Project 2: Counterfactual Disparity Measure for Race, Income, and High Blood Pressure in the NHEFS

Ashley I Naimi, PhD
Associate Professor
Director of Graduate Studies
Dept of Epidemiology
Emory University

✉ ashley.naimi@emory.edu

EMORY | ROLLINS SCHOOL OF PUBLIC HEALTH

Overview

Racial disparities in high blood pressure are well established. There are also know associations between race and income, and income and high blood pressure. We'd like to understand the extent to which racial disparities in high blood pressure are due to corresponding racial disparities in income.

- How much of the racial disparity in high blood pressure is explained by income?

- We'll use the NHEFS data, collected between 1971 and 1982, to answer this:
    - Exposure: Race, measured as "black or other" versus "white"
    - Mediator: Total Family Income (1971), measured categorically in increments of between $1,000 and $5,000, with a final category of $25,000+
    - Outcome: High blood pressure, defined as systolic ≥ 140 AND diastolic ≥ 90
    - Confounders (income and HBP): marital status, age, and education

Overview

- Three Phases - data management and exploration; analysis; interpretation

- Seven steps
  - Setup Project Folder
  - Data Management and Exploration
  - Import, Transform, Explore Raw Data
  - Regression Analysis
  - CDM Analysis
  - Interpretation

Overview: The Data

```
## # A tibble: 6 × 6
##   race  income marital   age school hbp  
##   <fct>  <dbl> <fct>   <dbl>  <dbl> <fct>
## 1 1         19 0          42      7 1    
## 2 0         18 0          36      9 0    
## 3 1         15 1          56     11 0    
## 4 1         15 1          68      5 0    
## 5 0         18 0          40     11 0    
## 6 1         11 1          43      9 0    
```

Phase 1: Data Management and Exploration

Step 1: Construct a Project Folder

- Create a folder called `Disparities_Project2`
- Create Subfolders: `data`, `code`, `figures`, `misc`, `sandbox`, and `report`
- Create an RStudio Project in the `Disparities_Project2` folder

Step 2: A Data Management and Exploration File

- In the `code` subfolder, create two code files:
  - `data_man.R`
  - `main_analysis.R`
- Open the `data_man.R` file and include the **preamble** at the top:

```
packages <- c("data.table","tidyverse","skimr","here", "lmtest", "sandwich")

for (package in packages) {
  if (!require(package, character.only=T, quietly=T)) {
    install.packages(package, repos='http://lib.stat.cmu.edu/R/CRAN')
  }
}

for (package in packages) {
  library(package, character.only=T)
}
```

Step 3: Import, Explore, and Transform Data

- Import the NHEFS into R. In this case, we will read the data directly from the web using the `url()` function:

```r
file_loc <- url("https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/1268/20/nhefs.csv")

nhefs <- read_csv(file_loc)
```

- Using the piping operator with tidyverse, complete the following data manipulations:

  - `select()` the following columns from the source data: `race, income, marital, age, school, dbp, sbp`
  - `mutate()` systolic and diastolic blood pressure into a measure of high blood pressure and convert it to a factor
  - `mutate()` race into a factor
  - `mutate()` marital status into a binary variable (marital > 2) and convert it to a factor
  - remove `sbp` and `dbp` from the dataset
  - omit any missing data using `na.omit()`

Note: whenever downloading data from the internet, consider saving a raw data file to your local computer/server. Otherwise, you may be jeopardizing the reproducibility of your results, should the original source data ever be removed or modified on the web.

Step 3: Import, Explore, and Transform Data

- explore the data you just imported using:
    - the `head()` and/or `tail()` functions
    - `skimr::skim()`
    - `GGally::ggpairs()`
    - `summary()`
    - basic descriptives and table functions: `mean(), median(), sd(), table()`
    - other

Step 3: Import, Explore, and Transform Data

- Conduct some basic regression modeling of the NHEFS data
  - Look at the relationship between HBP and race, HBP and income, and race and income
  - For the relationship between HBP and income, fit a logistic model with income coded as a continuous linear variable, a continuous variable with splines, and a categorical variable with three categories.
  - Base the categories on the histogram of income
  - Create a dose-response plot with the fits from each of these three models
  - Look at the relationship between race and income

Step 3: Import, Explore, and Transform Data

- Once you're satisfied with the income thresholds, you're ready to export the data

```
## # A tibble: 6 × 7
##    race income marital   age school   hbp income_cat
##   <dbl>  <dbl>   <dbl> <dbl>  <dbl> <dbl> <chr>
## 1     1     19       0    42      7     1 (17,20]
## 2     0     18       0    36      9     0 (17,20]
## 3     1     15       1    56     11     0 [11,17]
## 4     1     15       1    68      5     0 [11,17]
## 5     0     18       0    40     11     0 (17,20]
## 6     1     11       1    43      9     0 [11,17]
```

Step 4: Export the Data

- Create a new data file called `nhefs_analytic_data.csv`

- You should use the `write_csv()` function and the `here()` function to export to the `data` subfolder

Phase 2: Analysis

Step 1: Set Up and Import

- Open the `main_analysis.R` file and include the **preamble** at the top. Add the `lmtest`, `sandwich`, and `VGAM` packages to the list.

- Import the `nhefs_analytic_data.csv` data, and explore. For example:

```
head(a)
tail(a)

GGally::ggpairs(a[,c("c","y")])
```

Step 2: Basic Regression

- Fit an outcome model regressing `hbp` against all other variables in the data
    - use a linear and logistic model to explore on the additive and multiplicative scales
    - include an interaction between race and income
- Fit a propensity score model for the three category income variable adjusting for `marital`, `age`, and `school`
- You'll need to use the `vglm()` function from the `VGAM` package
    - Use the `family = multinomial` argument

Step 2: Basic Regression

- Generate a propensity score for the three category income variable*
    - Create a PS overlap plot for income and save the figure using `ggsave()`
    - Construct stabilized IP weights for the exposure and look at the distribution using `summary()`
- Interpret these models, figures, and summaries

* We'll go over the code needed to generate propensity scores for categorical variables in depth.

Step 3a: Exposure - Outcome Association

a. Fit a marginally adjusted regression model (g computation) for the association between `hbp` and `race`. Use the bootstrap to obtain standard errors.

- Use this marginally adjusted approach to estimate the **risk difference** and the **risk ratio** for the association between race and high blood pressure.

b. Create **two** indicator (dummy) variables for the three level income variable. Make the referent level `income` greater than level 20 (which makes the referent category an income of $\geq$ \$20,000)

c. Fit a regression model for `hbp` regressed against `race`, the two indicators for `income`, `marital`, `age` and `school`. Conduct two likelihood ratio tests for the interactions between race and the income variables.

Step 3b: CDM, Structural Transformation

a. Regress hbp against race, the two income indicators, the interaction between race and the indicator that income is between categories 11 and 17, as well as marital, age, and school. Use linear and log-linear models for risk differences and ratios.

b. Extract the coefficients from these models for each income category and the interaction between race and income.

c. Create transformed outcomes on the linear and log-linear scales

$$\tilde{Y} = Y - \hat{\beta} \times \text{income} - \hat{\beta} \times \text{income} \times \text{race}$$

$$\tilde{Y} = Y \times \exp(-\hat{\beta} \times \text{income} - \hat{\beta} \times \text{income} \times \text{race})$$

d. Estimate the CDM using linear and log-linear models with the transformed outcomes

e. Bootstrap to get standard errors

Step 3c: CDM, inverse probability weighting

a. Regress hbp against race, the two income indicators, the interaction between race and the indicator that income is between categories 11 and 17. Weight this model with the stabilized IP weights constructed earlier.

b. Estimate standard errors for the race coefficient from this model using the robust variance estimator.

Step 5: Combine the Results

a. Combine all the results into a single table

b. Export the table to a csv file

Step 5: Combine the Results

	Estimate	Std Error	LCL	UCL
ATE RD: Marginally Standardized	0.10	0.03	0.04	0.15
CDM RD: Structural Transformation	0.05	0.04	-0.02	0.13
CDM RD: IP Weighted	0.07	0.05	-0.03	0.18
----------------------------------	--------	-----------	------	-----
ATE RR: Marginally Standardized	2.30	0.20	1.55	3.41
CDM RR: Structural Transformation	1.72	0.33	0.90	3.32
CDM RR: IP Weighted	2.04	0.40	0.94	4.44

Phase 3: Interpretation

Step 6: Why are the CDMs different from the ATEs?

	mu1	mu0
ATE: Marginally Standardized	0.171	0.074
CDM: Structural Transformation	0.128	0.075
CDM: IP Weighted	0.168	0.068

- For the structural transformation method, mu1 goes down but mu0 stays the same.
    - Suggests having high income reduces risk of high blood pressure for "black or other"
- For IP weighting, both mu1 and mu0 go down

Step 7: Interpretation

- The risk difference for the association between race and high blood pressure is 0.10 (95% CI: 0.04, 0.15).

- The risk difference that would be observed if everyone had an income of $20,000 (in 1971 dollars) or more is between 0.05 (95% CI: -0.02, 0.13; structural transformation) and 0.07 (95% CI: -0.03, 0.18; IP weighting).

For every 100 people in the population, the "black or other" group had ten more cases of high blood pressure relative to the "white" group. Had everyone in the population had an income of $20,000 (in 1971 dollars) or more, the "black or other" group would have had 5 more cases of high blood pressure relative to the "white" group.

- Exercise: write out a paragraph interpreting the risk ratios.

Q&A

Project 2: Counterfactual Disparity Measure for Race, Income, and High Blood Pressure in the NHEFS

Ashley I Naimi, PhD
Associate Professor
Emory University

✉ ashley.naimi@emory.edu
ainaimi

EMORY | ROLLINS SCHOOL OF PUBLIC HEALTH