Income's Role in Explaining Black-White Differences in the Educational Gradient in Health: Evidence from the NLSY79 and G-computation

Michael Esposito
University of Washington, Department of Sociology

Background

Of the many topics of concern in population health, perhaps none is more well studied then the association between education and health. A substantial literature on the topic has shown that, in the United States (US), those with higher levels of educational attainment have better outcomes on a host of health indicators than their less educated counterparts (Cutler and Lleras-Muney 2006; Hummer and Lariscy 2011; Link 2008; Mirowsky and Ross 2003; Walsemann et. al. 2013). Moreover, this *educational gradient in health* is thought to be maintained by a large, diverse set of mechanisms, ranging form increased income to better developed metacognitive skills (Mirowsky and Ross 2003; Cutler and Lleras-Muney 2006). For this persistently positive, multifaceted, association with health, education is often positioned as one of the more promising venues through which changes in population health could occur (Cutler and Lleras-Muney 2006; Mirowsky and Ross 2003).

Given education's potential role in shaping population health, researchers have turned towards examining if, and to what degree the association between education and health is dependent on other social forces. In this area, researchers have been particularly active in examining how the link between education and health is dependent on race (Crimmins and Saito 2001; Farmer and Ferraro 2005; Hayward et al. 2000; Kimbro et. al. 2008; Masters et. al. 2012; Montez et. al. 2012). Indeed, research has consistently shown that education's association with health is weaker for Non-Hispanic Blacks than it is for Non-Hispanic Whites; on this matter, findings range from that the educational gradient is similar in shape, yet measurably more steep for Non-Hispanic Whites relative to Non-Hispanic Blacks (e.g., Montez et. al. 2012; Kimbro et. al. 2008), to that gradient is negligible to non-existent for Blacks, but quite steep for Whites (e.g., Farmer and Ferraro 2005; Kimbro et. al. 2008).[1]

Of the many potential explanations for *why* Black-White variation exists in terms of the the health returns to education, perhaps none is more conceptually intuitive than income (Walsemann et. al. 2013). In more detail, income is widely considered to be a important mechanism through which education operates to effect health (i.e, additional education often leads to greater income, and greater income often leads to better health) (Cutler and Lleras-Muney 2006; Kawachi et. al. 2010). At the same time, we know that Blacks in the U.S tend to make less than their similarly educated White counterparts (e.g., in 2010, the average income of college educated Black males was approximately $22,000 less then the average income of college educated White males) (Walsemann et. al. 2013; Williams et. al. 2010). Given the above two points, it could be the case that the racial variation in the *health returns* to education is largely a reflection of the racial variation in the *income returned* from increased education.

Though the notion that income is an important reason for why racial differences in the health returns to education exists is well reasoned, little empirical work has been produced to support said position (Walsemann et. al. 2013). The aim of this paper then, is to clarify the role income plays in maintaining Black-White differences in the shape of the educational gradient in health. In this endeavor, we employ a combination of rich data from the National Longitudinal Study of Youth 1979 (Bureau of Labor Statistics 2012), and G-Computation, a technique which allows us to quantify the

[1] From here on, we will refer to 'Non-Hispanic White' as 'White', and 'Non-Hispanic Black' as 'Black'

role income plays in maintaining Black-White differences in educational gradients while avoiding post-treatment bias (Lepage et. al. 2012). In addition, to alleviate the model specification concerns that come with a G-Computation approach, I make use of a nonparametric machine learning algorithm (Bayesian Additive Regression Trees) to estimate the regression models necessary to the G-Computation process (Chipman et. al. 2010).

Analytical Framework

To clarify the role income plays in maintaining race-based differentials in the health returns to education, we draw upon the following conceptual model:

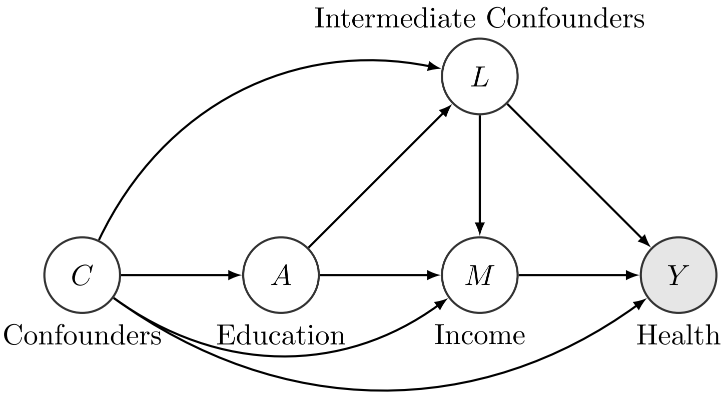

Figure 1: Diagram of Education, Income, and Health's Relationship

For the purposes of this paper, we are interested in estimating the the effect of education on health, or

$$E[Y|A=1,C]-E[Y|A=0,C], \tag{1}$$

which gives the difference in the average health of individuals with a particular level of education and the average health of individuals without said level of education, holding potential confounders constant. We are also interested in estimating the effect of education on health while holding income constant, or

$$E[Y|A=1,C,M]-E[Y|A=0,C,M], \tag{2}$$

which gives the same value as Eq. 1, save for that income is held constant across counterfactuals.

From Eq. 1 and Eq. 2, we have the materials needed to comment on the role income plays in explaining the difference in health returns to education for Blacks and Whites. That is, *if income plays a role in creating variation in the health returns to education between Blacks and Whites, the difference in the effect of education on health for Blacks and the effect of education on health for Whites should be greater than the difference in the effect of education on health while holding income constant for Blacks and the effect of education on health while holding income constant for Whites. Moreover, The degree to which the racial difference in educational gradients diminishes in the above scenario provides an approximation of how important income is in creating said racial difference in health. (e.g., if the racial difference in the effect of education on health completely disappears after accounting for income, we have evidence that income is completely responsible for the differential returns to education).*

G-Computation

Because of the presence of intermediate confounders (L) (e.g., features which are influenced by one's educational attainment, and which influence income and later health outcome), obtaining the effect of education on health while holding income constant via 'standard regression' can lead to post-treatment bias (Gelman and Hill 2007; Lepage et. al. 2012). Given this fact, we estimate Eq. 2 using a G-Computation procedure (Daniel et. al. 2011 or Lepage et. al. 2010). Assuming, for the sake of illustration, that L is only made up of one feature, the G-Computation procedure follows:

Algorithm 1 G-Computation Procedure (Lepage et. al. 2010)

```
Find $E(Y|C, M, L, A)$
Find $E(L|C, A)$
for each individual $i$ do
    Draw $L_0$ from $f(L|C = c_i, A = 0)$
    Draw $L_1$ from $f(L|C = c_i, A = 1)$
    Draw $Y_{00}$ from $f(Y|C = c_i, M = 0, A = 0, L = l_0)$
    Draw $Y_{10}$ from $f(Y|C = c_i, M = 0, A = 1, L = l_1)$
Find $E(\mathbf{Y_{10}}) - E(\mathbf{Y_{00}})$ [which $\approx (E[Y|A = 1, C, M] - E[Y|A = 0, C, M])$]
```

One potential drawback of the G-Computation procedure is its heavy reliance on modeling (e.g., each feature in the set L needs to be model conditional on A anc C for the procedure to work) (Daniel et. al. 2011). To alleviate concerns of model misspecification that come with reliance on such a method, we use Bayesian Additive Regression Trees to model the expectations required in the G-Computation process (Chipman et. al. 2010). This nonparametric machine learning algorithm is particularly adept at *letting the data decided* upon accurate, yet parsimonious, models of itself. That BART takes the model specification process out of the researchers hands, and instead leaves said process to a principled search through the set of all potential models, increase our confidence that the models necessary to the G-Computation process are properly specified representations of reality.

Data

We use data from the National Longitudinal Study of Youth 1979 (NLSY79). The NLSY79 is a nationally representative sample of over 10,000 individuals aged 14-22 in 1979 (Bureau of Labor Statistics 2012). From their first interview date, respondents where subsequently interviewed annually (and later biannually) through the year 2010.

Measurement

We begin by operationalizating our treatment variable, education [A]. In the endeavor, we subset our data to only include those individual whom completed their education between 1980 and 1986. Restricting the data in this way conceptually reflects *an experiment that we would like to see*; that is, under this design, the period between 1980 and 1986 can be considered a a "treatment period", where subjects in the study where assigned a dosage of education. Everything before this point then can, without bias, be considered a *pre-treatment* feature (and thus potentially a confounder), and everything after said period could be considered *post-treatment* (and thus potentially influenced by education). In the aforementioned subset of the data, we are left with 1,645 individuals whom received a high school degree, and 744 individual whom received a college degree. The specific question being examined in the effect of education on health quantity then is, *how different would an individual's health had been had they had remained at a high school degree, instead of completing* *a college degree?*

Next, we define our response variable, health $[Y]$. In this regard, we use the self-rated health indicator in the NLSY79 collected from each respondent when s/he turned 40. For this question, respondents were asked to describe their health as either 'excellent, very good, good, fair, or poor.' We choose this general measure of health because of its availability in the data and the substantial racial differences in the educational gradient in self-rated health reported by previous research (e.g., Farmer and Ferraro, 2005). With the timing of both treatment and response set, we can determine from which survey years confounders and intermediate confounders are pulled from:

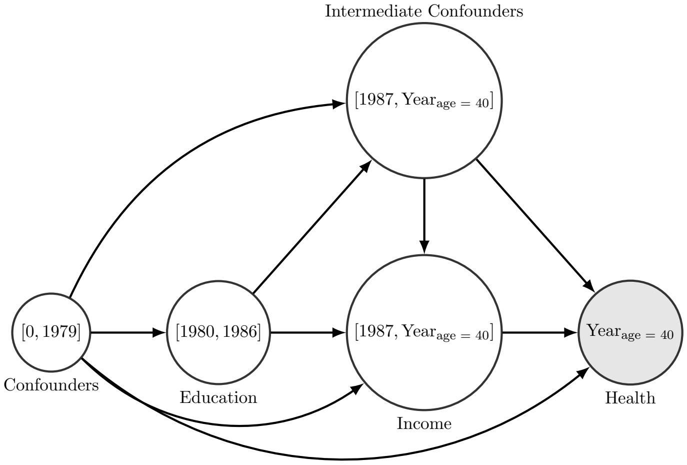
Intermediate Confounders
$[1987, \text{Year}_{\text{age}\,=\,40}]$
$[0, 1979]$
Confounders
$[1980, 1986]$
Education
$[1987, \text{Year}_{\text{age}\,=\,40}]$
Income
$\text{Year}_{\text{age}\,=\,40}$
Health

Figure 2: Diagram of NLSY79 survey years from which features are pulled

With this time ordering set, we defined one's post-education income (M) as one's average family income between the 1986 and the year they turned 40. Because of this, the effect of education on health holding income constant is equivalent to *net of the effect of family income, how different would an individual's health had been had they had remained at a high school degree, and not completed a college degree?*

As intermediate confounders $[L]$, we select several variables which simultaneously (1) are effected by the educational attainment earned in the treatment period, (2) effect income, and (3) effect self-rated health. This set includes: occupation, employment, marital status, and cognitive ability. As we did with income, we summarize each of these confounders by taking their average value across the period $[1987, \text{Year}_{\text{age}\,=\,40}]$ (e.g., for occupation, we find average one's occupational prestige score in 1987, 1988, 1989, ..., $\text{Year}_{\text{age}\,=\,40}$). Finally, for pre-treatment confounders $[C]$, we select every feature from the that occurred 1979 and prior that may influence the probability that one moves beyond a high school degree, and may also influence one's post-educational attainment self-rated health. In this regard, we select/control for 110 features which may serve as confounders.

Expected Results

Given that educational returns to income vary measurably between racial groups in the U.S., I expect to see that controlling for income diminishes racial disparities in the educational gradient to health (e.g., $[\text{Eq.1}_{\text{White}} - \text{Eq.1}_{\text{Black}}] > [\text{Eq.2}_{\text{White}} - \text{Eq.2}_{\text{Black}}]$). Given that Blacks in the United States face structural disadvantage in terms of translating their educational credentials into resources aside form income (e.g., occupation) though, I do not expect for racial disparities in the educational gradient to completely disappear after accounting for differences in income.

References

Bureau of Labor Statistics, U.S. Department of Labor. 2012. National Longitudinal Survey of Youth 1979 cohort, 1979-2010 (rounds 1-24). Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH.

Chipman, Hugh, Edward George and Robert McCulloch. 2010. "BART: Bayesian additive regression trees." *Annals of Applied Statistics* 4: 266-298.

Daniel, Rhian, Bianca L. De Stavola, and Simon N. Cousens. 2011. "gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula." *The Stata Journal* 11: 479-517.

Cutler, David and Adrianna Lleras-Muney. 2006. "Education and Health: Evaluating Theories and Evidence." NBER Working Paper 12352. National Bureau of Economic Research, Chicago, IL

Farmer, Melissa, and Kenneth Ferraro. 2005. "Are racial disparities in health conditional on socieconomic status?" *Social Science and Medicine*60: 191-204.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel Models.* Cambridge Press.

Hayward, Mark D., Eileen M. Crimmins, Toni Miles and Yu Yang. 2000. "The Significance of Socioeconomic Status in Explaining the Racial Gap in Chronic Health Conditions." *American Sociological Review* 65: 910-930.

Hummer, Robert and Joseph Lariscy. 2011. "Educational Attainment and Adult Mortality." In *International Handbook of Adult Mortality.* Vol. Chapter 12, edited by R.G Rogers, and E. Crimmins. New York: Springer.

Kawachi, Ichiro, Nancy Adler and William Dow. 2010. "Money, schooling, and health: Mechanisms and causal evidence." *Annals of the New York Academy of Sciences* 1186: 56-68.

Kimbro, Rachel, Sharon Bzostek, Noreen Goldman and German Rodriguez. 2008. "Race, Ethnicity, and the Education Gradient in Health." *Health Affairs* 27: 361-372.

Lepage, B., Dedieu, D., Savy, N., and Lang, T. 2012. "Estimating controlled direct effects in the presence of intermediate confounding of the mediator-outcome relationship: Comparison of five different methods." *Statistical Methods in Medical Research.* 0: 118

Link, Bruce. 2008. "Epidemiological Sociology and the Social Shaping of Population Health." *Journal of Health and Social Behavior* 49: 367-384.

Link, Bruce and Jo Phelan. 1995. "Social Conditions as Fundamental Causes of Disease." *Journal of Health and Social Behavior* 35: 80-94.

Masters, Ryan, Robert Hummer, and Daniel Powers. 2012. "Educational Differences in U.S. Adult Mortality A Cohort Perspective." *American Sociological Review* 77: 548-572.

Montez, Jennifer, Robert Hummer and Mark Hayward. 2012. “Educational Attainment and Adult Mortality in the United States: A Systematic Analysis of Functional Form.” *Demography* 49: 315-336.

Mirowsky, John, and Catherine E. Ross. 2003. *Education, Social Status and Health* New York: Aldine de Gruyter.

Walsemann, Katrina M., Gilbert C. Gee, and Annie Ro. 2013. “Educational attainment in the context of social inequality: New directions for research on education and health.” *American Behavioral Scientist* 57: 1082-1104.

Crimmins,E., and Saito,Y.(2001).Trends in disability free life expectancy in the United States, 1970 to 1990: Gender, racial, and educational differences. *Social Science and Medicine* 52: 1629-1641.

Williams, D. R., Mohammed, S. A., Leavell, J., and Collins, C. 2010. “Race, Socioeconomic Status and Health: Complexities, Ongoing Challenges and Research Opportunities. Annals of the New York Academy of Sciences” (1186): 69-101.