Marginal Structural Models for Causal effects of Higher Education Rates on Cancer Mortality Rates: *Should we make our society highly educated?*

Chanmin Kim † and Sooyoun Choi †*email* : lit777@stat.ufl.edu

† **Department of Statistics, University of Florida**

"College grads less likely to die from cancer"

– NBC news, 06 / 17 / 2011

It has been frequently reported that the least educated died of cancer at rates more than that of men with college degrees. Albano et al., (2007) explicitly showed the same result. In US (2001), the estimated relative risk for all-cancer mortality comparing the less educated people (≤12 years) with the educated people (>12 years education) categories was 2.38 (95% CI=2.33 to 2.43) for black men, 2.24 (95% CI= 2.23 to 2.26) for white men, and 1.76 (95% CI=1.75 to 1.78) for white women.

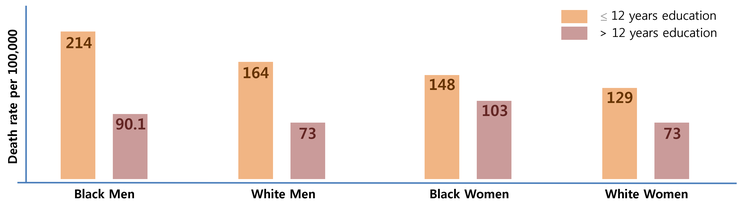


Figure 1 : Death rates per 100,000 persons from all cancers in the United States, 2001

The above graph shows that the gap between educated persons and less educated persons in terms of cancer mortality. In addition to this research, there are actually lots of clinical studies which confirm the relationship between education and cancer mortality. It seems reasonable in the sense that educated persons are more likely to have high income, have good environment, concern for their own health and spend on health related goods. Then, we can extend this idea and ask a question like this: "**Can increasing the post-secondary education rate in the society really reduce the cancer mortality rate of the society**?". We now exploit the answer.

Increasing the post-secondary education rate = Decreasing the cancers mortality rate?

If the previous studies are right, we might think the positive effect of higher (post-secondary) education rates in the society on cancer mortality. That is, if policy makers force more people to complete higher education, then cancer mortality rates might decrease in that society. Therefore, this is a matter of assessing the **causal effect** of treatment (or the main effect) on the outcome in the observational study.

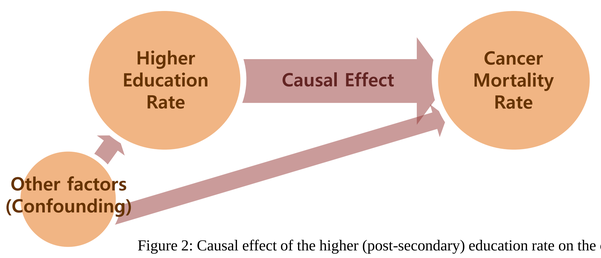


Figure 2: Causal effect of the higher (post-secondary) education rate on the cancer mortality rate.

In Figure 2, we are trying to estimate the causal effect of the treatment via Rubin Causal Model (RCM),

$$\text{Treatment Effect} = E(Y_z - Y_{z'}) \qquad \text{where } Y_z \text{ is a potential outcome under treatment } z. \qquad \textbf{(1)}$$

If the study is designed to have a randomized treatment, then **(1)** is well estimated by $E(Y \mid Z=z) - E(Y \mid Z=z')$ where the first expectation denotes the average of outcomes under treatment z and the second denotes the average of outcomes under treatment z'. Both of them are easily estimated by observed data. However, it is not the case when it comes to observational studies whose initial conditions are not randomized. Our case is one of those kind of studies. This problem is mainly caused by other factors, **confounders**, in Figure 2.

"So, the problem is how to adjust confounding in observational studies?"

To solve this issue, we need to make one strong assumption such as

Assumption 1. No unmeasured confounder other than measured confounders ***C*** on the causal pathway.

Then, $Y_z \perp Z \mid \mathbf{C}$. That is, Y_z and Z are conditionally independent given measured confounders ***C***.
With Assumption 1, we can now assess the treatment effect through the special inferencing method.

Marginal Structural Models (MSM)

To estimate the causal effect **(1)** of the treatment on the outcome, we use a marginal structural model (MSM). A marginal structural models for potential outcome Y_z is

$$E[Y_z] = \alpha_0 + \alpha_1 z \qquad \text{where } Y_z \text{ is } Y_1 \text{ if } z=1 \text{ and } Y_0 \text{ if } z=0.$$

This model differs from a regular regression model in that the former is for potential outcomes but the latter is for observed data. Here, α_0 and α_1 might be estimated by an ordinary regression approach $E[Y \mid Z=z] = \beta_0 + \beta_1 z$. However, estimates of β_0 and β_1 are unbiased for α_0 and α_1 only when we have the randomized treatment. In this study we need a strategy to adjust confounders with **Assumption 1**.

Inverse Probability of Treatment Weighting (IPTW)

In MSM, it controls confounding not by including covariates in the model but by giving certain weighs to individuals. Especially, Robins et al. (2000) proposed stabilized weights (sw_i) to avoid extreme variability,

$$sw_i = \frac{P(Z = z_i)}{P(Z = z_i \mid \mathbf{C} = \mathbf{c}_i)} \qquad \text{where } \mathbf{c}_i \text{ is a vector of measured confounders for subject } i.$$

Then, using Proc Genmod in SAS or GLM in R with weights, we could obtain unbiased estimates of marginal structural models, α_0 and α_1, from the model $E[\, Y \mid Z=z\,] = \beta_0 + \beta_1 z$. Of course, these estimates highly depend on the modeling structure of weights.

Data collection

Treatment: is the **post-secondary education rates (EDU)** by country in 2007. Society's stance on education usually doesn't change over years. Thus, data in 2007 actually well represent society's stance on education not only in 2007 but also over years. From UN datasets, we calculated EDU by

EDU = Tertiary education Enrollment / Tertiary School age population

where 184 countries have data for the denominator and 113 countries have data for the numerator. As the same reason stated above, past data are carried forward for missing values of the numerator in 2007 and we have 183 countries of complete data (discard 1 country for no information).

Outcome: is the **cancer mortality rates (MOR)** which are the number of deaths per 100,000 caused by all cancers from countries in 2008 . This data can be found in World Health Organization's data repository. The direct URL to data is 1) of DATA URL links. Countries in this data are completely matched to 183 countries from the list of EDU.

Confounder 1: is **outdoor air pollution (POL)** attributable DALYs per 100,000 capita in 2004. Here, DALY (disability-adjusted life year) means a measure of overall disease burden, expressed as the number of years lost due to ill-health, disability or early death. This data can be found on World Health Organization's data repository. It contains 18 missing values for the list of EDU. The direct URL to data is 2).

Confounder 2: is **urban population (URB)** which refers to people living in urban areas as defined by national statistical offices (% Total). This ratio can be attained on the World Bank Data website, 3). We consider data in 2007 and it has 3 missing countries for the list of EDU.

Confounder 3: is **health expenditure (EXP)** which indicates total health expenditure as a percentage of GDP in 2007 by country. It can be obtained on the World Health Organization's data repository, 4). This data contains 182 countries matched fully to the list of EDU.

Confounder 4: is **GNI per capita (GNI)** which means the gross national income per capita in 2007 measured as US$. This can be accessed through the World Bank's database, 5). This data contains 7 missing countries compared to the list of EDU.

Multiple Imputation

In the data set, all confounder variables, except EXP, have missing values. To overcome this issue, we use multiple imputation under MAR assumption as Stuart (2010)'s suggestion. To impute missing values for POL, URB and GNI, we use the following model

$$Confounder_i = region_{ij} + e_i \quad \text{where} \quad e_i \sim N(0, \sigma^2)$$

for each *Confounder*, POL, URB, GNI and j =1,2,…,14 and i =1,…,182. Here, ***region*** is a fixed effect of the categorical variable having 14 categories of world regions (e.g., East Africa=1, Middle Africa=2,…,Oceania=14) which are defined by UN country grouping, (6). All effect terms and parameters can be estimated by observed data.

We make **5 complete sets** from imputing since the efficiency of the estimate based on 5 imputations is

$$\left(1 + \frac{r}{m}\right)^{-1} = 0.98 \quad \text{where the missing rate, } r \approx 0.1 \quad \text{and the number of imputations, } m = 5.$$

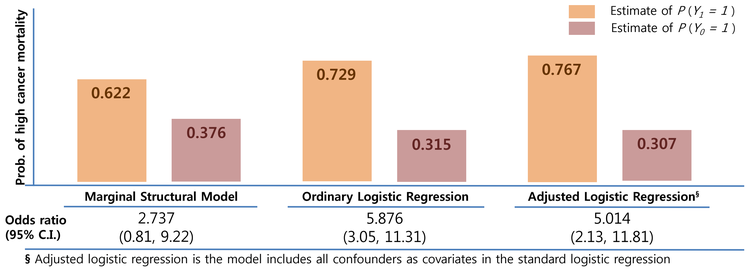
Using multiple imputation combining rules, we draw estimates from 5 imputed data sets.

Answer for the question is…

If we dichotomize the treatment and the outcome by their worldwide average (0.2738 for EDU and 223.5 for MOR), then 1 denotes a value higher than the average and 0 otherwise. Here, we use logit links

$$\text{logit } P(\, Y = 1 \mid z\,) = \beta_0 + \beta_1 z \quad \text{and}$$
$$\text{logit } P(\, Z = 1 \mid POL_i, URB_i, EXP_i, GNI_i) = \eta_0 + \eta_1 POL_i + \eta_2 URB_i + \eta_3 EXP_i + \eta_4 GNI_i$$

where weights $sw_i = P(\, Z = 1\,) \,/\, P(\, Z = 1 \mid POL_i, URB_i, EXP_i, GNI_i)$. Then, the cancer mortality rates are



§ Adjusted logistic regression is the model includes all confounders as covariates in the standard logistic regression

Figure 3 : Causal effects of post-secondary education rates on cancer mortality worldwide

In Figure 3, we compare results from MSM to those from the ordinal logistic regression (without any adjusting) and the adjusted logistic regression (including all confounding variables as covariates in the model). we can see that the marginal structural model attenuates the risk under treatment status1 (high post-secondary education rate) and augments the risk under treatment status 0 (low post-secondary education rate). When it comes to odds ratios, logistic regressions generate somewhat extreme values than MSM. Therefore, under assumption 1, we conclude that logistic regressions generate biased estimates. **See the supplemental material for detailed statistics. (with possible extension to the multilevel treatment case).**

Besides the above quantitative fact, we now realize the surprising result that

"The high post-secondary education rate in the society actually induces the high cancer mortality rate in the society."

Shouldn't it be the other way around? How can we explain this somewhat contradictory result?

In Figure 4, we simply depict this situation. Marginally (within country level), an individual completing higher education tends to have low risk of cancer mortality, which is proven by many studies such as Albano et al. (2007) using US data. However, by country, the society which has the high post-secondary education rate tends to have high cancer mortality rate. It might be interpreted as **making the society highly educated has the negative effect on individual's health, especially cancer mortality** possibly by more physical and psychological stress from competing environment , fast urbanizing and so on, which can be all affected by increasing post-secondary education rates.

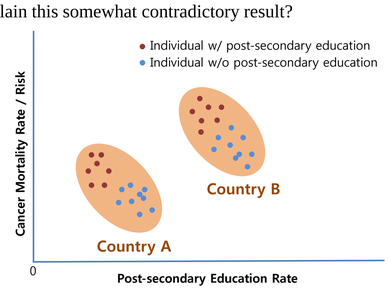


Figure 4 : Simple example of the adverse phenomenon marginally

On the other hand, within each society, people who completed post-secondary education might have high chances to earn more money and invest them to their own health, which leads to the adverse condition marginally.

Data URL links

1) http://apps.who.int/gho/data/node.main.A864?lang=en
2) http://apps.who.int/gho/data/node.main.157?lang=en.
3) http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS.
4) http://apps.who.int/gho/data/view.main.1900ALL?lang=en.
5) http://databank.worldbank.org/data/views/reports/tableview.aspx?isshared=true&ispopular=series&pid=4.
6) http://www.internetworldstats.com/list1.htm