
Total.Math	Integer	The average Math score of students in this state during this year.	559
Total.Test-takers	Integer	The number of test-takers in this state during this year.	3985
Total.Verbal	Integer	The average Verbal (Reading, not Writing) score of students in this state during this year.	567
Academic Subjects.Arts/Music.Average GPA	Float	The average GPA of all students in this state during this year in Arts/Music. Note that this is just the GPA within the subject, not across all academic subjects.	3.92
Academic Subjects.Arts/Music.Average Years	Float	The average number of years that a student has studied Arts/Music when they took the exam.	2.2
Academic Subjects.English.Average GPA	Float	The average GPA of all students in this state during this year in English. Note that this is just the GPA within the subject, not across all academic subjects.	3.53
Academic Subjects.English.Average Years	Float	The average number of years that a student has studied English when they took the exam.	3.9
Academic Subjects.Foreign	Float	The average GPA of all students in this state during this year in Foreign Languages. Note that this is just the GPA within	3.54

Figure 1: Table describing recorded variables in the dataset

dataset link:
https://corgis-edu.github.io/corgis/csv/school_scores/

RESEARCH QUESTION:

Does the number of years a student studies english affect the student's Math SAT scores?

Cause and effect

The intervention of a variable X that changes the distribution of another variable Y possibly show the cause-and-effect relationship between both variables.

Concepts

- Total and Direct Effects

> **To see if the variables are associated with each other either directly or indirectly**
  - **Direct effect: cause-effect relationship between variables WITHOUT mediators**
  - **Indirect effect: cause-effect relationship between variables WITH mediators**

> **How to achieve the calculations?**
  - **Draw the directed acyclic graphs (DAGs)**
  - **Use the back door criterion**

Concepts

- Directed Acyclic Graphs (DAGs)

> **Graphical model that represents an assumption of causal directions between variables in a system of variables**

> **The DAGs are essential to estimate causal effects when an experiment (or the Randomized Controlled Trial) is absent.**

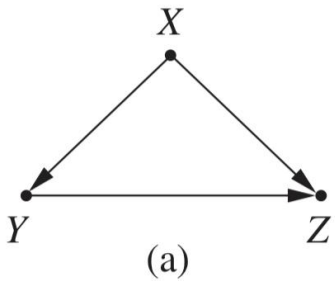

Figure 2: part (a) of Figure 1.7 (page 26 from textbook) on the acyclic graph

Concepts

- Back door criterion

> **Blocking "backdoor" paths of a pair of variables, variables A and B, allows us to calculate the direct effect between A and B (blocking a path is to condition other set of nodes *z* in the spurious path from A to B)**

Concepts (cont.)

- Back door criterion

> **We want to make sure that:**
>
> – **All indirect paths from A to B are blocked, and**
>
> – **We create no new indirect path when conditioning**

Definition 3.3.1 (The Backdoor Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the* backdoor criterion *relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

Figure 3: The definition of the back door criterion (page 61 from textbook)

Visuals

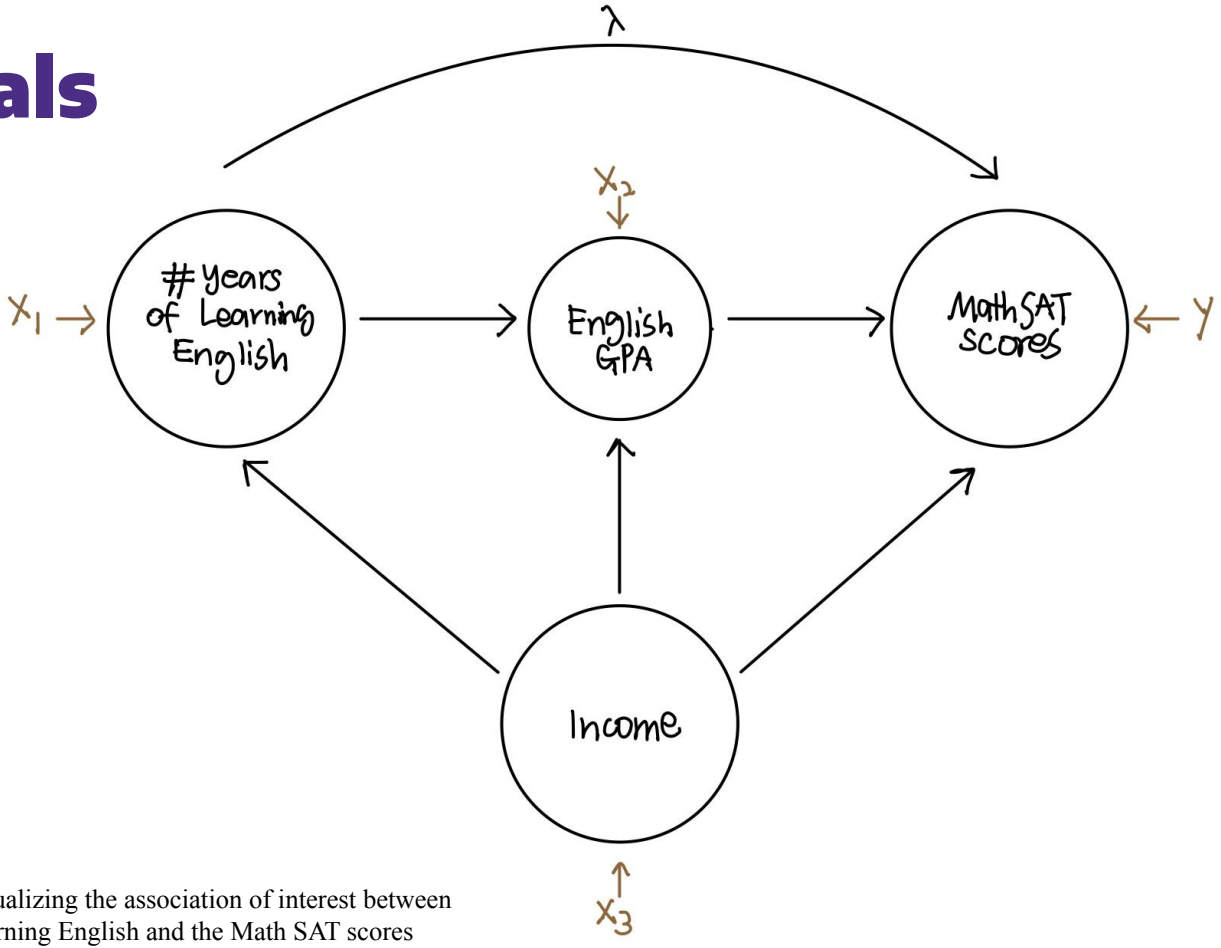

Figure 4: The DAGs visualizing the association of interest between numbers of years learning English and the Math SAT scores

Results/ Analysis

Regression equations for:

Total effect, τ

$$Y = B_1X_1 + B_3X_3$$

```
Call:
lm(formula = Total.Math ~ average.income + Academic.Subjects.English.Average.Years,
    data = sat)

Residuals:
     Min       1Q   Median       3Q      Max
-133.374  -17.052    1.695   19.065   74.792

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                           -7.929e+02  5.337e+01 -14.856   <2e-16 ***
average.income                         9.410e-04  1.081e-04   8.703   <2e-16 ***
Academic.Subjects.English.Average.Years  3.227e+02  1.408e+01  22.928   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.55 on 574 degrees of freedom
Multiple R-squared:  0.5917,    Adjusted R-squared:  0.5903
F-statistic: 415.9 on 2 and 574 DF,  p-value: < 2.2e-16
```

Direct effect, λ

$$Y = B_1X_1 + B_2X_2 + B_3X_3$$

```
Call:
lm(formula = Total.Math ~ Academic.Subjects.English.Average.GPA +
    Academic.Subjects.English.Average.Years + average.income,
    data = sat)

Residuals:
     Min       1Q   Median       3Q      Max
-130.958  -10.499    2.572   14.097   56.917

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                           -5.928e+02  4.744e+01 -12.497  < 2e-16 ***
Academic.Subjects.English.Average.GPA  1.217e+02  8.240e+00  14.773  < 2e-16 ***
Academic.Subjects.English.Average.Years  1.697e+02  1.584e+01  10.713  < 2e-16 ***
average.income                         5.507e-04  9.580e-05   5.748 1.47e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.17 on 573 degrees of freedom
Multiple R-squared:  0.7043,    Adjusted R-squared:  0.7028
F-statistic:   455 on 3 and 573 DF,  p-value: < 2.2e-16
```

Results/ Analysis

	Total Effect		Direct effect	
	Point Estimates	Standard Error	Point Estimates	Standard Error
Average Income	0.0009	0.0001	0.0006	0.0001
Number of Years learning english	322.70	14.08	169.70	15.84
English GPA	N/A	N/A	121.70	8.24

Total effect

Direct effect

Limitations

During the process, we noted that:

- **The dataset**
  a. **income s are recorded in bins-- we need to approximate the average state-level income to run the regression**
  b. **The data is state-level**
  c. **The dataset is longitudinal - state at a specific year being the observational unit**

Limitations

During the process, we noted that:

- **The DAG - we assume that the direction of causality is correct**
  a. **Potential unobserved, missing variables that could act as the additional confounders**

UNIVERSITY *of* WASHINGTON

Thank You!