Casual inference group assignment

Group 23[1]✉

[1] Statistics and Data Science, Leiden University

This study investigates the causal relationship between education level (EDUC) and long-term mortality (DEATH) using part of the Framingham Heart Study's dataset. To obtain the causal interpretation between EDUC and DEATH, we employed two approaches, i.e., G-computation and Propensity score method with inverse probability weighting. The results of both methods show that highly educated people have lower long-term mortality relative to less educated people, suggesting that better education can lead to lower long-term mortality.

1. Introduction

Literature has shown that there is a positive relationship between education level and length of life span [1, 2, 3]. However, whether a good education indeed can result in a longer life span remains to be addressed. In this study, a causal inference analysis is conducted to examine the relationship between education level and length of life span using data from Framingham Heart Study, which has been used in a long-term cohort study on cardiovascular diseases among people in the community of Framingham, Massachusetts in 1948.

1.1. Research Question.

Is a high education linked to low long-term mortality?

2. Methods

2.1 Information about the dataset.

This analysis used an enclosed dataset of the Framingham Heart Study, focusing on 4,434 participants who were selected from the 5209 volunteers since 1948. The data were collected through multiple ways such as laboratory, clinical, questionnaire, and event data on 4,434 participants during three examination periods, approximately 6 years apart, from roughly 1956 to 1968. Each participant was followed up by tracking a 24-year death, which can be caused by any reason.

2.2. Variables in the dataset. The dataset includes 16 variables with 10 exposures or confounds and 6 outcomes. The exposures/confounds are participant biological sex (SEX), age measured in years (AGE), systolic blood pressure with mmHg as unit (SYSBP), diastolic blood pressure with mmHG as unit (DIABP), use of anti-hypertensive medication (BPMEDS), current cigarette smoking (CURSMOKE), education (EDUC), body mass index with kilograms/heights in meters squared as unit (BMI) and heart rate in beats/min (HEARTRTE). For the aim of this study, we only focus on the outcome of death(DEATH).

2.3. Protocol of this study.

Eligibility criteria.

Adults in Framingham, Massachusetts consented to this study started in 1948.

Exposure definition.

highly educated (equal or more than college-level, EDUC = 1) vs. less educated (less than college-level, EDUC = 0).

Assignment procedures.

No randomization was described in this study.

Follow-up period.

24 years

Outcome definition.

mortality (death vs. still alive)

Causal contrast of interest.

$RD = E(Y(X = 1)) - E(Y(X = 0))$

Outcome Measures. The 24-year mortality measured in the current report is death from any cause (DEATH), which takes values of either 1 (dead) or 0 (alive).

2.4. Data Analysis.

2.4.1. Expert-defined direct acyclic graph (DAG).

Instead of learning from the current data, we defined an expertise-based DAG, which is shown in *Fig. 1* to prevent the so-called double dipping issues, i.e., using the same dataset to define the DAG and conduct further tests based on the DAG. With this expertise-based DAG, we obtained the minimal adjustment set when trying to get the total causal relationship between EDUC and DEATH with backdoor criteria (or backdoor adjustment rules). As a result, we obtained that the minimal adjustment set is $\{SEX, AGE\}$.

2.4.2. missing values issues checking.

Before conducting the analysis, we check whether there are any missing data issues in the variables we are interested in (i.e., EDUC, DEATH, AGE and SEX) in our dataset. As shown in *Table 1*, only the 'EDUC' variable has missing data, which is considered a predictor in the current study. We assume some less educated people are less willing to fill in their education information. That is, the missing data mechanism here is missing not at random (MNAR)

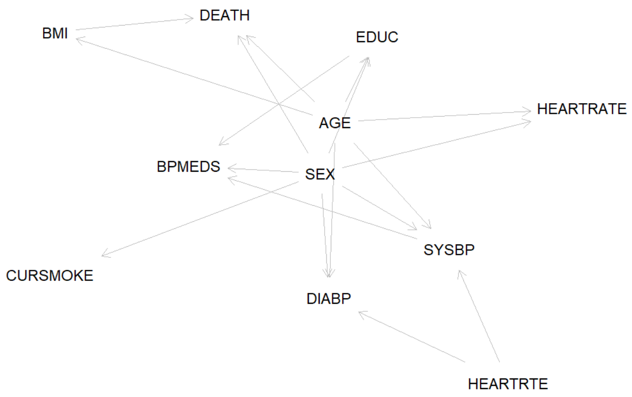

Fig. 1. DAG with all variables

depending on the predictors. Since the research interest in the current study (i.e., to obtain causal interpretation from EDUC to DEATH), this missing data mechanism will not affect our estimation. Thus, we will only consider cases with complete record, which is the remaining 97.5% of the records .

Table 1. Count of missing data

EDUC	SEX	AGE	DEATH
113	0	0	0

2.4.3. Assumptions of Identifiability Checking.

Note that the causal effect is based on the assumptions *consistency*, *positivity*, and *exchangablity*. Only if these three assumptions hold, the causal interpretation of the result would be valid and reliable. Therefore, it is crucial to check whether these three assumptions hold before conducting an analysis.

Consistency.

The exposure (EDUC) in this study is very well-defined. As all participants are adults, thus we assumed that there was no interference in the population. This means one person's mortality after 24 years does not depend on this person's education level.

Positivity & Exchangeability.

So far, we have two confounders (i.e., SEX and AGE). SEX is a categorical variable while AGE is a continuous variable. In the case of a combination of categorical and continuous variables, it becomes difficult to check whether EDUC = 1 vs. EDUC = 0 is exchangeable in terms of this combination. Therefore, we moved on to the G-computation and Propensity Score Method directly to adjust these two confounders. In that context, we only need to check whether the conditional positivity and exchangeability still hold.

2.4.4. Data Modelling.

To obtain the causal interpretation from observational data, we employed two analysis approaches (i.e., the G-computation and the propensity score method with inverse probability weighting)) to enable us to obtain the causal interpretation between EDUC and DEATH.

Approach 1: G-computation.

We used the G-computation to obtain the causal relationship between EDUC and DEATH. More concretely, we first modelled the $E(DEATH|SEX, AGE, EDUC)$ via the logistic regression model by including SEX, AGE, EDUC, and their interactions(See model coefficient in the appendix). The formula of the logistic model is DEATH ˜ SEX * EDUC * AGE. There is no need to check the model assumption as according to the procedure, we do not need to conduct the model selection here. With the logistic model, we obtained the $\hat{RD}$ and its corresponding standard error (i.e., se) with the *stdGlm*() function in *stdReg* library [4] by setting less educated group (EDUC = 1) being reference level.

Approach 2: Propensity Score Method with Inverse Probability Weighting.

As an alternative, we also used the Propensity Score Method with Inverse Probability Weighting to obtain the causal relationship between EDUC and DEATH. In detail, we first used a logistic regression model to predict the chance of each participant receiving the exposure they indeed received. Then, we set each individual's weight to one over the predicted probability. With each individual's weight, we checked the conditional exchangeability and positivity with the *love.plot*() function in the *cobalt* library [5]. We repeated the mentioned two steps until the assumptions of conditional exchangeability and positivity hold. In the end, the propensity score model was built by logistic regression with the formula of EDUC ˜ SEX + AGE (See model coefficient and diagnostic of this model in Appendix). With each individual's weight obtained through this model, we rebuild the model with *svydesign*() function, calculated RD, and its corresponding se with the *svyglm*() function in the *survey* library [6]. The 95% confidence interval was obtained with the *confint*() function.

3. Result

Result of G-computation.

According to the G-computation approach, $E(Y(\hat{X} = 1)$ = 0.311 and the $E(Y(\hat{X} = 0)$ = 0.365. *Table 2* shows that having college level or higher education results in lower 24-year mortality ($\hat{RD}$ = -0.054 with $se(\hat{RD})$ = 0.015, 95% CI = [-0.083, -0.025], p < 0.001) compared those who are not.

Table 2. Output of the G-computation modelling

	Estimate	se	L 0.95CI	U 0.95CI
HighSchoolorBelow	0.000	0.000	0.000	0.000
CollegeorMore	-0.054	0.015	-0.083	-0.025

Result of Propensity Score Method with Inverse Probability Weighting.

The results of our propensity score method with inverse probability weighting, which are reported in *Table 3*, are consistent with the results using G-computation. That is, being highly educated results in lower 24-year mortality ($\hat{RD}$ = -0.057 with $se(\hat{RD})$ = 0.016, 95% CI = [-0.083, -0.025], p < 0.001) compared to being lower educated.

Table 3. Output of the propensity score method with inverse probability weighting

	Estimate	se	t value	Pr(>|t|)
Intercept	0.364	0.009	42.34	< 0.001
EDUC CollegeorMore	-0.057	0.016	-3.52	<0.001

4. Discussion

This study used part of the data from the Framingham Heart Study to investigate the causal relationship between education and death. Two methods to estimate the ATE, which are the G-Computation and Propensity Score give consistent results that having a higher level of education will reduce the risk of 24-year death by 5.4% and 5.7% respectively.

Note that such interpretation holds only if the DAG was correctly defined. In the case of the DAG was not correctly defined, or the existence of other unmeasured variables which could result in confounding or selection basis, our conclusion would not be valid anymore. In that case, further study including those variables is necessary. Meanwhile, in terms of the G-computation, we assume that (1) conditional exchangeability, given age and gender; and (2) the linear model fits.
This study clarifies that being highly educated can result in a longer life span.

5. Acknowledgement

The code of this study can be seen in the attached file CI1Group23Assignment.Rmd.
All of the three authors equally contributed to the conceptualization, coding and writing.

Appendix

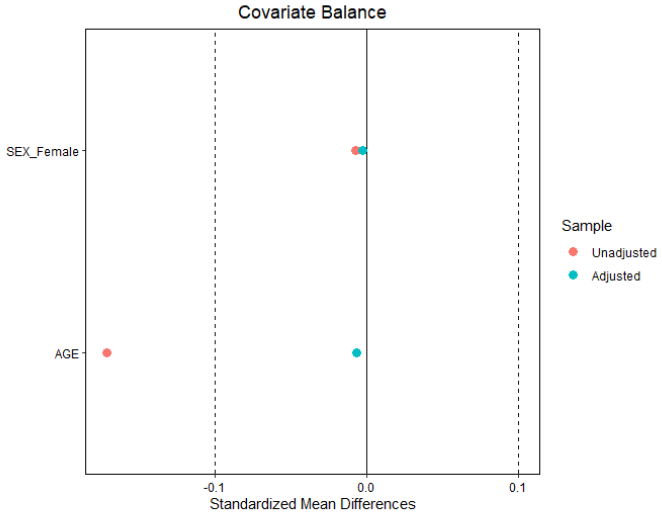

Fig. 2. Checking the Assumption of conditional exchangeability of Propensity Score method

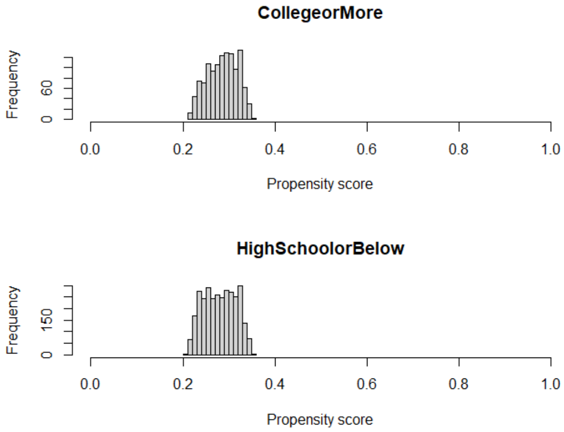

Fig. 3. Checking the assumption of positivity across levels of EDUC

Result of Assumption checking of Propensity Score Method with Inverse Probability Weighting.

Reference/literature cited

1. A Systematic Review of Major Cardiovascular Risk Factors: A Growing Global Health Concern. Cureus. 2022 Oct 10*Adhikary D, Barman S, Ranjan R, Stone H.* https://pubmed.ncbi.nlm.nih.gov/36381818/

2. The Links Between Education and Health. American Sociological Review, 60(5), 719–745.*Ross, C. E., Wu, C. (1995).* https://doi.org/10.2307/2096319

3. Ross, C. E., Wu, C. (1995). The Links Between Education and Health. American Sociological Review, 60(5), 719–745. *Ross, C. E., Wu, C. (1995).* https://doi.org/10.2307/2096319

4. Sjölander, A. (2016). Regression standardization with the R package stdReg. European journal of epidemiology, 31(6), 563-574.*Greifer, N. (2022).*

5. Greifer, N. (2022). cobalt: covariate balance tables and plots. 2020. URL https://CRAN. R-project. org/package= cobalt. R package version, 4(2), 299.*Greifer, N. (2022).*

6. Lumley, T. (2004). Analysis of complex survey samples. Journal of statistical software, 9, 1-19.*Lumley, T. (2004).*

R code.

```
# 1. The gobal DAG
library(dagitty)
library(ggplot2)
library(GGally)
library(gridExtra)
library(vioplot)
library(stdReg)
library(cobalt)
library(survey)

g.postulated <- dagitty('dag {
  SEX  [pos="2.500,3.000"]
  AGE  [pos="-4.000,-1.000"]
  SYSBP  [pos="2.500,-1.000"]
  DEATH  [outcome,pos="-1.000,-2.000"]
  BPMEDS [pos="-1.000,0.000"]
  CURSMOKE [pos="-4.000,-3.000"]
  BMI [pos="-4.000,3.000"]
  HEARTRTE  [pos="-4.000,1.000"]
  DIABP  [pos="1.000,-3.000"]
  EDUC [exposure,pos="-1.000,3.000"]
  SEX   -> {DEATH SYSBP DIABP BPMEDS
  CURSMOKE EDUC HEARTRATE}
  AGE   -> { DEATH SYSBP DIABP BMI
  HEARTRATE EDUC}
  SYSBP  -> { BPMEDS}
  BMI -> DEATH
  EDUC -> BPMEDS
  HEARTRTE -> {SYSBP DIABP}
  }
')

## Plotting DAG
plot(g.postulated)

## The minimal adjustment set is:
adjustmentSets(g.postulated,
type = "minimal", effect = "total")
```

```
## 1. Import the data and check the assumptions
df = read.csv("framingham_assignment.csv")
df$SEX = factor(df$SEX)
df$SEX = factor(df$SEX, levels = c(1, 2),
labels = c("Male", "Female"))
df$EDUC = factor(df$EDUC, levels = c(0, 1),
labels = c("HighSchoolorBelow", "CollegeorMore"))
df$DEATH = factor(df$DEATH, levels = c(0, 1),
labels = c("Alive", "Death"))

table(df$EDUC, df$DEATH)
mean(df$AGE[df$DEATH])
prop.table(table(df$EDUC, df$SEX), margin = 1)

vioplot::vioplot(df$AGE ~ df$EDUC, col = 2:3,
xlab = "", ylab = "AGE")

## Check the missing data issues
ExpDf = data.frame(EDUC = df$EDUC,
DEATH = df$DEATH, SEX = df$SEX, AGE = df$AGE)
colSums(apply(ExpDf, 2, is.na))

## outcome regression
model = glm(DEATH ~ SEX * EDUC*AGE, family =
binomial(link = "logit"), data = ExpDf)
fit.std = stdGlm(fit = model, data = ExpDf,
X = "EDUC")
summary(fit.std, contrast = "difference",
reference = "HighSchoolorBelow")
#Report the result at the probability
scale already.
summary(model) #### Assumption checking

## the propensity score method (inverse
probability weighing, matching,
stratification)
ExpDf2 = ExpDf[complete.cases(ExpDf), ]
model2 = glm(EDUC ~ SEX + AGE, family =
binomial(link = "logit"), data = ExpDf2)


ExpDf2$ps1 = fitted.values(model2)
range(ExpDf2$ps1)


ExpDf2$ipw1 <- (ExpDf2$EDUC ==
"CollegeorMore")/ExpDf2$ps1 +
(ExpDf2$EDUC ==
"HighSchoolorBelow")/(1 - ExpDf2$ps1)
summary(ExpDf2$ipw1)

# c: checking the balance
vars1 = c("SEX", "AGE")
covariates = ExpDf2[, vars1]

bal.tab(covariates,
treat =ExpDf2$EDUC,
weights = ExpDf2$ipw1,
method = "weighting",
un = TRUE)

love.plot(
  covariates,
  treat = ExpDf2$EDUC,
  weights = ExpDf2$ipw1,
  method = "weighting",
  binary = "std",
  threshold = .1
)

#histogram
par(mfrow=c(2,1))
hist(ExpDf2$ps1[ExpDf2$EDUC ==
"CollegeorMore"], xlim=c(0,1),
main="CollegeorMore",
xlab="Propensity score")
hist(ExpDf2$ps1[ExpDf2$EDUC ==
"HighSchoolorBelow"],
xlim=c(0,1),
main="HighSchoolorBelow",
xlab="Propensity score")


# f
ExpDf2$DEATH.num = as.numeric(
ExpDf2$DEATH) - 1

d.w = svydesign(~ 1,
weights = ExpDf2$ipw1,
data = ExpDf2)
fit.ipw = svyglm(DEATH.num ~
EDUC, design = d.w)
summary(fit.ipw)

confint(fit.ipw)
```