MIZUHO

One MIZUHO

Mizuho Economic Outlook & Analysis

October 20, 2020

Determinants of households' risk assets holding

The causal structure estimation through machine learning

< Summary >

- ◆ Using individual data from the *Japan Household Panel Survey* by Keio University, this paper attempts to identify the factors influencing households' risk asset holdings in Japan by inferring causal structure through machine learning techniques.

- ◆ The results of the analysis suggest that there are two primary channels through which households own risk assets. The first channel is to hold risk assets when households have a surplus in their savings (deposits) considering livings after retirement. As deposits are influenced by income, debt (such as mortgages), and age (life cycle), these factors are indirectly related to risk assets holding. The second channel is risk preference: risk assets are purchased if households have a higher risk preference. The analysis implies that the higher risk preference is formulated mainly through a higher level of financial literacy.

- ◆ The low ratio of risk assets holding seems to be mainly attributed to households' financial concerns after retirement. Although policy efforts are required, with the background of the declining birthrate and aging society in Japan, it is highly challenging to resolve such anxieties completely. Therefore, it will be imperative to enhance financial literacy—the second channel. Japan's financial literacy is relatively low compared to other countries, and it has much room for improvement. The government and financial institutions need to promote and facilitate financial education actively.

Shinya Kotera, Senior Economist,
Economic Research Department
Mizuho Research Institute Ltd.
shinya.kotera@mizuho-ri.co.jp

1. Introduction

This report conducts an empirical analysis to examine what factors (variables) would influence the households' holding of risk (financial) assets in Japan using individual data from the *Japan Household Panel Survey* by Keio University.[1] In particular, the report attempts to estimate a causal structure among various variables by using machine learning techniques and simulate which policy interventions effectively increase households' risk assets holding.

Compared to other countries, households in Japan tend to have a high proportion of cash & deposits and fewer risk assets. According to international comparison based on the Flow of Funds Statistics as of the end of March 2020, risk assets accounted for 13.0% of total financial assets in Japan, compared to 44.8% in the United States (US) and 25.9% in Europe, i.e., the ratio of the US is more than three times than that of Japan.[2] Further, according to the *National Survey of Family Income and Expenditure* conducted by the Ministry of Internal Affairs and Communications, only 23.5% of households owned securities in 2014, which means that households holding risk assets are in the minority.

Such sluggish household investments in risk assets are also regarded as a problem by the government. The Financial Services Agency (2017) reported that the low proportion of risk assets holding, which could lead to small investment returns, was likely to be the reason for the weaker growth rate of household financial assets in Japan than in the US and the United Kingdom (UK). Then, they emphasized the importance of shifting from "saving (cash & deposits) to investment" to help households' financial asset building and supply adequate funds for economic growth.

To encourage households to invest in risk assets and create a virtuous cycle of funds, it is necessary, in the first place, to understand the structures or mechanism of the risk assets holding by households. According to the classical theory such as Merton (1969), as long as the expected rate of return on risk assets exceeds the interest rate on safe assets, it is optimal for households to hold risk assets, regardless of their risk preferences. However, since most households in Japan do not hold any risk assets, the model fails to explain reality appropriately. The theoretical model can only be applied under certain simplified assumptions, and naturally, it cannot explain what is happening if these assumptions are different from reality.

Therefore, this report seeks to identify the factors that influence risk assets holding by reducing prior constraints and allowing the machine to learn from data. Through letting the machine learn the "patterns" of risk assets holding behaviors from observational data,

---

[1] The data used for this analysis, Japan Household Panel Survey (JHPS/KHPS), was provided by the Panel Data Research Center at Keio University. Any possible errors in this report are attributable to the author.
[2] See Bank of Japan (2020) for details. Here, risk assets are defined as the sum of equity and investment trusts.

more realistic models/factors are expected to be established or identified. Once the variables that directly (or indirectly) influence the risk assets holding are identified, one can change the variables that government or companies can intervene and simulate the effectiveness of these interventions in increasing the ratio of households holding risk assets. The identification of the causal structure and estimation of intervention effects are the ultimate objectives of this paper.

This report's machine learning technique consists of a two-step process, as adopted by the Bank of Japan (2017). The purpose of the first step is to compute each variable's "importance" using a method called Random Forest (RF). These importance values suggest which variables have more predictive power toward the explained variable (that is, whether households will own risk assets or not). However, RF can only discover the patterns between variables, and their relationships (causal structures) are not identified. Therefore, as a second step, the network structure graph is estimated by employing a machine learning technique called Bayesian network (BN). BN allows us to infer the causal relationship among variables and predict which factors could influence the risk assets holding.

The remainder of the report is organized as follows. Section 2 summarizes the previous studies (mainly empirical ones) and organizes factors used for the analysis. Section 3 describes the data and presents a list of variables to be used based on Section 2. Section 4 gives an overview of the two machine learning techniques used in this report. Section 5 presents the analysis results, which include estimates of a causal structure, simulation of interventions, and robustness check. The final section summarizes the findings.

2. Previous studies

The theoretical background on the holding of risk assets in households is summarized in works such as Iwaisako (2012), Shioji et al. (2013), and Ito et al. (2017). Under the assumption of a general utility function, it is always desirable to hold risk assets, albeit the amount of which may vary, as long as the expected rate of return on risk assets exceeds the interest rate on safe assets. However, since many households do not own risk assets in reality, some factors, including labor income, real estate, and entry costs, have been addressed to explain the discrepancy between the theoretical model and the reality. Labor income is a factor that increases the preference for safe assets when there is uncertainty in future income, as households avoid risks due to the motive for precautionary saving. Uncertainty about income varies according to attributes such as occupations and employment status, and they may also influence the portfolio choices. Further, real estate (housing) has lower liquidity than other types of assets, and a large balance of mortgage loans impose fixed payment obligations for households. These features work to discourage households from owning risk assets. Entry cost refers to cost in a broad sense and includes not only monetary cost such as transaction fees but also the cost of acquiring financial knowledge and the cost of gathering requisite information.

Based on the above-mentioned theoretical background, let us briefly review the empirical studies on households' portfolio choices using Japan's data. Ito et al. (2017) conducted an empirical analysis using panel data from Osaka University. They reported that risk assets holding was significantly impacted by classical theory factors[3], constraints in liquidity, motives for precautionary savings (due to unemployment risk and future anxieties), and entry costs (such as financial literacy). Shioji et al. (2013) conducted an empirical analysis via repeated cross-sectional data and showed that a significantly robust factor for stock ownership was the balance of financial assets. Kitamura and Uchino (2011), analyzing the repeated cross-sectional data, pointed out that 30 to 50% of the differences in risk assets holding between university-graduated and non-university-educated households could be explained by differences in their attributes (especially the size of the employer, income level, and amount of financial assets). They also argued that since there were still many portions that remained unexplained, the presence of entry costs (such as financial literacy) was implied. Kinari and Tsutsui (2009) referred to the survey data of the Postal Services Research Institute to conduct Tobit regression using the risk asset ratio (from 0 to 100) as explained variables. They reported that, in addition to the factors suggested by the Capital Asset Pricing Model (CAPM) (risk tolerance and expected rate of return), variables such as financial assets, income, age, confidence in banks and securities firms, and behavioral bias (overconfidence) were also having a significant influence in a manner consistent with the theory. Iwaisako et al. (2019) examined the possibility that equity ownership was negatively affected by homeownership and found that while higher home values increase equity ownership, higher mortgages were likely to lower equity ownership. Nakajo et al. (2017) reported that factors discouraging ownership of risk assets included disappointing equity market performance in Japan and inadequate life planning, such as the lack of a concrete picture of livings after retirement. Fukuhara (2016) investigated reasons for differences in risk asset holdings ratios between Japan and the US and suggested backgrounds such as liquidity constraints due to real estate ownership and institutional factors including defined contribution pension plans and mutual funds.

To summarize the empirical analysis mentioned above, the following 12 factors could be addressed as those that may be related to the holding of risk assets: (1) basic attributes (such as age and education), (2) employment status, (3) financial asset balance, (4)

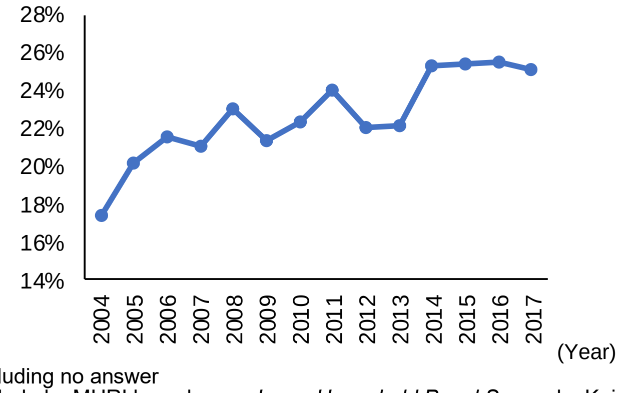
[3] That is, expected return on risk assets, an interest rate of safe assets, market volatility, and relative risk aversion.

precautionary saving, (5) liquidity constraints, (6) income, (7) real estate (homeownership), (8) entry costs (such as financial literacy), (9) confidence in financial institutions, (10) risk preference, (11) institutional factors, and (12) expected rate of return on risk assets. The next section provides an overview of the data used in this report and tries to select variables that can capture (or approximate) these factors.

3. Data

The data used for the analysis in this report comes from the *Japan Household Panel Survey*, collected by the Panel Data Research Center at Keio University.[4] The survey is conducted annually from 2004, adding new cohorts and filling in dropped samples. The sample size of 2017 is 4,626. In this paper, risk assets are defined as securities (shares, bonds, stock investment trusts, corporate & public bond investment trusts, loans in trust, money in trust, etc.) mentioned in the *Japan Household Panel Survey*. **Figure 1** illustrates the time series in the percentage of households holding risk assets according to the survey. The percentage has been on a moderate upward trend over the medium term but remained flat at around 25% since 2014. According to the *National Survey of Family Income and Expenditure* by the Ministry of Internal Affairs and Communications, 23.5% of the total households held securities in 2014. The holding ratio identified by the *Japan Household Panel Survey* is slightly higher, but it is at about the same level.

Figure 1: Ratios of households holding securities



Note: Excluding no answer
Source: Made by MHRI based upon *Japan Household Panel Survey* by Keio University

Based on the factors addressed in the previous section, 37 variables (as presented in **Table 1**) are selected for the analysis. Institutional factors and the expected rate of return on risk assets are excluded, as the survey data does not contain appropriate proxy variables for them, and these factors can be regarded as commonly applied to all households in the

[4] The details of the survey can be found at https://www.pdrc.keio.ac.jp/en/paneldata/datasets/jhpskhps/

case where only data from Japan are used. Although the risk aversion variables are often calculated from the way compensations are received (as suggested by Barsky et al. (1997)), there is no such question in the *Japan Household Panel Survey*, and this paper selects variables that are close to the risk preference. In particular, variable 33 (Preference) in **Table 1** is derived as follows. The survey asked respondents to rank the six financial products (stocks, stock investment trusts, corporate bonds, government bonds, bank deposits, and postal savings) to invest if they have 3 million yen to spare. The preference variable is calculated as the average ranking of stocks, stock investment trusts, and corporate bonds. Although this may not precisely capture the risk preference itself, this paper regards a respondent is risk aversion if the value is high (that is, the averaged ranking is low).

While the *Japan Household Panel Survey* follows the same households each year, which allows us to conduct dynamic panel analysis, this paper adopted it as cross-sectional data by averaging multiple years, as shown in **Table 1**. This is because some variables are not surveyed every year, and not a small number of missing values (no answer) are found in the items that are surveyed every year. Hence, the use of an average would increase the data's stability, and basically, the three-year average of the year 2015–2017 for each household is used. However, note that some variables are two-year averages (average of 2015 and 2016) or are pre-2015 surveyed items. Besides, the values for 2016 are used for variables that cannot be expressed in order of applicability (category answers), such as occupation.

As for the data-processing approach, firstly, since the household head's attributes are often used to capture household characteristics, the analysis here also follows this approach. However, as respondents of the *Japan Household Panel Survey* are not necessarily household heads, the sample is used if the surveyed person or the spouse of the surveyed person is the household head. Secondly, samples are excluded if the "Asset-holding" status, which is the primary concern of this report, is unknown. Third, the missing values are imputed using the algorithm called missForest,[5] as the method has been reported to have the highest accuracy in imputation (Waljee et al., 2013).[6] However, samples with four or more missing values among the 37 variables shown in **Table 1** are excluded from the analysis. The final number of samples processed as above is 4,267.

[5] R package, randomForestSRC was used. The missForest (Stekhoven and Bühlmann, 2012) multivariate version was applied.

[6] Further, when performing missForest, the variables that might be useful for imputation are added to the 37 variables mentioned in **Table 1**. To be specific, the added variables are such as the time when the house was purchased, the purchase price of the house, lot size, distance from the station, and amount of consumption.

Table 1: List of variables

		Variable name	Description	Average used	Type
Objective variable	0	Asset-holding	Holding / Not holding risk assets (securities) (2 choices)	2015-2017	F
Basic	1	Education	Final academic background of the household head (6 choices: Junior high school, senior high school, junior collage, four-year university, graduate school, and other)	2016	F
	2	Age	Age of the household head	2016	N
Work-related	3	Work status	Current work status of the household head (6 choices: Not working, unemployed, self-employed, non-regular employeee, regular employee, and executive officer)	2016	F
	4	Occupation	Occupation of the household head (7 choices: Note 1)	2016	F
	5	Industry	Industry of work of the household head (9 choices: Note 2)	2016	F
	6	Earned income	Previous years' annual income of the household head from the main job (ten thousand yen, before any deduction)	2015-2017	N
	7	Working hours	Working hours of the household head (number of hours, weekly average)	2015-2017	N
Precautionary savings	8	Anxiety 1	Been dissatisfied with the present life (4 levels)	2015-2017	O
	9	Anxiety 2	Felt anxiety over the future (4 levels)	2015-2017	O
	10	Satisfaction 1	Satisfaction level of household income (11 levels)	2015-2017	O
	11	Satisfaction 2	Satisfaction level of life overall (11 levels)	2015-2017	O
	12	Sufficiency	Have sufficient income and assets to live without any problems during the period after the mandatory retirement (5 levels)	2015 and 2016	O
	13	Pension	Will receive (receiving) pension payments that will be enough for the post-retirement life (5 levels)	2015 and 2016	O
Financial assets	14	Deposits	Deposits amount (unit: 10,000 yen)	2015-2017	N
Income	15	Household income	Previous year's household income (ten thousand yen, before any deduction, income from the sale of assets not included)	2015-2017	N
Liquidity constraints	16	Debt	Total present household borrowings balance (unit: 10,000 yen)	2015-2017	N
Housing-related	17	Housing	House ownership (3 choices: Rental housing, owned, and corporate housing)	2016	F
	18	House value	Market value of the owned house (unit: 10,000 yen)	2015-2017	N
	19	Land value	Market value of the owned plot (unit: 10,000 yen)	2015-2017	N
	20	Real asset	House value + Land value (unit: 10,000 yen)	2015-2017	N
	21	Mortgage	Total mortgage loan balance (unit: 10,000 yen)	2015-2017	N
	22	Plan	Housing purchasing plan (4 choices: None, detached house, condo, and other)	2016	F
Entry cost	23	Literacy 1	We should spend money now if the interest rate is 10% and the inflation rate is 20% (5 levels)	2015 and 2016	O
	24	Literacy 2	The price of government bonds that have a yield of 10,000 yen after one year should be 10,000 yen (5 levels)	2015 and 2016	O
	25	Internet	Number of devices connected to the internet (4 levels)	2015 and 2016	O
	26	Securities 1	(Impressions of the securities market) Profits cannot be made with certainty. (5 levels)	2015 and 2016	O
	27	Securities 2	(Impressions of the securities market) There is a possibility of heavy loss. (5 levels)	2015 and 2016	O
	28	Securities 3	(Impressions of the securities market) One should not buy or sell by reacting to a temporary price change (5 levels)	2015 and 2016	O
Trust	29	Securities 4	Securities firms provide clients with advice that is useful for making investment decisions (5 levels)	2012-2014	O
	30	Securities 5	Securities firms suggest stocks with potentially substantial returns (5 levels)	2012-2014	O
Risk	31	Precipitation	When you go out to a place you have never been to before with your family or friends, what percentage of chance of rain makes you decide to take an umbrella? (4 levels)	2015-2017	O
	32	Conservative	I feel more comfortable with buying whatever I buy at a familiar shop (4 levels)	2015 and 2016	O
	33	Preference	Risk assets preference (Investment method preferred by the household when having spare in financial status: Note 3)	2015 and 2016	N
Other	34	Understanding 1	Understanding of the situation concerning income (5 levels)	2015 and 2016	O
	35	Understanding 2	Understanding of the situation concerning financial assets (5 levels)	2015 and 2016	O
	36	Rational	I have a reasonable life (3 levels)	2015 and 2016	O
	37	Discount	Time discount factor (8 levels: Note 4)	2015-2017	O

Notes: 1. Occupation is categorized into seven categories: salesperson/service worker, manager, clerical worker, transportation or communications worker/maintenance worker, mine worker/manufacturing worker, specialized or technical worker, and others.
2. The industries are categorized into nine categories: construction, mining/manufacturing, wholesale and retail/restaurants/accommodations/transportation, finance/insurance, real estate, information and telecommunications, medicine/welfare/education and learning support/public service, utilities/other services, and others.
3. The preference variable is derived based on the question of "If you have 3 million yen to spare, in which of the following financial products would you invest?". The variable value is the average ranking of three financial products (stocks, stock investment trusts, and corporate bonds).
4. "Discount" shows the answers to the question, "Instead of receiving 10,000yen one month later, at least how much would you like to receive 13 months later? "
5. Types are N (numeric): numeric value; O (ordered): Factors with order; and F (factor): Factors without order.
6. When the household head and his/her spouse's values are available, that of the household head is given priority.
7. Some answers (choices) with a small sample size have been merged with the other close answers.

Source: Made by MHRI based upon *Japan Household Panel Survey* by Keio University

4. Analysis method

The analysis strategy here is similar to the Bank of Japan (2017), where they conducted a machine learning analysis to uncover the mechanism of an inflation expectation by firms. The purpose of this report is to learn about the causal structure of risk assets holding through a machine learning technique known as the Bayesian Network (BN). However, the complex and lower interpretability structures may be estimated if one used all the variables that appear in **Table 1**, and we might not be able to understand the causal structure adequately. Therefore, before performing BN analysis, the analysis took an extra step (specifically, a learning method called Random Forest (RF)) to select more important variables for risk assets holding.

RF is a widely known method in machine learning and is famous for its high predictive performances.[7] The main reason for using RF is that the analysis method can calculate variable importance. Using this variable importance, one can infer which explanatory variables have a strong relationship with risk assets holding. The variable importance used here is called "permutation importance", and it is calculated by measuring how much the model's prediction error will increase when a certain variable is randomly shuffled among the samples. If a certain explanatory variable and risk assets holding are highly correlated, the RF's predictive performance will substantially decrease when the explanatory variable is randomly sorted across the samples. On the other hand, in the case of an unrelated explanatory variable, little change will be observed in prediction accuracy before and after the variable's shuffling, which indicates that the variable's importance for risk assets holding is low.

RF can suggest the importance of variables but not a causal structure between the variables. Therefore, the next step is to learn the BN using the variables with higher

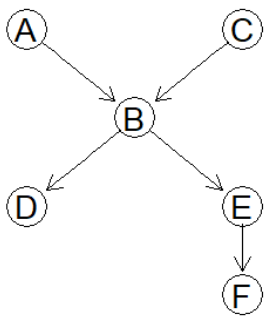
[7] The method builds multiple tree-structured models called decision trees and predicts the outcome based on the majority vote of these trees. In general, the RF chooses random $\sqrt{k}$ variables out of k independent variables to build a decision tree, and the paper also follows this approach. Details of the RF algorithm are available at works such as James et al. (2013).

importance obtained from RF analysis.[8] The BN applies the conditional probability test between variables and produces a graph called a Directed Acyclic Graph (DAG). For example, when the joint distribution of six random variables from A to F is shown below, the DAG can be expressed as **Figure 2**.

$$P(A)\ P(C)\ P(B|A,C)\ P(D|B)\ P(E|B)\ P\ (F|E)$$

Suppose if the government wants to influence D, but they can only intervene C or E. The RF will show both C and E are important variables (higher correlation) for D, but the DAG can indicate the government should intervene on C. In **Figure 2**, each variable (circle) is called a node, and the lines connecting the nodes are called an edge.

Figure 2: Example of a DAG



Source: Made by MHRI

The learning of the BN in this report consists of two steps. The first step is to learn a network structure, such as in **Figure 2** using the observational data (structure learning), leading to the next step of learning the conditional probability of each node given the network structure as learned in the previous step (parameter learning).

For structure learning, a method called the Peter & Clark (PC) algorithm is used. The algorithm is mainly employed to understand the causal structure between variables and is considered to fit this paper's purpose. The algorithm consists of two steps. In the first step, a graph referred to as a skeleton is estimated. The skeleton is the graph where the nodes are connected by edges but without any directions. Then, in the second step, the algorithm estimates the edges' directions given the skeleton structure. However, in the second step, it should be noted that the directions of all edges may not always be determined uniquely. Further, while the network structure in the PC algorithm is estimated via the repeated testing of conditional independence, each variable needs to be integrated into either a discrete or continuous variable. In the case of continuous variables, the distribution of each variable is often assumed to be Gaussian, but this assumption does not seem to hold, as

[8] The descriptions about BN in this report is based on the works by Scutari & Denis (2014), Hunermund & Bareinboim (2019), and Kalisch et al. (2012).

the variable of "Asset-holding", which is of interest in this report, is already binary. Hence, it seems more appropriate to perform the BN analysis by aligning all variables to discrete ones (i.e., continuous variables are discretized).[9]

The network structure like **Figure 2** will be learned via the above process; however, we must be cautious here such that the estimated structure should not be lightly interpreted as indicating a true causal relationship. To ensure that the network learned by the PC algorithm is a causal structure, the condition that there must be no latent variables is especially required.[10] However, latent variables are usually unknown, and it is quite challenging to control latent variables, especially in observational data. Although the true causal structure is unknown, it is pointed out that characteristics of BNs that are likely to represent a causal structure are fairly sparse and clearly interpretable (Scutari and Denis, 2014). It is essential not to believe in the results of learning simply but to carefully examine the validity of the learned causal structure, possibly with the help of experts' insight.

After a reasonable DAG has been estimated, each node's conditional probabilities are estimated (parameter learning). The analysis here adopts a maximum likelihood estimation methodology. These calculated probabilities are used to simulate the changes in ratios of households holding risk assets by interventions.

5. Result of analysis

(1) Random Forest

The variable importance identified by the analysis of RF is as follows.[11] The dependent variable is whether households hold the risk assets or not, and the independent variables are 37 variables, as shown in **Table 1**. The analysis uses two types of samples to ensure robustness: the whole sample (sample size: 4,267) and a sample in which the household head is employed (sample size: 3,298). The number of decision trees is 1,500. Further, to compare the results between the two samples easily, the variable importance is assessed in relative importance, with the highest value being 1.

The analysis result is presented in **Figure 3**, in which the top 15 variables are plotted in order of the average of the two outcomes of relative importance. The figure illustrates that the relative importance of "Preference" and "Deposits" is remarkably high. After that, "Sufficiency" is slightly higher than others, and the importance of fourth place down below is almost the same level. The result reveals that a balance of deposits (a safe asset) and risk preference are useful in predicting securities holding. Although previous studies

---

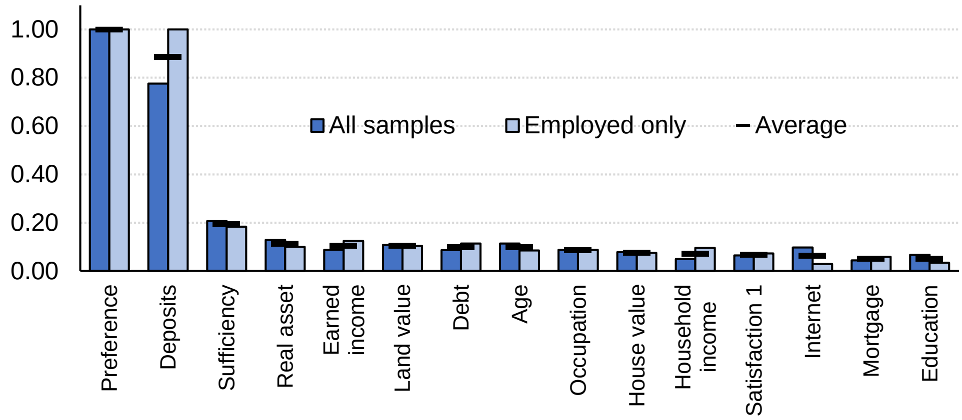
[9] Pearson's X2 test is used to test the conditional independence of the discrete variables. The significance level set for the test in this report is 5%.

[10] In addition to this condition, two other assumptions (known as the "faithfulness assumption" and the "causal Markov assumption") are required. However, these two assumptions are automatically violated if there are latent variables.

[11] For the analysis, the R package, randomForestSRC is used.

indicated the significant effects such as real estate, entry cost, education, and age, the RF analysis shows that these variables' predictive power is not exceptionally high.[12]

Figure 3: Relative importance (explained variable: "Asset-holding")



Note: The top 15 are plotted based on an average.
Source: Made by MHRI based upon *Japan Household Panel Survey* by Keio University

As the importance of "Preference" and "Deposits" is prominently high, let us analyze the two variables in more detail. Even variables that are not rated as having high importance in **Figure 3** may cause an indirect effect on risk assets holding by affecting these two variables. Hence, additional analysis (RF) is conducted, setting each of these two variables as explained variables and the remaining 35 variables (excluding "Preference" and "Deposits") as the explanatory variables. Again, RF is performed for both the whole sample and sample with employed. The number of decision trees is set at 1,500. The results of the (relative) variable importance of "Preference" and "Deposits" are plotted in **Figure 4 and 5**, respectively.

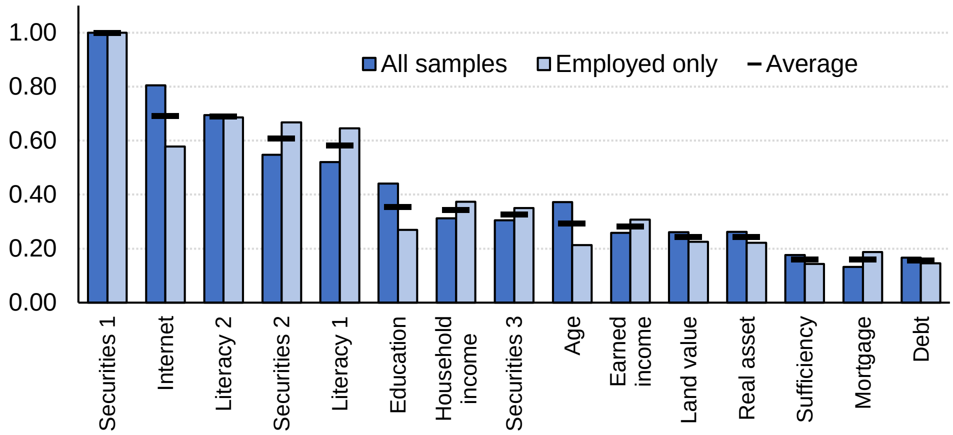
In terms of importance to "Preference", the variables such as "Securities 1 to 3", "Internet", "Literacy 1 and 2", "Education", "Age", and "Household income" have higher values. It seems that the variables related to the entry costs have high importance, and risk preference might be influenced by the perceptions of the securities market, financial literacy, and gathering information from the internet. Previous studies (e.g., Kitamura and Uchino (2011), van Rooij et al. (2011)) have pointed out that the significant effect of entry cost such as financial literacy on securities ownership, but the factor may not directly impact the risk assets holding but may have an indirect impact through risk preference.

By using "Deposits" as an explained variable, it can be spotted that "Sufficiency", "Household income", "Debt", and "Age" are the variables with higher importance. The variable with the highest importance on average is "Sufficiency", suggesting that the

[12] Note that RF analysis is not for testing significance.

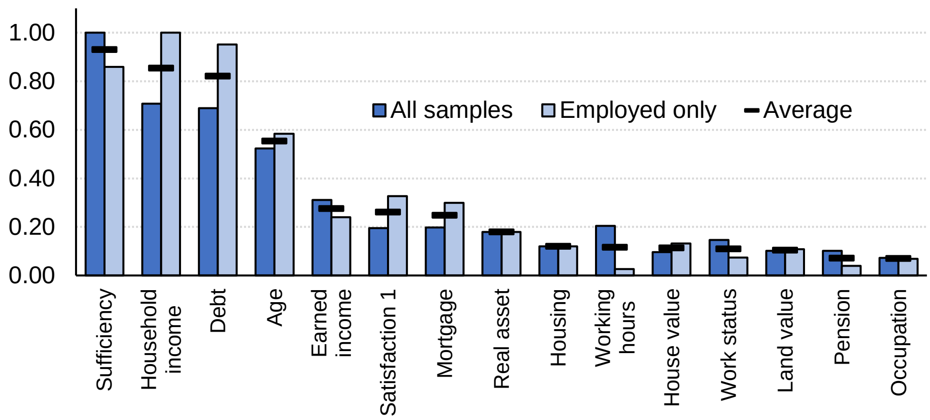
households' perception toward their financial status after retirement is closely related to the deposit balance. "Household income" and "Debt" are exceptionally high in the sample of employed only. Households with higher incomes are more likely to save money, and households with higher debt are less likely to save. As suggested by Iwaisako et al. (2019), the outstanding debt, including the mortgage loan, may discourage risk assets holding since the debt will prevent households from accumulating their saving (deposit).

Figure 4: Relative importance (explained variable: "Preference")



Note: The top 15 are plotted based on an average.
Source: Made by MHRI based upon *Japan Household Panel Survey* by Keio University

Figure 5: Relative importance (explained variable: "Deposits")



Note: The top 15 are plotted based on an average.
Source: Made by MHRI based upon *Japan Household Panel Survey* by Keio University

(2) Bayesian Network

Using the variables with high importance, the causal structure among these variables is estimated.[13] Although it is difficult to define a "high-importance" variable, this paper

[13] R package, pcalg is used for PC algorithm learning, and bnlearn is used for other BN analyses.

sees that a variable is high-important if an averaged relative importance is 0.3 or more in **Figure 3, 4, and 5**. The number of relevant variables is two from **Figure 3**, nine from **Figure 4**, and four from **Figure 5**. Adding the "Asset-holding" variable and eliminating duplicates, 14 variables are selected. Besides, to improve the interpretability of the causal structure and clarify the discussion, the variables of higher importance are used concerning "Securities 1 and 2" and "Literacy 1 and 2", as they may capture similar factors (that is, "Securities 1" and "Literacy 2" are used). Based on the above process, 12 variables are determined for the BN analysis.[14]

As mentioned in the methodology section, all variables must be either continuous or discrete variables in the BN analysis here. Based on the nature of the data, the variables are discretized. As the conditional independent test could not be adequately conducted (due to the sample size) if discrete variables have a high number of layers, and to clarify the discussions after the learning, all variables except "Asset-holding" are expressed in three levels (such as high, medium, and low) for discretization. Specifically, the continuous variables of "Deposits", "Household income", and "Debt" are divided into three levels via the algorithm by Hartemink (2001). The other variables are summarized into three levels, considering the sample size distribution and meaning of variables.

The PC algorithm does not guarantee that every edge's direction can be determined uniquely. In fact, after performing the PC algorithm, the directions of eight edges are found to be indistinguishable (see appendix 1 for the estimated DAG). Among them, five edges are related to "Education" and "Age", and these are the variables that represent the attributes of each household and are exogenous. Therefore, the edge's arrow should not point to these two variables, and the directions of these edges are determined accordingly. After this procedure, there are still three undirected edges, which are "Deposits ○–○ Debt", "Household income ○–○ Debt", and "Deposits ○–○ Household income". Among these, as for the "Household income"-related edges, "Household income → Deposits" and "Household income → Debt" seem to hold, since it is more natural to assume that households with higher income have a higher amount of deposits and can borrow higher amounts of money.[15] Although the relationship between deposits and debt is not necessarily straightforward, this paper assumes that the balance of deposits does not increase due to debt (repayment obligation) (that is, "Debt → Deposits"), as households are considered for prioritizing debt repayment over savings in general.

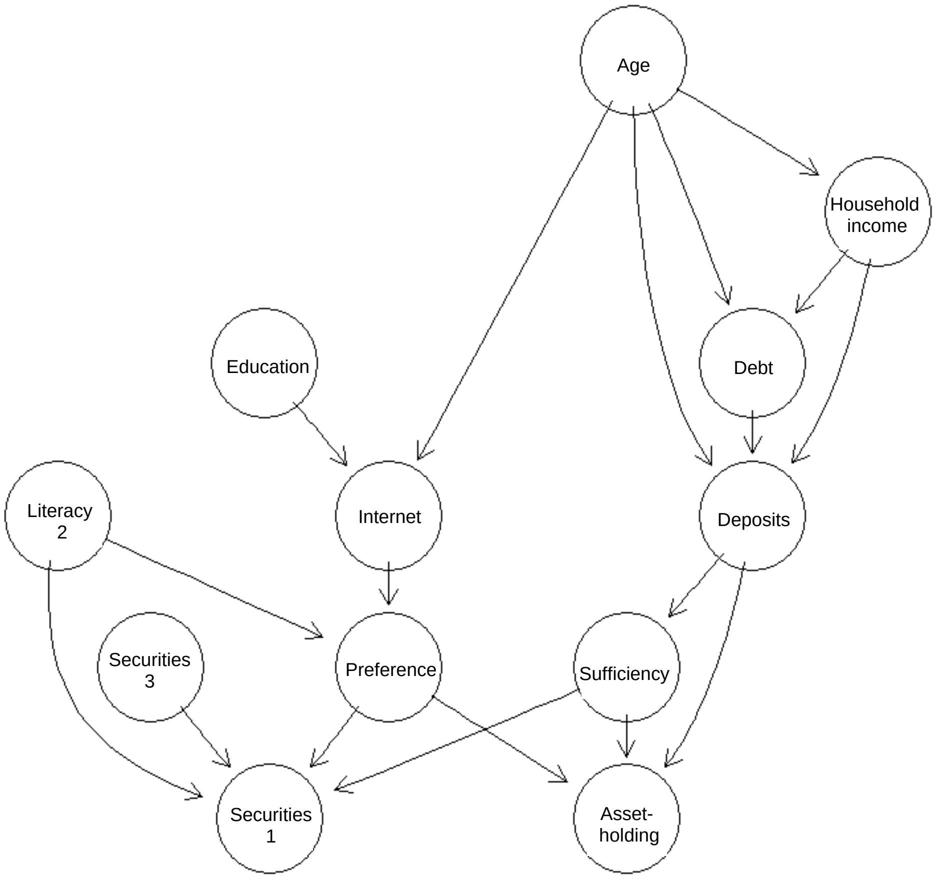
Based on the above process, the DAG with the final plotting is presented in **Figure 6**. As mentioned in the analysis method, the interpretation of the causal structure requires

[14] The 12 variables are Asset-holding, Education, Age, Sufficiency, Deposits, Household income, Debt, Literacy 2, Internet, Securities 1, Securities 3, and Preference.

[15] Usually, it is considered that the balance of deposits and debt do not determine the household income.

additional consideration, but as discussed below, the relationships between the variables are generally within a reasonable range. The author also asked professionals who have experience in customer service at a financial institution to examine the causal relationship of the estimated DAG, and they replied that they were comfortable with the graph. As such evaluations are highly subjective, they do not assure that the learned graph is free from problems. However, at least the learned DAG appears to capture one of the possible causal structures. The conditional probabilities at each node are also calculated with a maximum likelihood estimation, and posterior distributions at the main nodes are illustrated in appendix 2.

Figure 6: Bayesian Network (PC algorithm)



Source: Made by MHRI based upon *Japan Household Panel Survey* by Keio University

Figure 6 describes that three variables directly influence "Asset-holding", which are "Preference", "Deposits", and "Sufficiency". Although the balance of deposits and risk preference are the variables with the highest importance in the RF analysis, the third-ranked variable in the same analysis, "Sufficiency", may also directly impact risk assets holding.

The balance of deposits influences "Sufficiency", and household income and debt influence the balance of deposits. From the conditional probabilities, one can observe that households with higher household income tend to have a higher balance of deposits, and those with higher debt tend to have a lower balance of deposits. Considering these structures, as also suggested by Iwaisako (2012), risk assets seem to be purchased when households have funds to spare, and they do not tend to hold risk assets until they fully repay a debt (such as a mortgage) and secure enough safe funds for livings after retirement. Households with higher incomes and no debt (or fully repaid debt) have a higher probability of securing adequate safe assets, and this may be the reason for the significant effects of income and debt-related variables on risk assets holding in previous research. Although age does not have a direct influence on "Asset-holding", it has an impact on household income, debt, and deposit amount. It, therefore, forms the background against which a positive correlation (more precisely, a bell-shaped relationship) can be observed between age group (life cycle) and the risk assets holding ratio (Shioji et al., 2013).

Another factor influencing "Asset-holding" is risk preference, which is influenced by financial literacy and the number of devices used to access the internet. The result suggests that accurate knowledge of financial markets contributes to a higher preference for risk assets and that those that can collect information widely also have a higher preference for risk assets. The conditional probabilities of financial literacy for "Preference" suggest that, if literacy is high, the probability of "risk preference = high" is about 40%. On the other hand, when literacy is low, the probability is halved to about 19%. The conditional probability of "Internet" indicates that the probability of a high-risk appetite is about 28% when two or more devices for an internet connection are used, about 24% where only one device is used, and about 17% if no such device is used.

To summarize, there seem to be two main channels for the holding of risk assets. One is to hold risk assets when households have some spare money after securing sufficient funds (deposits) for livings after retirement. The balance of safe funds is influenced by household income and debt. Since sufficient deposits are not accumulated if households earn low income and have debt (including a mortgage), the probability of holding risk assets is inevitably lower in such households. The second channel is through risk preference. If a household has high financial literacy and is capable of gathering requisite information, the household can form a more-rational risk preference, and therefore the probability of holding risk assets is also increased.

(3) Effects of intervention

Given the above estimated DAG, one can simulate what kind of government policies would effectively increase the ratio of households holding risk assets. **Table 2** shows specific intervention assumptions conducting here. The simulation examines how the percentages of households holding risk assets will change when the probability distributions of the "Literacy 2", "Internet", and "Sufficiency" variables are modified. The purpose of interventions 1 and 2 is to increase the risk asset holding ratio by increasing risk appetite, with intervention 1 increasing financial literacy and intervention 2 increasing information literacy (use of the internet). As the government cannot control the balance of deposits, interventions 3 and 4 aim to change households' perceptions of financial availability after retirement (the "Sufficiency" variable). As for intervention 3, it is assumed, for example, that the government's drastic structural reforms will reduce uncertainty about pensions and increase expected economic growth, thereby making everyone perceive that "Sufficiency is high". Intervention 4 supposes that the households with the perception of "Sufficiency is low" will disappear (with all "Low" households shifting to "Medium"). For example, as pointed out by Nakajo et al. (2017), it is assumed that the intervention is made to reduce the number of households that are overly concerned about lack of money after retirement by encouraging long-term life planning (i.e., helping households build a concrete vision of livings after retirement).

Table 2: Assumptions

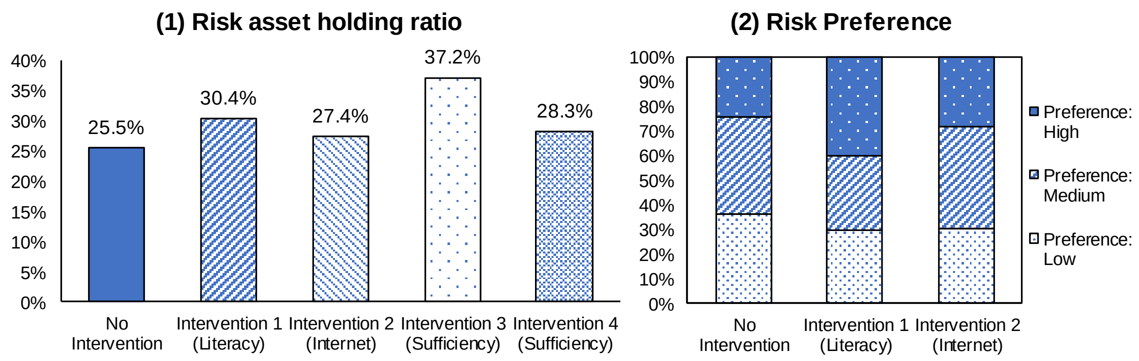
Intervention 1: Financial Literacy				Intervention 2: Internet				Intervention 3&4: Sufficient funds			
	High	Medium	Low		2 +	1	0		High	Medium	Low
Current	12%	37%	51%	Current	48%	31%	22%	Current	21%	30%	49%
Intervention 1	100%	0%	0%	Intervention 2	100%	0%	0%	Intervention 3	100%	0%	0%
								Intervention 4	21%	79%	0%

Source: Made by MHRI based upon *Japan Household Panel Survey* by Keio University

It should be noted that the effects of the interventions are calculated by using conditional probabilities based on observational data, and the variables used at each node include proxy variables. For example, financial literacy includes knowledge such as interest rate and risk hedge, but this paper only uses the knowledge of the government bond price as a proxy for financial literacy. Hence, the proxy variable may not capture the effect of financial literacy fully, and the simulated result could be under-valued. Therefore, it may seem more appropriate to focus on which interventions are likely to be more effective by comparing the margin of increase relative to each other, rather than to pay attention to the absolute value of the increase in the risk asset holding ratio.

The results of the simulations are plotted in **Figure 7 (1)**. As the effects from interventions 1 and 2 are obtained through changing risk preference, **Figure 7 (2)** plots the distribution of risk preference after the interventions. The outcomes suggest that intervention 3 has the highest effect, with an increase in a risk asset holding ratio of more than 10%pt. However, as dispelling the financial concerns after retirement needs to be achieved in intervention 3, it may be difficult for the government to realize this intervention in terms of feasibility. The second most-effective intervention would be to enhance financial literacy, for which about a 5%pt increase is expected in the risk asset holding ratio. The fact that only 12% of households answered correctly to the financial knowledge question suggests that there will be plenty of room for the government and financial institutions to promote and facilitate financial education. The third most-effective intervention is to eliminate the segments that think they may not have enough money to survive after retirement. By changing this confidence, the risk asset holding ratio is expected to increase by about 3%pt. The possible intervention approach may be to convince the overly concerned households that they are not in a severe shortage of money after retirement by showing the life plan simulations. Further, despite the least effective, increasing the number of devices connecting to the internet can improve the risk asset holding ratio by about 2%pt.

Figure 7: Intervention effects

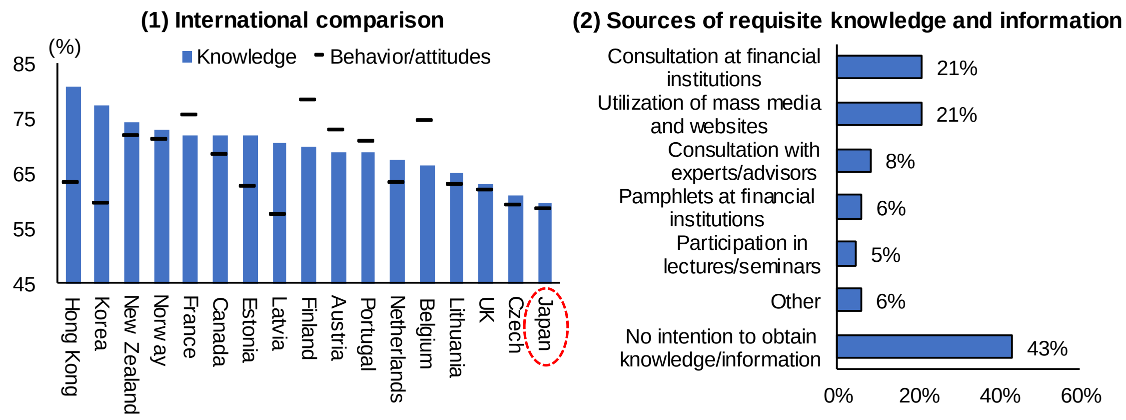


Source: Made by MHRI based upon *Japan Household Panel Survey* by Keio University

In Japan, it has been regarded as a problem that not enough cash & deposits are shifting to the investment. The main reason for this problem may be that many people believe they do not have enough money for retirement (or are determined not to invest until they have sufficient funds for retirement). As the above simulations suggest, if people perceive that they have adequate assets for livings after retirement, the holding ratio increases significantly more than other interventions. However, Japan's potential growth rate is likely to decline in the medium-to-long term due to a declining birthrate and an aging population. To ease financial concerns about livings after retirement against such a background will require fundamental policy measures such as pension reform and dramatic productivity improvements, which will be highly challenging to be achieved practically.

Therefore, although it may be the second-best option from the perspective of effectiveness, considering feasibility, changing households' risk preference by improving their financial literacy would be the effectual action for raising the risk asset holding ratio. According to the *Financial Literacy Survey* in 2019 by the Central Council for Financial Services Information, as compared to the OECD survey, the percentage of correct answers given to questions on financial knowledge[16] in Japan ranked 22nd of the 30 countries for which data are available and is the lowest of 17 countries if the data are limited to advanced economies (**Figure 8 (1)**). The above survey also provided an international comparison of the percentage of respondents that selected desirable financial behaviors and attitudes[17], and Japan ranked 16th out of 17 countries. International comparison implies that the low level of financial literacy in Japan is an issue to be addressed.

Figure 8: Financial literacy



Note: (1) Showing the result of advanced economies only; (2) Multiple answers allowed.
Source: Made by MHRI based upon *Financial Literacy Survey (2019)* by the Central Council for Financial Services Information and *Japan Household Panel Survey* by Keio University

The *Japan Household Panel Survey* in 2017 researched how households obtain the requisite knowledge and information for investment in risk assets (**Figure 8 (2)**), and many respondents mentioned the consultation at financial institutions as well as utilization of the internet. Hence, it is considered that financial institutions' proactive efforts for literacy improvement are essential. However, the panel survey also confirms that many households are reluctant to improve their financial literacy, as about 40% of respondents answered, "No particular intention to obtain either knowledge or information". Because formulating a more rational risk preference is a desirable change for both households and society, reaching such reluctant households will be an imperative challenge for literacy improvement.

[16] Average of five items, including interest rates and risk.
[17] Average of six items, including the long-term financial goal and prioritizing saving.

(4) Robustness

Finally, let us check the robustness of the network structure learned by the PC algorithm. To be specific, this section compares the result of the PC algorithm with other algorithms for learning network structures (structure learning). However, as mentioned above, it is not appropriate to interpret lightly that the learned structure is a causal structure, and some learning algorithms do not confirm the directions of the edges with certainty. Hence, this section examines only the robustness of the skeleton (a graph with undirected edges).

The paper employs six algorithms, three from the Constraint-based algorithm, which is in the same category as the PC algorithm, two from the Score-based algorithm, and one from the Hybrid algorithm.[18] Based on the estimated six skeletons, this paper views that an edge is robust if the edge is observed in four or more skeletons. See appendix 3 for the detailed settings and results of this exercise.

The robustness check indicates that three relationships, "Preference ○–○ Internet", "Literacy 2 ○–○ Security 1", and "Sufficiency ○–○ Security 1", are not observed, although they were by PC algorithm. This result suggests that the implication between “Preference” and "Internet" may need to be modified. Although the PC algorithm argued an "Internet → Preference" relationship, this relationship is only confirmed in two of the six learning algorithms. It does not mean that the relationship has no robustness at all, but since this is the relationship with the smallest intervention effect, as suggested in **Figure 7**, it may be more appropriate to evaluate the relationship with some skepticism. On the other hand, three new relationships are detected, which are "Sufficiency ○–○ Household income", "Age ○–○ Sufficiency", and "Internet ○–○ Income", but it seems that one cannot clearly deny the presence of these edges.

Although the directions of edges are not in the scope of this analysis, about 70% of edges learned by the PC algorithm are found to be robust. These include, for example, the relationships of "Asset-holding" and "Deposits, Sufficiency, and Preference", which are the main results of the analysis performed here, as well as the relationship of "Deposits" and "Household income, Debt, and Age" and the relationship of "Preference" and "Literacy 2". However, the remaining 30% of edges, such as the relationship between "Internet" and "Preference", are not necessarily robust based on the analysis results.

[18] Specifically, Constraint-based algorithm: Hilton-PC, Grow-Shrink (GS), Incremental Association (IAMB); Score-based algorithm: Hill-Climbing (HC), Tabu Search (Tabu); and Hybrid algorithm: Restricted Maximization (RMAX2).

6. Conclusion

This paper attempted to estimate a causal structure relating to the risk assets holding of households in Japan. In order to find out the crucial factors for risk assets holding, the analysis used various variables and adopted machine learning techniques to reduce prior constraints and learn a causal structure from observational data.

The analysis results suggested that two channels influence the holding of risk assets: the household's affordability from safe assets (deposits) and the risk preference. The former channel is to hold risk assets when there is (or, when the household thinks that there is) a surplus in safe funds for livings after retirement. Low income and debt (including a mortgage) are factors that prevent households from making enough deposits, and thus they indirectly discourage households from holding risk assets. Further, income, debt, and deposits are also influenced by age, and this seems to be the background of a positive correlation between the age group (life cycle) and the risk assets holding ratio. Regarding the latter channel, the holding ratio increases when the risk appetite is higher. The result implied that higher financial literacy and internet usage could influence the formation of higher risk preferences. However, it was also suggested that the relationship between the use of the internet and risk appetite may not be robust.

According to the simulations, the most-effective way to increase the risk asset holding ratio for a virtuous cycle of funds is to alter households' perception of insufficient funds after retirement, possibly by implementing government policies that lead to the higher expected growth rates and the credible pension system. Such households' financial anxieties after retirement are likely to be the main reason why Japanese households are not investing in risk assets. Policy efforts are indeed required in this regard, but it will be difficult to eliminate financial concerns for livings after retirement against the background of a declining population and aging economy in Japan. Therefore, although this may be the second-best option, it is vital to increase preference for risk assets by helping households enhance their financial literacy. The international comparison demonstrated that Japan's financial literacy level is low, and it has much room for improvement. The government and financial institutions need to promote and facilitate financial education actively. It is imperative to help households understand their optimal risk tolerance through accurate financial knowledge. Because many households are reluctant and do not feel a necessity for acquiring financial knowledge, approaching these households is also a significant challenge.

Reference

Bank of Japan (2017), "New Facts on Corporates' Inflation Expectation: Part II -Machine Learning Approach-", *Bank of Japan working paper series*, No.17-J-4 (in Japanese)

Bank of Japan (2020), "Flow of Funds -Overview of Japan, the United States, and the Euro area-", August 21, 2020

Barsky, R. B., Juster, F. T., Kimball, M. S., and Shapiro, M. D. (1997), "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study", *Quarterly Journal of Economics*, 112(2), pp. 537-579.

Financial Services Agency (2017), "About NISA" (in Japanese) https://www.fsa.go.jp/policy/nisa/20170614-2/12.pdf (accessed on 25 September 2020)

Fukuhara, T. (2016), "Summary of Issues regarding Risk Asset Holdings of Japanese and the US Households", *BOJ Report & Research Papers* (in Japanese)

Hartemink, A. J. (2001), "Principled Computational Methods for the Validation Discovery of Genetic Regulatory Networks", Doctoral Dissertation, Massachusetts Institute of Technology

Hünermund, P., and Bareinboim, E. (2019), "Causal Inference and Data-Fusion in Econometrics", *arXiv: Econometrics*

Ito, Y., Takizuka, Y., and Fujiwara, S. (2017), "Household Asset Selection Behavior: Verification of Asset Selection Based on Dynamic Panel Analysis", *Bank of Japan Working paper series*, No.17-J-2 (in Japanese)

Iwaisako, T. (2012), *Financial Behavior of Households and Companies and the Japanese economy*, Nihon Keizai Shimbun, Inc. (in Japanese)

Iwaisako, T., Ono, A., Saito, A., and Tokuda, H. (2019), "Disentangling the Effect of Home Ownership on Household Stock-Holdings: Evidence from Japanese Microdata", *RIETI Discussion Paper Series*, 19-E-007

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, New York: Springer

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M., and Bühlmann, P. (2012), "Causal Inference Using Graphical Models with the R Package pcalg", *Journal of Statistical Software*, 47(11), 1 – 26

Kinari, Y. and Tsutsui, Y. (2009), "Determinants of Share of Risky Assets in Japan", *Japan Society of Monetary Economics*, 29 (in Japanese)

Kitamura, Y. and Uchino, T. (2011), "The Effect of Academic Background on Household Portfolio Selection: Evidence from Japanese Repeated Cross Section Data", *Japan Society of Monetary Economics*, 33 (in Japanese)

Merton, R. C. (1969), "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case", *Review of Economics and Statistics*, 51(3), pp. 247-257.

Nakajyo, M., Shino, J., and Imakubo, K. (2017), "Behavioral Characteristics of Japanese Households' Financial Asset Selection", *Bank of Japan Review*, 2017-J-7 (in Japanese)

Scutari, M. (2020), "Package ‘bnlearn’", Reference Manual https://www.bnlearn.com/documentation/bnlearn-manual.pdf (accessed on 18 September 2020)

Scutari, M., and Denis, J. B. (2014), *Bayesian Networks with Examples in R*, Chapman and Hall/CRC

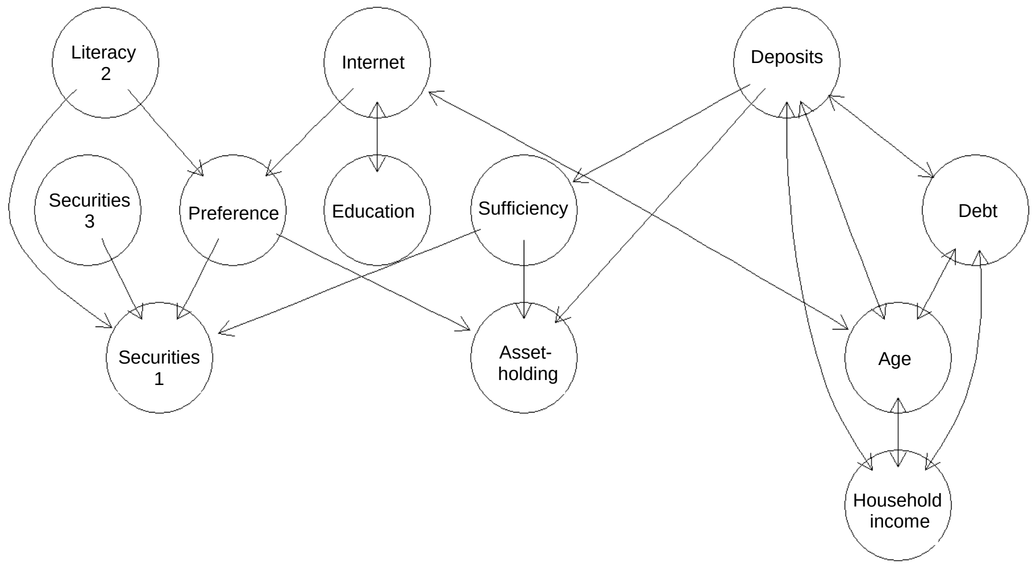
Shioji, E., Hirakata, N., and Fujiki, H. (2013), "Aging and Household Stockholdings: Evidence from Japanese Household Survey Data", *Monetary and Economic Studies*, 32-2 (in Japanese)

Stekhoven, D. J., and Bühlmann, P. (2012), "MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data", *Bioinformatics*, 28(1), 112-118.

van Rooij, M., Lusardi, A., and Alessie, R. (2011), "Financial Literacy and Stock Market Participation", *Journal of Financial Economics, Elsevier*, vol. 101(2), pages 449-472

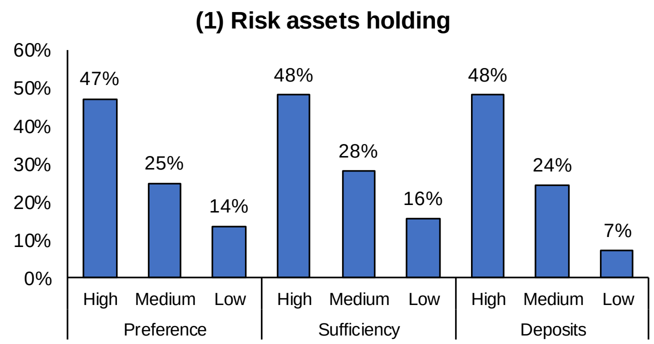
Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Ji Zhu, J., and Higgins, P. D. (2013), "Comparison of Imputation Methods for Missing Laboratory Data in Medicine", *BMJ open*, 3(8).

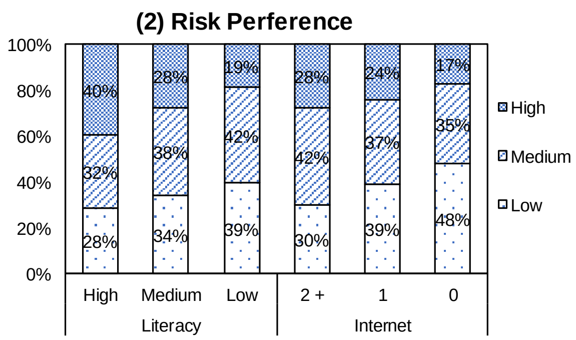
Appendix 1: Bayesian Network (PC algorithm)

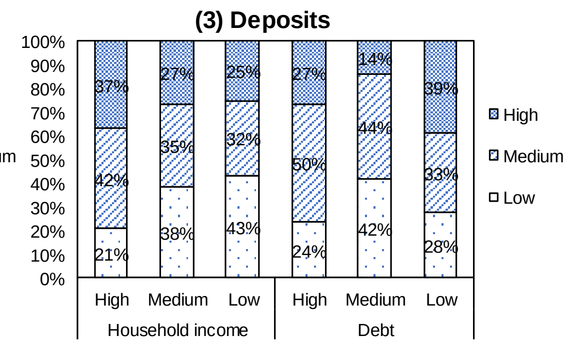


Source:Made by MHRI based upon *Japan Household Panel Survey* by Keio University

Appendix 2: Conditional Probabilities (posterior distribution)







Source: Made by MHRI based upon *Japan Household Panel Survey* by Keio University

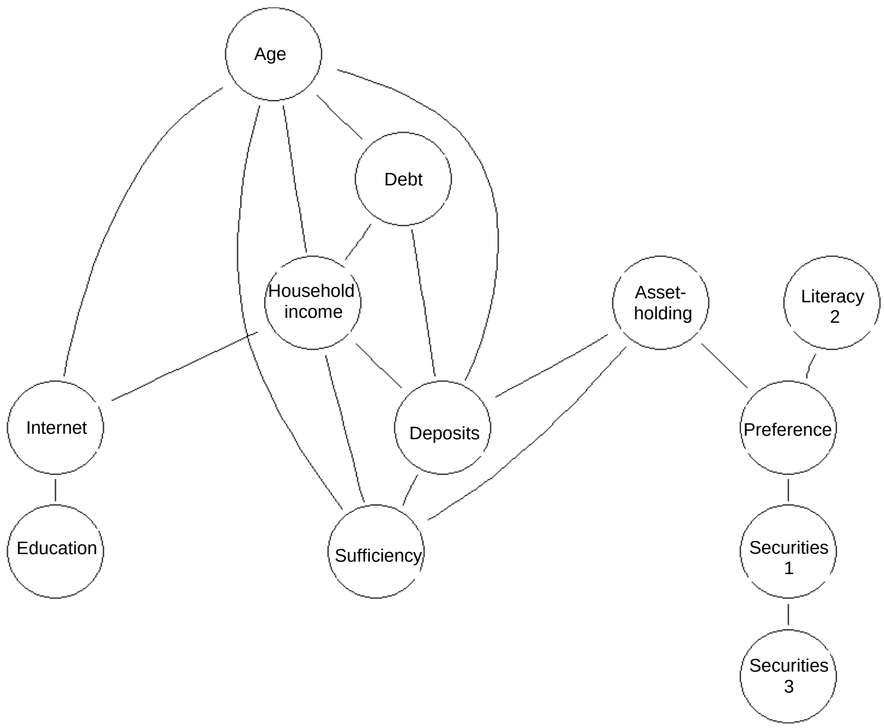
Appendix 3: Robustness

The table below summarizes the settings of six structure learning algorithm used for checking the robustness. In this paper, an edge is regarded as robust if the edge is observed at four or more models. The graph (skeleton) below shows the averaged graph, where only robust edges are plotted.

algorithm	type	test or score	# of edges	Comparison with PC algorithm			notes
				TP	FP	FN	
Hiltion Parents and Children	Constraint-based	Test: Mutual Information (Monte Carlo Permutation)	17	14	4	3	alfa=0.05, B=5000
Grow=Shrink	Constraint-based	Test: Mutual Information (Monte Carlo Permutation)	22	16	2	6	alfa=0.05, B=5000
Incremental Association	Constraint-based	Test: Mutual Information (Monte Carlo Permutation)	21	15	3	6	alfa=0.05, B=5000
Hill-Climbing	Score-based	Score: Bayesian Dirichlet (BDe)	27	17	1	10	iss=200, piror="marginal", restart=1500
Tabu Search	Score-based	Score: Bayesian Dirichlet (BDe)	27	18	0	9	iss=200, piror="marginal", tabu=500
Restricted Maximization	Hybrid	Test: Mutual Information Score: BIC	14	14	4	0	Hilton-PC+Tabu Search
average	—	—	18	15	3	3	threshold=0.65

Note: alfa: type 1 error threshold, B: number of permutations for Monte Carlo tests, iss: imaginary sample size, "marginal": marginal uniform priors, restart: the number of random restarts, tabu: the length of the tabu list

Source: Made by MHRI based upon Scutari (2020) and *Japan Household Panel Survey* by Keio University



Source: Made by MHRI based upon *Japan Household Panel Survey* by Keio University