*GPH-GU 2363 Causal Inference Final Project*

**Causal Relationship Between Average Education Level and Life Expectancy**

**Across Different Countries**

*Yimin Yuan (yy4413) Xinyu Diao (xd2182)*

**Introduction**

Life expectancy is a statistical measure of the estimate of the span of a life and life expectancy at birth (LEB) is the most commonly used measurement. This demographic indicator estimates the average number of years a newborn is expected to live under current mortality levels.1 It is calculated from the life table.2 During the decades, worldwide LEB has increased from 66.8 years in 2000 to 73.4 years in 2019.3 LEB can be influenced by a wide range of factors like disease, environment, and sociological factors. This study aims to evaluate the education level's impact on LEB to determine to what extent education can affect LEB across different countries.

**Step 1: Data Preparation**

The life expectancy dataset of this study comes from the Global Health Observatory (GHO) data repository under World Health Organization (WHO) from 2000 to 2015 (<https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>). Health factors for 193 countries were collected. The initial sample size is 2,938, after dropping the missing values, the final sample size is 1,649. The outcome variable of this study is `life expectancy`, which indicates the LEB of a country in age. The treatment variable is `schooling`, which indicates the number of average years of schooling in the country. We set 12-year as the cutoff, countries with an average years of schooling less than 12 is recoded as 0 (controlled group), otherwise, recoded as 1 (treated group). Predictors we used are `year`, `adult mortality`, `infant deaths`, `alcohol`, `percentage expenditure`, `Hepatitis B`, `Measles`, `BMI`, `under five deaths`, `Polio`, `total expenditure`, `Diphtheria`, `HIV/AIDS`, `GDP`, `population`, `thinness 1-19 years`, `thinness 5-9 years`, and `income composition of resources` (Details can be found in Appendices 1).

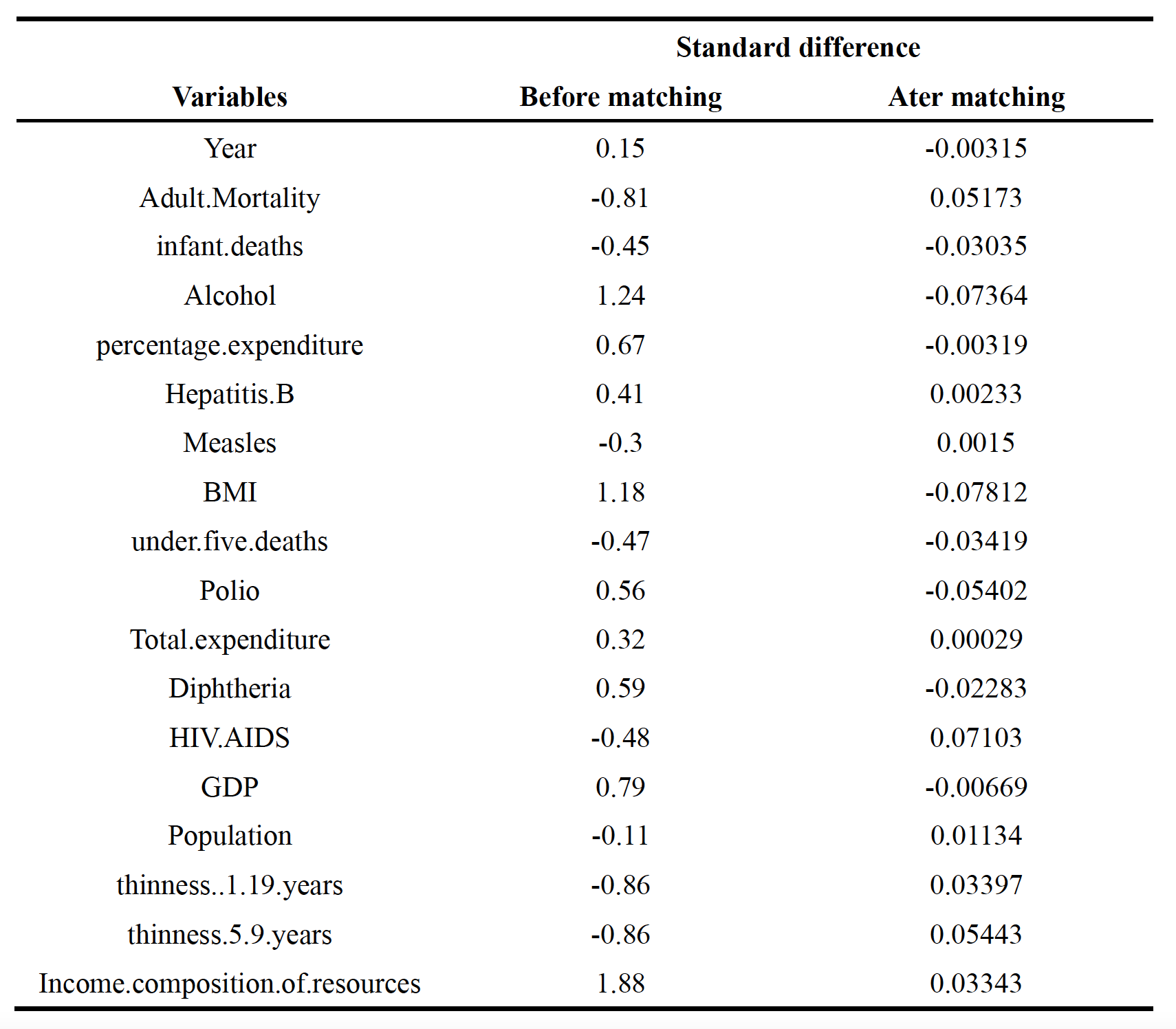
**Step 2: Full Matching**

First of all, we fit a propensity score model on the treatment variable using logistic regression and then matched based on the propensity score. Since we have 895 treated groups but only 754 controlled groups, we chose full matching as our matching method.

**Step 3: Check the Standardized Difference**

After full matching, we checked the standard difference before and after the matching to check if matching could solve the problem of measured confounders. Before matching, all variables have absolute standardized difference greater than 0.1, and after matching, all standardized differences are smaller than 0.1 (Table 1), which indicates that there are no significant differences in these predictors between the control group and the treated group and any observed differences in the outcome are not due to these potential confounders.

Table 1 Standard differences of variables before and after matching



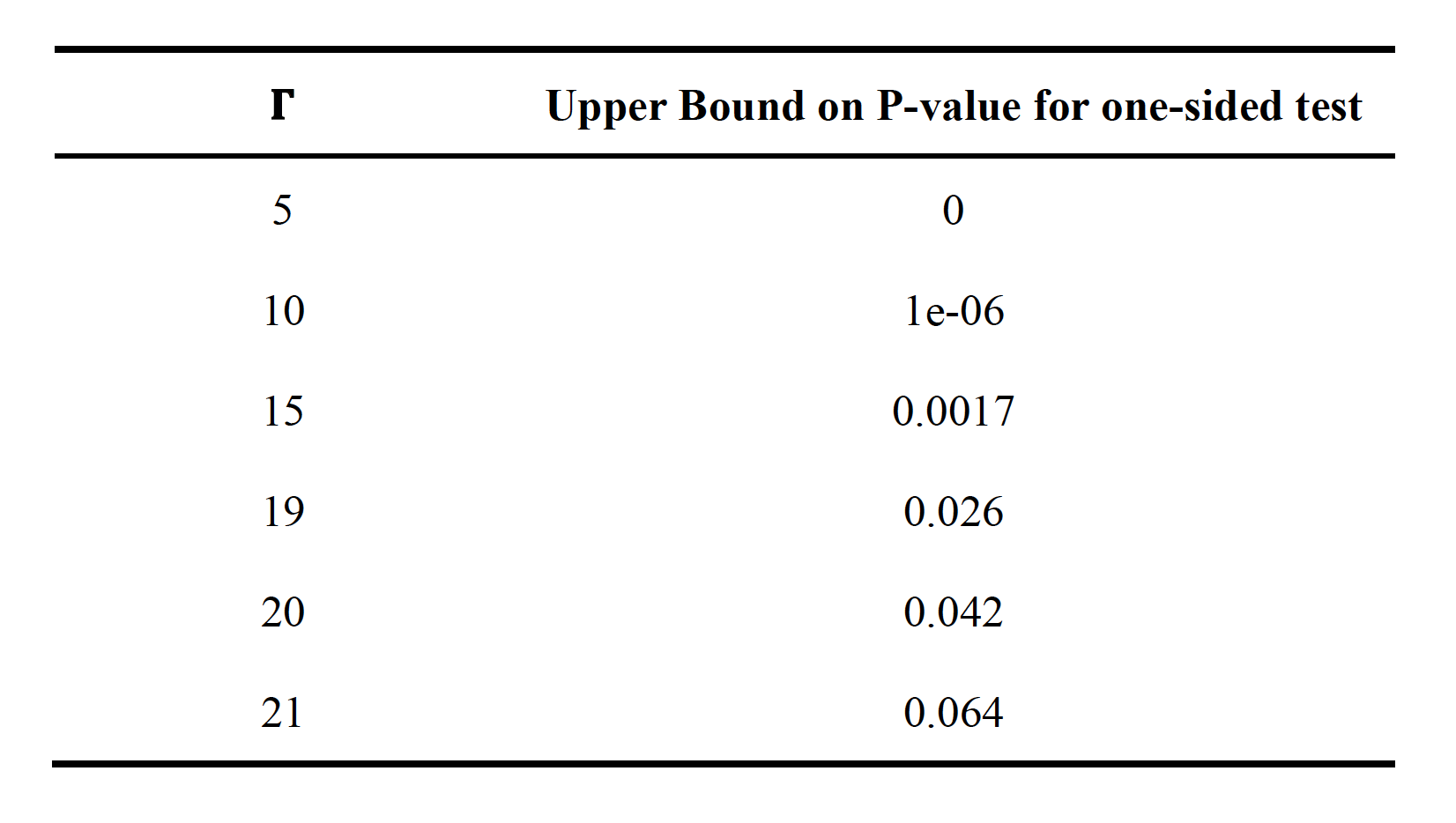
**Step 4: Check the Effect of Education on Life Expectancy**

After full matching, we utilized the M-test randomization inference approach here to test whether there is a difference of life expectancy between the treatment group and the control group. The reason for selecting M-test here is that M-statistics have better performance on heavily tied data. The point estimate is 0.98, with a 95% confidence interval of (0.081, 1.885), which indicates that countries with an average schooling year exceeding 12 years have an increase in life expectancy of 0.98 years on average.

**Step 5: Sensitivity Analysis**

At last, we generated a sensitivity analysis for the M statistics to see how the unmeasured biases may affect the causal effect of education level. Table 2 shows the upper bounds of P-values for different values of 𝚪.

Table 2 Result of sensitivity analysis



For 𝚪 up to 20 , there is still evidence that higher education level improves life expectancy. The large value of 𝚪 suggests that there may be some unmeasured covariates that produce biases in the result.

**Conclusions and Limitations**

The study's findings suggest a positive causal association between a country's average schooling age exceeding 12 years and a longer life expectancy, with countries meeting this criterion having an average of 0.98 years longer life expectancy compared to those with less than 12 years of schooling. This result reveals the potential health benefits associated with longer education years at country level.

Nevertheless, this study faces several limitations. Firstly, the dataset's variables exhibit a high degree of correlation, as education, healthcare, and economic indicators tend to be interrelated in countries. Additionally, the sensitivity analysis suggests the presence of unmeasured covariates that may influence the observed causal associations. Further examination on other covariates outside this dataset may be needed to generate a more precise conclusion.

**References**

1. WIKIPEDIA. Life Expectancy. Accessed December 20, 2023. https://en.wikipedia.org/wiki/Life\_expectancy

2. CDC. Life Expectancy. Accessed December 20, 2023. https://www.cdc.gov/nchs/nvss/life-expectancy.htm

3. World Health Organization. GHE: Life expectancy and healthy life expectancy. Accessed December 20, 2023. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy

**Appendices**

**1 - Detail information of variables**:

Outcome**:**

`Life expectancy`: life expectancy at birth in age.

Treatment**:**

`Schooling`: Number of average years of schooling. If the average number of schooling years is less than 12 years, then recoded as 0, otherwise, 1.

Predictors**:**

`year`: year of the data collected.

`Adult mortality`: adult mortality rates of both sexes (probability of dying between 15 and 60 years per 1,000 population).

`infant death`: number of infant deaths per 1,000 population.

`alcohol`: recorded per capita (15+) consumption (in litres of pure alcohol).

`percentage expenditure`: expenditure on health as a percentage of Gross Domestic Product per capita (%).

`Hepatitis B`: Hepatitis B immunization coverage among 1-year-olds (%).

`Measles`: number of reported cases per 1,000 population.

`BMI`: average body mass index of entire population.

`under-five deaths`: number of deaths under 5 years per 1,000 population.

`Polio`: Polio immunization coverage among 1-year-olds (%).

`Total expenditure`: general government expenditure on health as a percentage of tatal government expenditure (%).

`Diphtheria`: diphteria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-old (%).

`HIV/AIDS`: deaths per 1,000 live births HIV/AIDS (0-4 years).

`GDP`: gross domestic product per capita in USD.

`Population`: population of the country.

`Thinness 1-19 years`: prevalence of thinness among children and adolescents for age 10 to 19 (%).

`Thinness 5-9 years`: prevalence of thinness among children for age 5 to 9 (%).

`Income composition of resources`: human developmen index in term of income composition of resources (range from 0 to 1).

**2 - Codes**

Step 1. Data Preparation and Cleaning

```{r}

library(MASS)

library(optmatch)

life\_expectancy <- read.csv("/Users/diaoxinyu/Document/NYU/23 Fall/GU2363 Causal Inference/Final Project/Life Expectancy Data.csv")

#View(life\_expectancy)

life\_exp <- na.omit(life\_expectancy)

summary(life\_exp)

dim(life\_exp)

dim(life\_exp[life\_exp$Schooling >= 12,])

life\_exp$Schooling >= 12

# recode if schooling >= 12,scho\_over\_12 = 1, else = 0 --- treatment

life\_exp$scho\_over\_12 <- as.factor(ifelse(life\_exp$Schooling >= 12, 1, 0))

table(life\_exp$scho\_over\_12) # 895 data with 1, 754 with 0

```

Step 2: Full matching (because there are not enough control groups to do pair matching or matching with multiple groups)

```{r}

# Fit a propensity score using logistic regression

propscore.model <- glm(scho\_over\_12 ~ Year + Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP + Population + thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources, family = binomial, data = life\_exp)

summary(propscore.model)

life\_exp$outcome <- life\_exp$Life.expectancy

life\_exp$treated <- propscore.model$y

treated <- life\_exp$treated

life\_exp$logit.ps <- predict(propscore.model)

diff.propensity.score.mat <- outer(life\_exp$logit.ps[life\_exp$treated==1], life\_exp$logit.ps[life\_exp$treated==0],'-')

distmat.propensity <- abs(diff.propensity.score.mat)

rownames(distmat.propensity) <- rownames(life\_exp)[life\_exp$treated==1]

colnames(distmat.propensity) <- rownames(life\_exp)[life\_exp$treated==0]

# Matrix of covariates, excluding intercept

Xmat <- model.matrix(scho\_over\_12 ~ Year + Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Hepatitis.B + Measles + BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP + Population + thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources, family = binomial, data = life\_exp)

Xmat <- Xmat[,-1]

head(Xmat)

# Full Matching

matchvec <- fullmatch(distmat.propensity, data = life\_exp)

life\_exp$matchvec <- matchvec

# Number the strata

matchedset.index=substr(matchvec,start=3,stop=10)

matchedset.index.numeric=as.numeric(matchedset.index)

```

Step 3: Check the standardized difference before and after a full match

```{r}

# Calculate standardized difference before and after a full match

# Drop observations with missing values from the calculations

# stratum.myindex should contain strata for each subject, 0 means a unit was not

# matched

# Use harmonic mean weights

standardized.diff.harmonic.func=function(x,treatment,stratum.myindex,missing=rep(0,length(x))){

xtreated=x[treatment==1 & missing==0];

xcontrol=x[treatment==0 & missing==0];

var.xtreated=var(xtreated);

var.xcontrol=var(xcontrol);

combinedsd=sqrt(.5\*(var.xtreated+var.xcontrol));

std.diff.before.matching=(mean(xtreated)-mean(xcontrol))/combinedsd;

nostratum=length(unique(stratum.myindex))-1\*max(stratum.myindex==0);

if(max(stratum.myindex==0)==0){

stratumlist=sort(unique(stratum.myindex))

}

if(max(stratum.myindex==0)==1){

templist=sort(unique(stratum.myindex))

stratumlist=templist[-1]

}

diff.in.stratum=rep(0,nostratum);

number.in.stratum=rep(0,nostratum);

harmonic.weight=rep(0,nostratum)

for(i in 1:nostratum){

if(sum(stratum.myindex==stratumlist[i] & treatment==1 & missing==0)==0 | sum(stratum.myindex==stratumlist[i] & treatment==0 & missing==0)==0){

number.in.stratum[i]=0

}

if(sum(stratum.myindex==stratumlist[i] & treatment==1 & missing==0)>0 & sum(stratum.myindex==stratumlist[i] & treatment==0 & missing==0)>0){

diff.in.stratum[i]=mean(x[stratum.myindex==stratumlist[i] & treatment==1 &

missing==0])-mean(x[stratum.myindex==stratumlist[i] & treatment==0 &

missing==0]);

number.in.stratum[i]=sum(stratum.myindex==stratumlist[i])

harmonic.weight[i]=1/(.5/sum(stratum.myindex==stratumlist[i] &

treatment==1)+.5/sum(stratum.myindex==stratumlist[i] & treatment==0))

}

}

std.diff.after.matching=(sum(harmonic.weight\*diff.in.stratum)/sum(harmonic.weight))/combinedsd;

list(std.diff.before.matching=std.diff.before.matching,std.diff.after.matching=std.diff.after.matching);

}

std.diff.before=rep(0,ncol(Xmat));

std.diff.after=rep(0,ncol(Xmat));

names(std.diff.before)=names(Xmat[1,]);

names(std.diff.after)=names(Xmat[1,]);

for(i in 1:ncol(Xmat)){

missing.temp=is.na(Xmat[,i])

temp.stand.diff=standardized.diff.harmonic.func(Xmat[,i],life\_exp$treated,matchedset.index.numeric,missing.temp);

std.diff.before[i]=temp.stand.diff$std.diff.before.matching;

std.diff.after[i]=temp.stand.diff$std.diff.after.matching;

}

sd.bf=std.diff.before

sd.af=std.diff.after

options(digits=2)

cbind(sd.bf,sd.af)

```

Step 4: Check the effect of schooling on life expectancy

```{r}

# M-test

# Put data into format for senfmCI function

library(sensitivityfull)

stratum.myindex=matchedset.index.numeric

nostratum=length(unique(stratum.myindex))-1\*max(stratum.myindex==0);

if(max(stratum.myindex==0)==0){

stratumlist=sort(unique(stratum.myindex))

}

if(max(stratum.myindex==0)==1){

templist=sort(unique(stratum.myindex))

stratumlist=templist[-1]

}

treated1=rep(0,nostratum)

stratumsize=rep(0,nostratum)

for(i in 1:nostratum){

stratumsize[i]=sum(stratum.myindex==stratumlist[i])

}

y=matrix(rep(NA,nostratum\*max(stratumsize)),nrow=nostratum)

for(i in 1:nostratum){

no.treated.in.stratum=sum(stratum.myindex==stratumlist[i] & life\_exp$treated==1)

no.control.in.stratum=sum(stratum.myindex==stratumlist[i] & life\_exp$treated==0)

treated.in.stratum=which(stratum.myindex==stratumlist[i] & life\_exp$treated==1)

control.in.stratum=which(stratum.myindex==stratumlist[i] & life\_exp$treated==0)

if(no.treated.in.stratum==1){

y[i,1]=life\_exp$outcome[treated.in.stratum]

y[i,2:(no.control.in.stratum+1)]=life\_exp$outcome[control.in.stratum]

treated1[i]=1

}

if(no.treated.in.stratum>1){

y[i,1]=life\_exp$outcome[control.in.stratum]

y[i,2:(no.treated.in.stratum+1)]=life\_exp$outcome[treated.in.stratum]

treated1[i]=0

}

}

treated1=as.logical(treated1)

senfmCI(y,treated1)

## schooling >= 12, on average 0.98 more life expectancy

```

Step 5: Test unmeasured confounders - Sensitivity analysis

```{r}

library(sensitivitymv)

# Create a data frame with columns for treated and control life expectancy

matched\_data <- data.frame(

Treated\_Life\_Expectancy = life\_exp\_treated,

Control\_Life\_Expectancy = life\_exp\_control\_matched[life\_exp$treated == 1]

)

senmv(sensitivity\_dataset\_matched,gamma = 1)

senmv(sensitivity\_dataset\_matched,gamma = 5)

senmv(sensitivity\_dataset\_matched,gamma = 10)

senmv(sensitivity\_dataset\_matched,gamma = 15)

senmv(sensitivity\_dataset\_matched,gamma = 18)

senmv(sensitivity\_dataset\_matched,gamma = 19)

senmv(sensitivity\_dataset\_matched,gamma = 20) # 0.042

senmv(sensitivity\_dataset\_matched,gamma = 21) # 0.064

```