Paper: RELATIONSHIP BETWEEN HEALTH CONDITION AND INDIVIDUAL INCOME AMONG DISABLED PEOPLE

Ha Cao, Jeny Kwon, Maddie Haines

4/20/2018

Abstract

In this project, we investigated the relationships between income and health in the United States within a more specific context, that of disabled individuals to whom health related issues are especially important. We modeled not just the effect of income on health, but also how this outcome was influenced by gender, type of disability, marital status, education, and race to get a more nuanced understanding of the association. Our research can be used to motivate future policy decisions to better accommodate health-related needs and address income inequality issues for the disabled community and to serve as a jumping-off point for further, more in-depth studies about this community.

Introduction

The association between health and income is as well documented as it is vital to understand when looking to improve the living standards of any subset of our society (1). In this paper, we studied this issue in the context of disabilities because disabled people make up a considerable proportion of the U.S. population and deserve our special attention. In 2017, the U.S. population was 325.7 million and approximately 1 in 7 people lived with some kind of disability (2). Additionally, health issues are extremely important for disabled people because individuals with disabilities are more likely than people without disabilities to report poorer overall health, less access to adequate health care, and physical inactivity (3, 4). Also, income is likely to affect their access to healthcare and social activities and their quality of life, so we wanted to study the relationship between health and income for disabled individuals, while adjusting for potential confounding factors, such as age or race. This study used data from IPUMS-CPS (Current Population Survey), and we focused only on data from 2017. Moreover, we speculated that the association between self-rated health and income among disabled adults would vary by types of difficulties because different difficulties are likely to make people feel different about their health and limit them to certain jobs only. We also wanted to check whether the gender pay gap exists in the disabled community; if so, female disabled individuals are among the most vulnerable and disadvantaged, because they have to suffer from both disabilities and gender inequality prevailing in our society. We did not compare disabled and non-disabled populations because many studies have been conducted about this, which showed that the pay gap between non-disabled and disabled workers is 13.6%, which makes logical sense because disabled people are likely to have limitations in their employability as opposed to non-disabled people (5). Instead, we wanted to focus our study only our disabled people to understand their community better and find good ways to support them.

Hypotheses and Methods

- Primary hypothesis: among those with any disabilities, people with higher self-reported general health have higher income than those with lower self-reported general health
- First secondary hypothesis: the difference in income between people with higher and lower self-reported general health will be larger for people with physical difficulty than people with remembering disability
- Second secondary hypothesis: the difference income between people with higher and lower self-reported gender health will be larger for disabled males than disabled females

We used the following 6 main variables: `DIFFANY`, `INCTOT`, `HEALTH`, `DIFFREM`, `DIFFPHYS`, and `SEX` (`SEX` would be included as a confounder in the first and second models and a main explanatory variable in the third model). For confounding variables, we consistently used `AGE`, `RACE`, `MARST` (marital status), and `EDUC` (educational attainment) in our models. We dropped levels for factor variables, created binary variables `isHealthy`, `isFemale`, `isWhite`, `isMarried`, `hasDegree` in lieu of categorical variables by grouping similar values, and created `age` and `income` as numeric variables based on original variables `AGE` and `INCTOT`. We used the data from the Annual Social and Economic Supplement (ASEC) in CPS instead of the monthly data. The raw sample size of our study was 185,914, but then we filtered out missing values in the main variables, so the sample size became 16,464. For each hypothesis, we built a multiple regression model and tested the significance of

the main explanatory variables in that model. We also conducted t-tests to conclude whether the potential confounders in the model were statistically significant.

Results

Health and Income

The fitted model is:

$\widehat{income} = 3180.946 + 1414.723 \cdot isHealthy + 85.444 \cdot age - 3626.189 \cdot isFemale + 1714.271 \cdot isWhite + 1097.502 \cdot isMarried + 6113.146 \cdot hasDegree$

The fitted model shows that among people with disabilities, those who reported generally good health are predicted to earn 1414.723 dollars more than those who reported generally poor health on average, adjusted for age, sex, race, marital status, and educational attainment. The t-stat for `isHealthy` is 10.47, p-value is <2e-16 << 0.05, which provides statistically significant evidence that there is a positive linear relationship between good health and income. The 95% confidence interval is [1149.97093, 1679.47432], which means we are 95% confident that the true association between health and income is between 1149.97093 and 1679.47432 dollars on average, holding everything else constant. The fitted model also gives us the p-values for each of the confounders. All of the p-values are less than 0.05, so we can conclude that all of the confounders are statistically significant in predicting an individual's total income.

Health, Type of Disability, and Income

We filtered data so that we would only test this hypothesis on people with one kind of difficulty, either remembering or physical difficulty, and created a new binary variable `hasDiffRem`. Thus, the sample size for this model is reduced to 10,017.

The fitted model is:

$\widehat{income} = 2810.2 + 1401.840 \cdot isHealthy - 339.851 \cdot hasDiffRem + 88.672 \cdot age - 2792.97 \cdot isFemale + 1334.409 \cdot isWhite + 518.45 \cdot isMarried + 6113.957 \cdot hasDegree - 753.494 \cdot isHealthy * hasDiffRem$

The p-value of `hasDiffRem` is 0.16, which means we cannot find statistically significant evidence that the income of people with remembering difficulty is less than that of people with physical difficulty. Meanwhile, the p-value of the interaction term `isHealthy*hasDiffRem` is 0.037, which means that we can find statistically significant evidence at the 5% significance level that the relationship between reported health and life quality status and income does vary by the kind of difficulty the person has, whether it is physical or remembering difficulty. We can interpret the coefficient of the interaction term, -753.494, as: the difference in income between people with generally good health and generally bad health is 753.494 dollars less for people with remembering difficulty than those with physical difficulty on average, holding everything else constant. **In simpler terms, health is expected to affect income of people with physical difficulty more than those with remembering difficulty.** As the coefficient of `isHealthy*hasDiffRem` is significant, we want to construct its 95% confidence interval, [-1460.17907, -46.80939], which means we are 95% confident that the difference in income between people with generally good health and generally bad health is between 46.81 and 1460.18 dollars (on average) less for people with remembering difficulty than those with physical difficulty, holding everything else constant.

Health, Gender, and Income

This model has the same sample size of 16,464 as that of the primary hypothesis.

The fitted model is:

$\widehat{income} = 3065.399 + 1707.715 \cdot isHealthy + 85.485 \cdot age - 3424.067 \cdot isFemale + 1715.941 \cdot isWhite + 1098.683 \cdot isMarried + 6116.136 \cdot hasDegree - 548.445 \cdot isHealthy * isFemale$

The p-value for `isFemale` is <2e-16 << 0.05, which means there is statistically significant evidence that the income of disabled women is less than the income of disabled men on average, holding everything else constant. Also, the p-value for the interaction term `isHealthy*isFemale` is 0.0418 < 0.05, which means there is statistically significant evidence that the relationship between health and income does vary by gender. We can interpret the coefficient of the interaction term as: the difference in income between people with generally good health and generally bad health is 548.445 dollars less for disabled females than disabled males on average, holding everything else constant. **In simpler terms, health is expected to affect income of disabled males more than disabled females.** The 95% confidence interval of `isFemale` is [-1076.60858, -20.28173], which means we are 95% confident that the difference in income between people with generally good health and generally bad health is between 20.28 and 1076.61 dollars (on average) less for disabled females than disabled males, holding everything else constant.

Discussion

In summary, our research is aimed at exploring the relationship between health and income among the community of disabled people, to whom health-related issues are extremely important because disabled people might need greater access to medical services and their health care costs might also be greater. After analyzing, we have found a positive relationship between health and income for disabled people (i.e. the more healthy people are, the more money they are expected to earn). Also, the gender pay gap does carry through into the disabled community (i.e. women still earn less than men), as per the intersectionality theory. Furthermore, health is found to affect income of people with physical difficulty more than that of people with remembering difficulty, and affect income of disabled men more than that of disabled women.

The strength of our analysis lies in the careful consideration of many potential confounders to achieve more precise prediction models. However, there are some limitations that we wished to remove in the future:

- The data about health are actually perceptions of people who did the survey about their own health conditions. We are concerned because people's own perceptions may not reflect the real conditions of their health. We are not quite sure if `isHealthy` means one's true health conditons or one's self-perception of health conditions (which might be affected by people's perspectives on life as well).
- We only considered two types of disability, physical and remembering, and we wish to consider more types of difficulties to see how they affect a person's income.

References

1. Dubey, Lisa, Kim X. Luk, Laudan Aron, Emily Zimmerman, Steven H. Woolf, and Sarah H. Simon. How Are Income and Wealth Linked to Health and Longevity? Issue brief. Virginia Commonwealth University. N.p.: Urban Institute, April 2015. Web.

2. https://disabilitycompendium.org/sites/default/files/user-uploads/AnnualReport_2017_FINAL.pdf

3. Crossley, Mary. "BECOMING VISIBLE: THE ADA'S IMPACT ON HEALTH CARE FOR PERSONS WITH DISABILITIES." Alabama Law Review, 2000, pp. 355–370.

4. https://www.cdc.gov/ncbddd/disabilityandhealth/relatedconditions.html

5. https://www.theguardian.com/commentisfree/2018/apr/11/gender-pay-gap-disability-disabled-people-job

Data Appendix

Structure and names

There are 18 variables and 185914 observations. The variables are the following:

1) YEAR gives a 4-digit indicating the year in which the survey was conducted
2) SERIAL gives a 5-digit numeric value identifying number unique to each household in a given survey month and year
3) ASECFLAG indicates whether the respondent is part of the ASEC or the March Basic
4) MONTH is a 2-digit numeric value indicating the calendar month of the CPS interview
5) CPSID gives a defined variable of 14-digits that uniquely identifies households
6) AGE contains the age of the individual at last birthday (number)
7) SEX contains the sex of the individual with male as 1 and female as 2
8) RACE gives a 3-digit code identifying the race of the individual
9) MARST gives a 1-digit code identifying the marital status of the individual
10) OCC gives a 4-digit numeric value reporting the person's primary occupation
11) EDUC gives a 3-digit code indicating the person's educational attainment
12) DIFFREM contains whether the person has difficulty remembering with 1 as no difficulty and 2 as having difficulty remembering
13) DIFFPHYS contains whether the person has physical difficulty with 1 as no difficulty and 2 as having physical difficulty
14) DIFFANY contains whether the person has any difficulty with 1 as no difficulty and 2 as having a difficulty
15) CPSIDP gives a defined variable of 14-digits that uniquely identifies individuals
16) PERNUM gives a 2-digit numeric value numbering every person within each household consecutively starting with "1" in the order in which they are listed in the original CPS data
17) INCTOT contains each person's total pre-tax personal income or losses for the previous calendar year
18) HEALTH gives the self-determined health status on a five-point scale (1=Excellent, 2=Very good, 3=Good, 4=Fair, 5=Poor)

Variable analysis

YEAR

All the values are 2017, which is totally expected, because we only selected the sample from 2017.

SERIAL

SERIAL represents household serial number generated by IPUMS-CPS to differentiate households from each other in a given month or year and is automatically included in the dataset, so we shouldn't care much about these values.

ASECFLAG

ASECFLAG indicates whether the respondent is part of the ASEC or the March Basic and is automatically included in the dataset. We also shouldn't care much about these values because it is not relevant to our research.

MONTH

MONTH indicates the month when the CPS interview was conducted. The values are all 3, so it means for all observations, the CPS interviews were conducted in March.

CPSID

CPSID is an IPUMS-CPS defined variable that uniquely identifies households across CPS samples, and we don't have to worry about it because it's not relevant to our research.

AGE

AGE gives each person's age at last birthday. It ranges from 0 to 85 years old, so it means the dataset covers people from newborn babies to elderly people. The mean is nearly 37 and the median is 36, roughly the same as each other, so the dataset should be normally distributed.

SEX

This one is a categorical variable, so we'll have to go back and make it into a factor. Therefore, these min, median, and quartiles don't really have any meaning in context. Thankfully, there aren't any missing data observations.

MARST

This one is a categorical variable, so we'll have to go back and make it into a factor. Therefore, these min, median, and quartiles don't really have any meaning in context. Thankfully, there aren't any missing data observations. This will probably end up transformed into 'married' and 'not married.'

OCC

Occupation will also probably end up as a factor variable of some sort. Once again the min, max, and quartiles don't have real meaning here, but thankfully none of the observations are missing.

EDUC

Educ will also probably end up as a factor variable of some sort. Once again the min, max, and quartiles don't have real meaning here, but thankfully none of the observations are missing.

DIFFREM

'Difficulty remembering' will also probably end up as a binary variable. Once again the min, max, and quartiles don't have real meaning here, but thankfully none of the observations are missing.

DIFFPHYS

'Physical difficulty' will also probably end up as a binary variable. Once again the min, max, and quartiles don't have real meaning here, but thankfully none of the observations are missing.

DIFFANY

'Any difficulty' will also probably end up as a binary variable. Once again the min, max, and quartiles don't have real meaning here, but thankfully none of the observations are missing.

CPSIDP

This one is a linkeage variable, and has no meaning in context. We probably won't be using this variable itself. No observations are missing.

PERNUM

This one is an identifying variable meant to identify individual people in the dataset, so we probably won't end up using this one itself. No observations are missing.

INCTOT

The min is -9999, meaning that person lost quite a bit of money that year. The User Extracts indicated that there could be negative values, so this was expected. The max is a number so big it has to be expressed in scientific notation. That also makes sense- some people make a lot of money. There is no missing observations.

HEALTH

This one will also be a categorical variable of some sort, though since the options are listed logically, with 1 being Excellent Health and 5 being Poor Health, the mean isn't totally meaningles- 2.15 is believable, as that would indicate most people saying they were in good health. There are no missing observations.