Risky Assets and Household Investment Behaviors: The Causal Effect of Income on Risk Preferences

Generated by Deepai

Zixia Huang

June 02, 2024

1. Introduction

Finance theory predicts that rational households are expected to allocate their risky assets in a well-diversified portfolio (Markowits, 1952; Rubinstein, 2002). Understanding the extent to which households conform to this prediction is crucial for regulating risky financial products and promoting financial well-being. In recognizing the importance, policymakers have pledged many policies to shore up individual confidence and support the recovery of investment in the risky financial markets, and explore how national institutional factors impact households' investment decisions in 15 countries(Apergis & Bouras, 2023).

Yet, there is a research gap to explore how households' socio-economic characteristics influence risky asset investment behavior, and effectiveness of this policy is a novel subject that deserves thorough research. Existing research suggests that higher accessibility to financial inclusion encourages households to invest in risky assets (Bian et al., 2023). Households that have medical insurance exhibit a higher probability of raising their position within the national financial asset distribution, and are more inclined to invest in the risky financial assets (Liu et al., 2022). Household financial decisions are complex, interdependent, and heterogeneous.

Therefore, this research aims to explore what socio-economic factors drive household risky asset investment behavior, and take China as an example. **In particular, this research would like to study the causal effect of households' annual income on the households risky investment behavior.** Although there is a large amount of existing studies on households' limited participation in financial consumption and investment, few previous studies focus on the causal effect of annual income level on the household's risky assets portfolio.

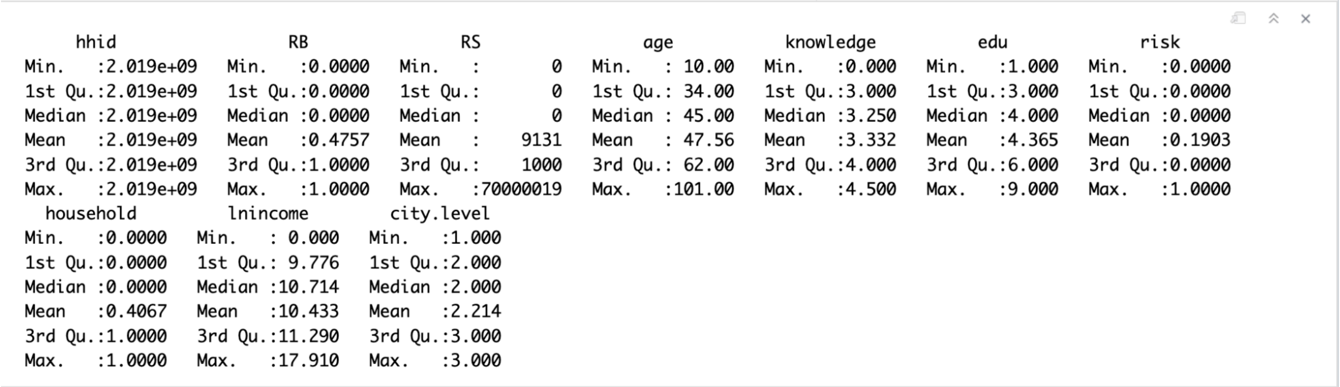
2. Data

This research uses the 2019 Chinese Household Financial Survey (CHFS) dataset, which explores income and expenditure, social programs and insurance involvement, preferred financial tools, and investment behaviors with socio-demographic characteristics(CnOpenData, 2019). The dataset consists of 34,643 households and provides different geo-spatial information across provinces, districts and counties, and villages (neighborhoods) in China. Xu (2024), Bian et al. (2023), and Liu et al. (2022) utilize the same dataset to explore household income allocation.

This research will bring some results from the 2019 CHFS survey and convert to several variables for our research question. Since there are lots of incomplete responses and missing values, our research team conducted further data cleaning and have 17,149 valid households for analysis use. Table 1 and Figure 1 show the definition of several variables from the dataset and include descriptive statistics as below.

Table 1. Variable Definition and Data Manipulation in the 2019 CHFS Survey

Variables	Description
hhid	Household ID
RB	RB could see a household's willingness to own/participate in stocks, bonds, financial derivatives and non-RMB assets. This research sums up D3103 (Stock), D4103 (bonds), D6100a (financial derivatives), D8104 (non-RMB assets) columns for each household. Then, we convert it to a binary variable. If the sum is larger than 0, then RB=1. otherwise RB=0
Age	Age is defined as the median age in the household. This research uses the A2005 column in the dataset to compute the age of each family member.Then, we take the median of the age in each household.
Knowledge	Knowledge means the level of a household's financial knowledge. The CHFS asked a series of questions related to financial investment and its related knowledge, and this research made an index using the responses from column H3101 and D9203, for each household. The range is from 0 to 4.5. When a household has a larger number of the index, it represents the household has more financial knowledge.
Education	The educational level of a household. Education is defined as the highest level of education that one household ever reached. We use data from column A2012, but change primary school and below into 1, junior high school =2, senior high school =3, and college and above =4.
Risk	Risk means the household representative's risk preference. We use the H3104 column and code it as a binary variable, setting the responses below 3 equals 1(prefer risk), and answers 3 and above as 0(avoid risk).
Household type	Household type represents where the household is located. We use columns A2022 and A202201, and then set rural type=0, and non-rural type(city) =1.
City level	City level tries to categorize what level of city a household lives in. We use the column A2016, and A2019, and we make categorical variables to group a city that the survey respondent at. 1 means level 1 city, 2 means level 2 city, and 3 means level 3 city.
income	Income means a household annual income. We sum up all individual values in the A2023lc column within one household.



```
      hhid                RB               RS                 age           knowledge          edu              risk       
 Min.   :2.019e+09   Min.   :0.0000   Min.   :        0   Min.   : 10.00   Min.   :0.000   Min.   :1.000   Min.   :0.0000  
 1st Qu.:2.019e+09   1st Qu.:0.0000   1st Qu.:        0   1st Qu.: 34.00   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:0.0000  
 Median :2.019e+09   Median :0.0000   Median :        0   Median : 45.00   Median :3.250   Median :4.000   Median :0.0000  
 Mean   :2.019e+09   Mean   :0.4757   Mean   :     9131   Mean   : 47.56   Mean   :3.332   Mean   :4.365   Mean   :0.1903  
 3rd Qu.:2.019e+09   3rd Qu.:1.0000   3rd Qu.:     1000   3rd Qu.: 62.00   3rd Qu.:4.000   3rd Qu.:6.000   3rd Qu.:0.0000  
 Max.   :2.019e+09   Max.   :1.0000   Max.   :70000019   Max.   :101.00   Max.   :4.500   Max.   :9.000   Max.   :1.0000  
   household         lnincome        city.level   
 Min.   :0.0000   Min.   : 0.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.: 9.776   1st Qu.:2.000  
 Median :0.0000   Median :10.714   Median :2.000  
 Mean   :0.4067   Mean   :10.433   Mean   :2.214  
 3rd Qu.:1.0000   3rd Qu.:11.290   3rd Qu.:3.000  
 Max.   :1.0000   Max.   :17.910   Max.   :3.000  
```

Figure 1. Descriptive statistics of the 2019 CHFS Survey

3. Method

3.1 Causal effect identification

To estimate the causal effects, we first leverage what is designated as the arbitrary low-income cutoff, set at 6,450 yuan by The State Council of the People's Republic of China (2023), to estimate the local average treatment effect (LATE) in a fuzzy regression discontinuity design (RDD). Our analysis uses a bandwidth ranging from 5,400 to 7,000 yuan. Preliminary analyses indicate that there are 193 households below this threshold and 111 above it. We plan to employ alternative bandwidth specifications for robustness checks. Due to the specific nature of our research design, our findings cannot be generalized to the population beyond our sub-sample.

The independent variable in this study is the annual income level, which serves as the running variable (X) in our RDD framework. The specific income threshold used to define the treatment and control groups will be determined based on a policy or natural cutoff that influences financial behavior (D). The dependent variable captures various aspects of risk-related financial behavior, including participation in stocks, bonds, financial products, financial derivatives, and non-RMB assets. This is treated as a binary indicator (RB) where 1 denotes participation in risky assets investment and 0 denotes non-participation.

To isolate the effect of income on risk preferences, we will control for a set of variables that may confound the relationship between the independent and dependent variables. These control variables include: age, knowledge, education level, city level, risk preference and household type.

3.2 Regression model of the variables

Since our outcome variable is binary, we used logistic regression to explore the potential effects of earnings. We believe that the treatment only affects the level of the outcome, not the rate of change, so the interactive term between the treatment and running variable is not included in these models. We used several models with different sets of independent variables and covariates. The following four models are considered:

1. $ln(\frac{P(RB=1|D,X,Z)}{P(RB=0|D,X,Z)}) = \beta_0 + \beta_1 D + \beta_2 X,$

Where we do not control for any covariates.

2. $ln(\frac{P(RB=1|D,X,Z)}{P(RB=0|D,X,Z)}) = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 Z,$

where $Z = \{age, knowledge, edu, household, city.level, risk\}$ (full model)

3. $ln(\frac{P(RB=1|D,X,Z)}{P(RB=0|D,X,Z)}) = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 Z,$

Where $Z = \{age, knowledge, edu, household, city.level\}$

4. $ln(\frac{P(RB=1|D,X,Z)}{P(RB=0|D,X,Z)}) = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 Z,$

Where $Z = \{knowledge, edu, household\}$

We first fit the logistic regression models for different sets of variables. The first model only includes the treatment variable and the running variable income. The second model considers all the variables, including the treatment variables, the running variable and other covariates. The third model excludes the household's risk preference, while the fourth model excludes age of the household, risk preference and the city level. By comparing the estimates of these models, we could find the conditional independence between the variables. We choose the best model based on the goodness of fit tests.

In order to further explore and have a visualization on the causal relationship between those variables, after the model selection, an RDD plot calculated from the selected model is considered.

4. Results

4.1. Estimation of the logistic regression model

The estimates of the logistic regression models for the variable are shown in Table 2. From the table, we can see that the significant coefficient for the treatment variable indicates that there is a statistically significant difference in the risky assets investment behavior between the treatment and control groups at the cutoff. From the second column of the table where we have the regression result when we control for all covariates, p-value corresponding to the household's risk preference is relatively large, suggesting the relationship between the outcome and city level may not be linear inside the logistic regression, so we cut it out in model 3. Same reasoning also applied when we change model 3 into model 4. Our preferred model is model 4, which looks at the effects of annual income on household risky assets investment behaviors, controlling for household type, educational level and financial knowledge. We do not detect any statistically significant effects of earnings on investment behaviors. Again this may be due to the true functional form between a household's annual income and the risky assets investment behavior is not linear inside the logistic regression. Also another reason is the bandwidth that we choose is too wide which may violate the assumption of similarity between units just above and just below the cutoff. Therefore, different models with non-linear assumptions are tested in section 4.2. And different bandwidths are used to check the robustness in section 4.3.

Table 2. Model estimates of the logistic regression

	Model 1:	Model 2:	Model 3:	Model 4:
Treatment(cutoff)	1.813***	1892***	1.896***	1.911***
	(0.582)	(0.590)	(0.602)	(0.598)
Annual income	-0.0005	-0.0004	-0.0004	-0.0004
	(0.0005)	(0.0006)	(0.0006)	(0.0005)
Age of the household		0.002	0.002	
		(0.010)	(0.010)	
Household's type		-0.398	-0.400	-0.400
		(0.401)	(0.400)	(0.400)
Risk preference		0.073		
		(0.795)		
Education level		0.396***	0.396***	0.383***
		(0.128)	(0.128)	(0.111)
Financial knowledge level		0.333*	0.333*	0.332*
		(0.186)	(0.186)	(0.186)
City level		0.017	0.016	
		(0.170)	(0.170)	
Note:				*p<0.1; **p<0.05; ***p<0.01

4.2 Non-linear assumptions between the running variable and outcome

No significance is detected on the previous four models we use with linear assumptions between the running variable annual income and the log-odds of risky assets investment behavior. We now try two more different models with non-linear assumptions.

The first additional model we fit is a polynomial model with degree two on the running variable annual income, in order to detect any significant quadratic relationship between the log-odds and the running variable. The model 5 we used:

5. $$ln(\frac{P(RB=1|D,X,Z)}{P(RB=0|D,X,Z)}) = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 X^2 + \beta_4 Z$$

,Where $Z = \{knowledge, edu, household\}$.

Finally, we do not wish to make a specific functional form assumption, so we use non-parametric methods. For the second additional model, we use splines to flexibly model the relationship between the running variable and the log-odds. The model 6 we used:

6. $$ln(\frac{P(RB=1|D,X,Z)}{P(RB=0|D,X,Z)}) = \beta_0 + \beta_1 D + \sum_{i=1}^{4} \beta_i B_i(X) + \beta_4 Z$$

, fitted with B-splines with order 4, Where $Z = \{knowledge, edu, household\}$.

The estimates of the logistic regression models with non-linear assumptions are shown in Table 3. From the table, we can see that the non-parametric B-spline method failed to detect any significance relationship between the risky assets investment behavior to any of the covariates including the treatment variable cutoff. However, The model with polynomial of order 2 actually performs better than the original model where we assume linear relation between the running variable and the log-odds. The result suggests a possible quadratic relationship between the running variable and the log-odds of the outcome, after controlling for the treatment effect and other confounders.

In other words, the effect of the household's annual income on the log-odds of participating in risky assets investment is not constant but changes at a rate that also depends on the amount of annual income itself.

Table 3. Model estimates of different functional form

	Model 4:	Model 5:				Model 6:	
Treatment(cutoff)	1.911**	1.691*				2.42***	
	(0.598)	(0.936)				(0.676)	
Annual income	-0.0004	1	2	3	4	1	2
		2.1	-0.6	2.6	0.4	-7.442	-4.826*
	(0.0005)	(2)	(1.2)	(2.2)	(1.4)	(5.5)	(2.6)
Age of the household							
Household's type	-0.400	-0.453				-0.462	
	(0.400)	(0.407)				(0.406)	

Risk preference			
Education level	0.383**	0.382***	0.390***
	(0.111)	(0.112)	(0.112)
Financial knowledge level	0.332*	0.331*	0.349*
	(0.186)	(0.189)	(0.187)
City level			
Note:	*p<0.1; **p<0.05; ***p<0.01;”1,2,3,4” is the order of corresponding functional form		

4.3 Alternative bandwidth for robustness check

The choice of bandwidth (i.e., the range of the running variable around the cutoff used for the analysis) can affect the results. A bandwidth that is too wide may violate the assumption of similarity between units just above and just below the cutoff, while a bandwidth that is too narrow may lead to imprecise estimates due to a small sample size. In this section we will consider both cases. First we choose a narrower bandwidth according to the original data set, with households having annual income 5800 to 6800. The sub-data set has 185 observations in total. Then a wider bandwidth obtained from the full data set is considered, with households having annual income 5100 to 7400. Finally, an even wider bandwidth with a household's annual income 4500 to 8000 is chosen to further explore the robustness of our model.

The estimates of the preferred model with alternative bandwidth are shown in Table 4. From the table, we can see that with narrower bandwidth we failed to detect any significance relationship between the risky assets investment behavior to any of the covariates of interests. This is an imprecise estimate due to lack of samples because considering the results from a wider bandwidth as shown in the table, it suggests our results from the primary model is robust, with the polynomial of annual income of order 2 significant at 0.01 level. However, the estimates from the model with the largest bandwidth are misleading. The abnormal significance at 0.01 level from the covariate household type suggests this size of the bandwidth may violate the assumption of similarity between units just above and just below the cutoff. The suddenly improved performance of the model is purely due to the excessively increasing sample size.

Table 4. Model estimates of alternative bandwidth

	Narrow:		middle:		Wider:	
Treatment(cutoff)	2.440***		2.872***		3.149***	
	(0.946)		(0.634)		(0.509)	
Annual income	1	2	1	2	1	2
	-1.74	-2.79	-16.2***	-12.1***	-32.5***	-22.6***
	(4.80)	(2.6)	(6.01)	(2.98)	(7.31)	(3.76)
Age of the household						
Household’s type	-0.088		-0.370		0.726***	
	(0.450)		(0.390)		(0.133)	
Risk preference						
Education level	0.282*		0.416***		0.482***	
	(0.145)		(0.110)		(0.035)	
Financial knowledge level	0.496**		0.334***		0.459***	
	(0.238)		(0.181)		(0.079)	
City level						
Note:	*p<0.1; **p<0.05; ***p<0.01;”1,2,3,4” is the order of corresponding functional form					

4.4 Results and local average treatment effect (LATE)

As the RDD plot shows, we can detect a dramatic cutoff between the control and treatment group corresponding to the threshold annual income 6450 yuan. Based on our preferred model (model 6), the causal effect of annual income on household risky assets investment behavior is identified by the estimated local treatment effect, which suggests 30% more chances of investment in risky assets if the household is not identified as a low-income household.

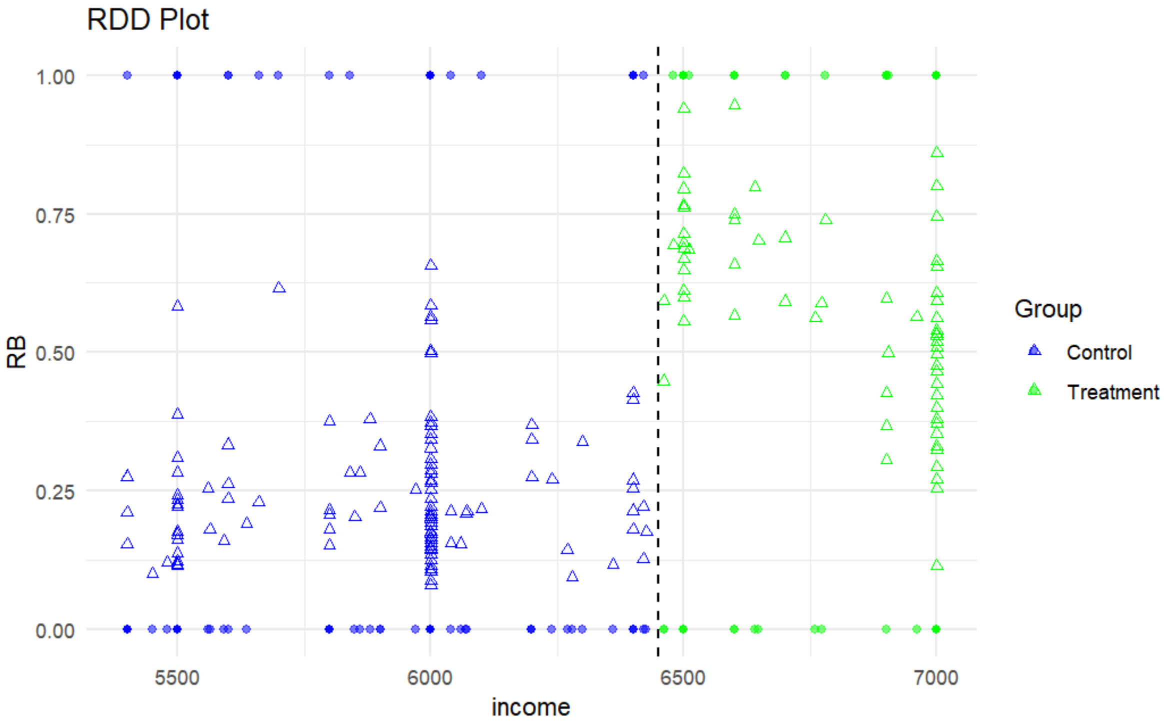


Figure 2. RDD Plot based on model 6

5. Conclusion

5.1 Discussion

This study utilized 2019 data to analyze the causal effect of household income on risky asset investment behaviors in China. However, data from 2022 and the impact of the COVID-19 pandemic on household income and investment behaviors would undoubtedly be of greater interest. Future research should incorporate these data to gain a more comprehensive understanding of how economic fluctuations and public health emergencies influence financial decision-making over the long term.

Additionally, there are latent variables in our study that were not fully considered. These include household expectations of future income and confidence in financial markets. These latent variables may significantly impact household investment behavior and should be included in subsequent research.

5.2 Limitations

Some limitations also need to be noted. Firstly, the data sample is from a specific year (2019) and is limited in size, which may introduce data bias. The households in the sample may not fully represent the situation of all households in China, thus limiting the external validity of the results. Secondly, key macroeconomic indicators such as GDP and interest rates were not included in this study. These indicators may significantly influence household investment behaviors, and their exclusion could lead to an incomplete understanding of the factors at play. Future research should consider incorporating these macroeconomic variables to provide a more comprehensive explanation of risky asset investment behaviors. Thirdly, our analysis did not fully account for households' expectations regarding future economic conditions and income changes. These expectations can significantly influence current investment decisions. Quantifying and including these expectation factors in future research is crucial for a more accurate assessment of the causal effects of household income on investment behaviors.

Despite these limitations, this study provides important insights into the impact of household income on risky asset investment behaviors in China. By utilizing a fuzzy regression discontinuity design (RDD), we were able to preliminarily identify the causal effect of income thresholds on investment behaviors. Model 4 indicates that controlling for household type, educational level, and financial knowledge provides robust results. However, no statistically significant effects of earnings on investment behaviors were detected, possibly due to non-linear relationships between household income and investment behaviors or the wide bandwidth used.

The exploration of non-linear models, including polynomial and non-parametric B-spline methods, our preferred model 6 suggested a potential quadratic relationship between income and the likelihood of investing in risky assets. Additionally, robustness checks with alternative bandwidths highlighted the importance of selecting appropriate bandwidths to avoid bias and ensure accurate estimates.

Overall, this study underscores the complexity of household financial decision-making and the need for further research to incorporate more recent data, account for latent variables, and include broader macroeconomic indicators. Future studies should aim to refine the models and methods used to provide a more comprehensive understanding of the factors influencing household investment behaviors in risky assets.

Reference

Apergis, Nicholas, and Christos Bouras. 2023. “Household Choices on Investing in Financial Risky Assets: Do National Institutional Factors Have Their Own Merit?” *International Journal of Finance & Economics* 28 (1): 405–20. https://doi.org/10.1002/ijfe.2427.

CnOpenData. 2019. “CHFS Chinese Household Financial Survey Data CHFS China Household Financial Survey Data.” https://www.cnopendata.com/en/data/m/survey/chfs.html.

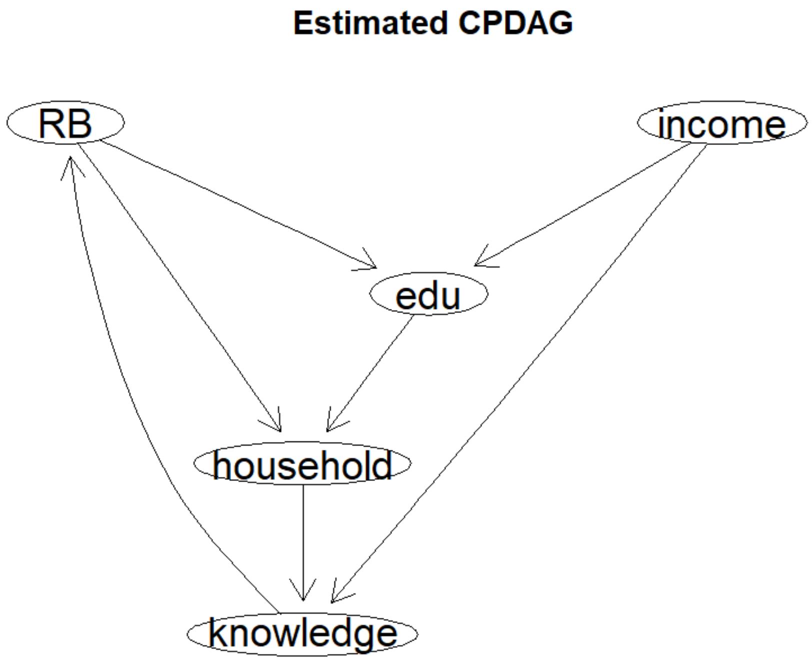
Markowitz, Harry. 1952. “Portfolio Selection.”

Rubinstein, Mark. 2002. “Markowitz’s ‘Portfolio Selection’: A Fifty-Year Retrospective.” *The Journal of Finance* 57 (3): 1041–45. https://doi.org/10.1111/1540-6261.00453.

Xu, Xinzhe. 2024. “The Impact of the Investment Expectation Gap on Households’ Risky Financial Asset Investment.” *Investment Management and Financial Innovations* 21 (1): 331–42. https://doi.org/10.21511/imfi.21(1).2024.25.

Appendix

1. PC-DAG for the full data



2. Relevant R code

~~~
```{r}
data <- read.csv("https://drive.google.com/file/d/1bIGHW8_tsi7nqwMyx5xckX6tbz6kxSJo/view?usp=drive_link",header = TRUE)
data <- data[,-1]
 # Create the treatment variable
data$data_win <- ifelse(data$income > 6450, 1, 0)

  # Tabulate the treatment variable
  table(data$data_win)

  hist(data$income)
  ~~~
  ```{r,warning=FALSE}
```

```
library(ggplot2)
# Specify the cutoff
cutoff <- 6450

# Run a logistic regression model with the outcome as the dependent variable and the treatment and running variable as independent variables
model <- glm(RB ~ data_win + income, data = data, family = binomial())
summary(model)
```

```{r}
# full model
model <- glm(RB ~ data_win +age+ income+household+risk+edu+knowledge+city.level, data = data, family = binomial())
summary(model)
```

```{r}
# eliminate risk
model <- glm(RB ~ data_win +age+ income+household+edu+knowledge+city.level, data = data, family = binomial())
summary(model)
```

```{r}
# eliminate age and city.level
model <- glm(RB ~ data_win+ income+household+edu+knowledge, data = data, family = binomial())
summary(model)
```

```{r}
# Predict the probabilities using the logistic regression model
data$predicted_prob <- predict(model, type = "response")

# Create a new dataframe for plotting
data$data_win <- as.factor(data$data_win)
plot_data <- data

# Plot the observed outcomes and predicted probabilities
ggplot(plot_data, aes(x = income)) +
  geom_point(aes(y = RB, color = data_win), alpha = 0.5) +
```

~~~
   geom_point(aes(y = predicted_prob,color=data_win), pch = 2)+
  geom_vline(xintercept = cutoff, linetype = "dashed") +
  scale_color_manual(values = c("blue", "green"), labels = c("Control", "Treatment")) +
  labs(x = "income", y = "RB", color = "Group", title = "RDD Plot") +
  theme_minimal()
```

```{r}
#Polynomial with order 2
model <- glm(RB ~ data_win+poly(income, degree = 2)+edu+household+knowledge, data = data, family = binomial())
summary(model)
```

```{r}
#non-parametric B-splines
library(splines)

model <- glm(RB ~ data_win + bs(income, df = 4)+edu+household+knowledge, data = data, family = binomial())
summary(model)
```

```{r}
# Predict the probabilities using the logistic regression model
data$predicted_prob <- predict(model, type = "response")

# Create a new dataframe for plotting
data$data_win <- as.factor(data$data_win)
plot_data <- data

# Plot the observed outcomes and predicted probabilities
ggplot(plot_data, aes(x = income)) +
  geom_point(aes(y = RB, color = data_win), alpha = 0.5) +
   geom_point(aes(y = predicted_prob,color=data_win), pch = 2)+
  geom_vline(xintercept = cutoff, linetype = "dashed") +
  scale_color_manual(values = c("blue", "green"), labels = c("Control", "Treatment")) +
  labs(x = "income", y = "RB", color = "Group", title = "RDD Plot") +
  theme_minimal()
```

```{r}
~~~

```
#LATE
trt <- subset(data, income > 6450)
con <- subset(data, income<6450)
mean(trt$predicted_prob)-mean(con$predicted_prob)
```

```{r}
#Robustness check

#Smaller bandwidth
sub_small <- subset(data, income >= 5800 & income <= 6800)
sum(sub_small$income<6450)
```

```{r}
data_full <- read.csv("https://drive.google.com/file/d/1-JYPxuRh6BLMRDbgW_E_LpNJHKJ35wmI/view?usp=drive_link",header = TRUE)

#Lager bandwidth
sub_big <- subset(data_full, income>=5100& income<= 7400)
model <- glm(RB ~ data_win+poly(income, degree = 2)+edu+household+knowledge, data = sub_small, family = binomial())
summary(model)
```

```{r}
sub_big$data_win <- ifelse(sub_big$income > 6450, 1, 0)
model <- glm(RB ~ data_win+poly(income, degree = 2)+edu+household+knowledge, data = sub_big, family = binomial())
summary(model)
```
```{r}
#More Lager bandwidth
sub_big2 <- subset(data_full, income>=4600& income<= 7900)
sub_big2$data_win <- ifelse(sub_big2$income > 6450, 1, 0)
model <- glm(RB ~ data_win+poly(income, degree = 2)+edu+household+knowledge, data = sub_big2, family = binomial())
summary(model)
```

```{r,message=FALSE,warning=FALSE}
#Implement PC DAG
install.packages("BiocManager")
```

```
library("Rgraphviz")
library("RBGL")
library("abind")
library("corpcor")
library("sfsmisc")
library("robustbase")
library("pcalg")
library("graph")

plotcpdag <- "Rgraphviz" %in% print(.packages(lib.loc = .libPaths()[1]))
```

```{r}
##### Using the PC Algorithm to alternative bandwidth 2:
data_full <- read.csv("https://drive.google.com/file/d/1-JYPxuRh6BLMRDbgW_E_LpNJHKJ35wmI/view?usp=drive_link",header = TRUE)
sub_big2 <- subset(data_full, income>=4600& income<= 7900)
dag <- data.frame(RB=sub_big2$RB,income=sub_big2$income,edu=sub_big2$edu,knowledge=sub_big2$knowledge,household=sub_big2$household)
n <- nrow(dag)
p <- ncol(dag)
indepTest <- gaussCItest
suffStat <- list(C=cor(dag), n = n)

## estimate CPDAG
alpha <- 0.1
pc.fit <- pc(suffStat, indepTest, p = p, alpha = alpha, verbose = TRUE)
showAmat(pc.fit)

if (plotcpdag) {
  plot(pc.fit, main = "Estimated CPDAG",labels=c("RB","income","edu","knowledge","household"))
  ## Note undirected edges are represented here as  <->
}
```