A Dataset and Structure

A.1 Source Datasets {asec:srcdatas

For each household, we potentially have information from month 1 to month 160. Many households, however, are not observed for all months. For 472 households, there was some monthly specific household information for all 160 months. For 630 households, there was some monthly specific household information for at least 133 months. There were 785 unique household IDs in the survey for households that had at least information in one of the 160 survey months.

For the 785 household IDs, there are 9689 unique household–years, or 8993 unique household–years if we do not count 1998 in which there were 4 months of data. In 95.24 percent (8565) of the 8993 household–years after 1998, there was some household financial information in all 12 months of a calendar year. In 120 of the 8993 household–years, there was less than 6 months of information in a household–year. Including 1998 as well, there are 108742 unique household months in which we have some information about a household.[1]

Dataset A: Skeleton Given $N = 785$ households, and $M = 160$ number of months, we create dataset **A** with $N \cdot M$ rows and 3 variables: a household ID variable, *hhid*, a survey month variable, *month*, and a key that joins household ID with survey month, *hhid_month*. Not all rows will have matching observed data given that we do not observe information about all households in all 160 months, but this skeleton frame, which contains no data, will be used as basis for data merging.

Dataset B: Household Demo Edu From the raw dataset, we constructed dataset **B** which contains several variables from various modules that include information on the number of individuals present in a household in a given month, the highest grades completed for a member of the household in a given month, and several other variables. Dataset **B** is household and survey month specific. The dataset has *hhid*, *month* and *hhid_month*.

Dataset C: Household Financial From White/Rob financial modules, we have dataset **C**, which is also household and months specific, and contains dozens of

1. See `data_log\DataY_CountingMthPerYear.html`

assets and income variables. The dataset has *hhid*, *month* and *hhid_month*.

Dataset D: Borrowing and Saving From survey modules on borrowing, savings, lending and transfers, we construct dataset **D**, which is household and financial transaction specific. The dataset has *hhid*, *month* and *hhid_month*. Here, *month* indicates the month in which a loan was taken out, or when a deposit or withdraw was made. If household i in month m acquired 2 loans, there would be two rows in dataset **D** containing information about these two loans.[2] If the household took out deposits from two deposit taking institution in that month, there would be two additional rows for that household for that month that records how much was taken out and from which account. If a household made two deposits in two different depositing institutions in that month, there would be two more rows for that household in that month that records how much was deposited, at what interest rate, into which deposit taking institution. For this paper, we only use information from the borrowing and savings modules, but this dataset also contains additional information on lending as well as transfers.

A.2 Combined Monthly Data

{asec:combined

Dataset Y: Merge A B C D Datasets A, B, C and D all have *hhid_month*, we merge them based on this key to create dataset **Y**.[3] There will be at least one row for each household in any month when there is any household demographic and financial information for that month from datasets **B** and **C**. In some months, there might be no financial transactions at all for some households, if that is the case, then the variables corresponding to financial information for these household months would contain no values. During months in which a household made more than 1 financial transaction, there would be multiple rows that contain different loan and deposit information from dataset **D** for each loan taken out, deposit or withdraw made; these rows for the same household and month would contain identical household demographic/edu and financial information from datasets **B** and **C**. In each household month, there could be two rows for household i in month m that correspond to lender l if the household obtained two loans from lender l.

2. There are information from multiple months about each loan as households pay back loan principle and interests, but we aggregate these information up and have one observation for each loan, at the month when the loan was taken out.

3. `\data_month160\thaiMthly_DataY.dta`

Dataset Y Keys String Province, Village and Household IDs: `id_p`, `id_v`, `id_h`. And numeric versions of these IDs (for sorting purposes): `id_p_num`, `id_v_num`, `id_h_num`. Variables for year, 1 to 160 survey month, and 1 to 12 calendar month: `ts_yrr`, `ts_mth`, `id_m12`. Indicators for first occurances of panel units (generated from `egen=tag(id_x,ts_y)`): tag of province year `S_py_tag`, tag of household month `S_hm_tag`, tag of househodl year `S_hy_tag`, and tag of household (all years) `S_hl_tag`

B Dataset Z: Annualized Model Variables {asec:keyvari

Given the model, in our annualized dataset we have several key variables that map to the state-space and choice-space of the household maximization problem. The annualized dataset has one observation for each household year that the household was in the survey. This is dataset **Z**.[4] The total number of observations is 9689, which are in terms of household–years, the number of variables is 142. Out of these variables, 9 variables from dataset **A** originally, 6 variables from dataset **B** originally, 91 variables from dataset **C** originally, 14 variables for saving from dataset **D** originally, and 20 variables for borrowing from dataset **D** originally, and several other variables.

Annualize Variables from Dataset A: Skeleton We inherit 9 key variables from here: `id_p`, `id_p_num`, `id_v`, `id_v_num`, `id_h`, `id_h_num`, `ts_yrr`, `id_h_ts_yrr`, `S_py_tag`, `S_hy_tag`, `S_hl_tag`.

Annualize Variables from Dataset B: Household Demo Edu From household demo education dataset, we keep 6 main variables: the educational attainment of the head of the household over all years (max over all years), `EDh_gradeAtt_head`, the highest educational attainment achieved by a member of the household (max over all years), `EDh_gradeAtt_max`, the number of individuals who was at home for some periods within the calender year, `DMh_popAtHome_yr`, and also variables `DMh_mthInSvy_cntMth`, `DMh_mthInSvy_begMth`, `DMh_mthInSvy_endMth` for total number of months in survey, start mont and end months.

4. `\data_month160\thaiMthly_Annulized_DataZ.dta`

Annualize Variables from Dataset C: Household Financial Households also enter a period y with K_{iy}, the productive physical capital held at the beginning of the period. Inside the period, given realized productivity shocks, output is realized from the production function and we have Y_{iy}.

From White/Rob financial modules, we have 43 `agg_IS_*` variables, which measure monthly income/costs etc–potential Y_{iy} variables. We sum up for each of these 43 variables across all months in a calendar year.[5] These 43 variables start with the prefix `hy_agg_IS_*`.

The dataset also contains 24 `agg_BS_*` variables, which measure monthly capital and asset levels–potential K_{iy} variables. We generate 24 annualized averages of these variables[6]: named as `hy_agg_IS_*_avg`. We generate 24 additional annualized versions of these variables which equaling to the first month of the year values of these variables[7]–these variables are named with this structure `hy_agg_IS_*_1st`.

Annualize Variables from Dataset D: Saving and Borrowing Within each period y[8], households i could save, and could also make borrowing choices over three borrowing options, with the BAAC and VF options coming from a menu of loans. We describe the construction of the saving variables in Section C.1, and the borrowing variables in Section C.2. For savings, the amount saved is S_{iy}.

For borrowing, households can borrow from informal sources, $P_{iy,inf}$, they can borrow from formal traditional sources (BAAC+), $P_{iy,baac+}$, and they can also borrow from village fund, $P_{iy,vf}$–with the letter P for principle, and $P \leq 0$, more negative is more borrowing. Savings amounts have associated savings interest rate R^S_{iy}, and the three sources have borrowing interest rates, $R_{iy,inf}$, $R_{iy,baac+}$, and $R_{iy,vf}$. In our model, the informal borrowing rate is individual specific, and the rates on BAAC+ and VF loans and savings rate are not i specific. The average location and time specific savings rate and borrowing rates

5. Sum only once per month, sum conditional on if `S_hm_tag==1` given that there are potentially multiple transactions per month. Additionally, for households where we observe less than 12 months in the year, using average for months in which household is observed multiplied by 12.

6. Need to use one number from each month, take average conditional on if `S_hm_tag==1` given that there are potentially multiple transactions per month.

7. January if January if the first month that the household appears in the survey for that year, other months if otherwise.

8. There is a few household IDs that only exist for initial month, but the household seems to not have been surveyed after perhaps a baseline survey. See `data_log\DataY_mth0_issue.html`

from BAAC+ and VF should be based on averaging these individual specific rates.

As described in the model section, given the borrowings and savings choices made in the previous period, we have $B_{i,y+1}$ for period $y+1$:

$$B_{i,y+1} = (P_{iy,inf} \cdot R_{iy,inf} + P_{iy,baac+} \cdot R_{iy,baac+} + P_{iy,vf} \cdot R_{iy,vf}) + S_{iy} \cdot R^S_{iy} \quad (1) \quad \texttt{\{eq:bnext\}}$$

(Note that we use the individual specific rates for this calculation).

C Borrow and Savings Variables Construction

Main Savings In the dataset **Z** (see section A.2), where each row is a household year, we have 1 variable for savings amount, and one for savings interest rate:

$$\left\{ S_{iy}, R^S_{iy} \right\} \quad (2) \quad \texttt{\{eq:savemain\}}$$

And we have 14 additional savings amount and savings rate variables for different deposit institutions which we keep for the sake of completeness

Main Borrowing In dataset **Y** (see section A.2), where each row is a loan, we have 8 main variables for loan information. In dataset **Y**, ewhere each row is a household year, we have 6 main interest rate and principle variables which we use for the 3 borrowing categories in our model. And we have 14 additional interest rate and principle variables which we keep for the sake of completeness (and checking and generating summary statistics if needed)

See Section C.2.1 for descriptions of the 8 individual loan j variables. The most important variable variables are, R_j and P_j. And see Section C.2.2 for how we calculate the annualized rates for the household and year observation based dataset. We have 6 main variables that are household i, year y, and lender category type l, specific variables:

$$\left\{ \left\{ P_{iyl}, R_{iyl} \right\}_{l \in (\text{inf, baac+, vf})} \right\}_{iy} \quad (3) \quad \texttt{\{eq:borrmain\}}$$

We have 14 additional household i, year y, and lender type $\hat{l}$ specific variables, these are not used in the model, but we keep in dataset for completeness.

C.1 Constructing Savings Variables {asec:savevars

C.1.1 Net Annual Deposits and Interest Rates

Household reports for each deposit institution whether in each month money was deposited or taken out of an account. We calculate net savings in a year across savings institutions, taking the difference between deposits and withdraws. Specifically, we use two pieces of information to calculate savings rate and savings amount: 1, net changing in savings at each depositing institution in each month; 2, interest rate on savings.

For individual i in month m of year y with deposit institution d, we observe withdraw and deposits, the net of withdraw and deposit in each month is, s^{Δ}_{imyd}. Unlike borrowing information, for savings, which happens at the monthly level for many accounts, there is often no interest rate recorded with each deposit made. To obtain the savings rate, we take the average savings rate, whenever an annual savings rate is observed when new deposits are made, for all individual i in province p who made deposits at institution d in year y. This rate is: R^{S}_{pyd}.[9]

C.1.2 Stock of Savings by Deposit Institution

Let S_{iyd} denote savings by household i in year y at depositing institution d, it is:

$$s^{\Delta}_{iyd} = \sum_{m=1}^{12} \left[s^{\Delta}_{imyd} \right] \tag{4} \quad \{eq:s1sdiyd\}$$

$$S_{iyd} = s^{\Delta}_{iyd} + \left[S_{i,y-1,d} \cdot R^{S}_{p,y-1,d} \right] \tag{5} \quad \{eq:s2siyd\}$$

This means that in each year y, we have a vector of deposits S_{iyd}[10] at different deposit taking institutions for each household[11], and have a vector of associated

9. For the common deposit institutions, we have interest rate information for each province in each year. For institution d that does not have interest rate information in year y and province p, if there are deposits made in dyp, we will use average rate from provinces where rates are observed $R^{S}_{pyd} = R^{S}_{yd}$

10. Variables: `D_net_BankAAC_yr,D_net_PCGroup_yr,D_net_AgrCoop_yrD_net_GovSave_yr,D_net_Commerc_yr,D_net_Otherrr_yr`, see file `GEN_SAVE_eq_s1sdiyd_s_delta_iyd.log` in folder `..\data_log`, generated by invoking do file: `..\code\stata\data_process\proc_data_save_bybank_agg.do`

11. Suppose a household starts the year 2000 with 10 Baht in savings. If the household deposits 1 Baht in January of year 2000 into one savings account, undertakes no other savings actions for the rest of the year, the savings choice for year 2000 would be 11 Baht. Suppose a households starts the year 2000 with 10 Baht in savings. If the household withdraws 10 Baht in March of year 2000 from savings account, undertakes no other savings actions for the rest of the year, then the savings choice for year 1999 was 10 Baht and the savings choice for year 2000 would be 0 Baht.

savings rate R^S_{pyd}.[12]

We can directly calculate from raw data s^{Δ}_{imyd} as well as R^S_{pyd}. Once we have these, we can drop individual deposit and withdraw information, and make calculations at the household year level. Specifically, we use the following procedure to obtain the sequence of $\{S_{iyd}\}^{2011}_{y=1999}$. We first set $\widehat{S_{i,1998,d}} = S^{init}_{i,1998,d}$, the initial reported deposits in an account, then we can follow Equation 5 and, year by year, add $(\widehat{S_{i,y-1,d}} \cdot R^S_{p,y-1,d})$ to s^{Δ}_{imyd} starting from $y = 1999$. The sequence $\{\widehat{S_{iyd}}\}^{2011}_{y=1999}$ is likely to contain negative numbers if household initial deposit was greater than $S^{init}_{i,1998,d}$ and the household has taken out some of that unreported initial deposit over the years. We replace initial deposit by, $S_{i,1998,d} = S^{init}_{i,1998,d} + \widetilde{S^{init}_{i,1998,d}}$, where $\widetilde{S^{init}_{i,1998,d}}$ is:

$$\widetilde{S^{init}_{i,1998,d}} = (-1) \cdot \min\{0, \frac{\widehat{S_{i,1999,d}}}{R^S_{pyd}}, ..., \frac{\widehat{S_{i,2011,d}}}{\prod^{2011}_{y=1999} R^S_{pyd}}\} \tag{6}$$ {eq:s2}

In Equation 6, we adjust for interest rates. Now, we can compute sequence $\{S_{iyd}\}^{2011}_{y=1999}$ based on $S_{i,1998,d}$. The idea here is that if for whatever reason the initially reported deposits in an account is less than actual (for some accounts the initial amount could be 0), we can still try to estimate the initial amount by looking at how much is taken out from the account in subsequent years. This will provide us with the correct sequence of S_{iyd} if the household has withdrawn all deposits from an account at some point of the survey, which means that this calculation is more accurate if the survey has been in place longer.

C.1.3 Annual Household Stock of Savings

In our model, there is one aggregate savings choice for each calendar year, which aggregates over different depositing institutions d, it is:

$$S_{iy} = \sum_d S_{iyd} \tag{7}$$

along with savings rate:

$$R^S_{iy} = \sum_d R^S_{vyd} \frac{S_{iyd}}{S_{iy}} \tag{8}$$

Along with the individual annual aggregate savings amount and interest

12. Use deposit institution and year specific average rate R^S_{yd} when R^S_{pyd} does not exist for a particular province but there are deposits from that year and depositor.

rate, we also keep the the vector of deposits at individual institution along with savings rate:

$$\left\{ \left\{ S_{iyd}, R^S_{vyd} \right\}_{d \in (1,..,5)} \right\}_{iy} \tag{9}$$

where $d = 1$ are deposits at BAAC, $d = 2$ are deposits at PCG (Production Cooperative Groups), $d = 3$ are deposits at Government Savings Bank, $d = 4$ are deposits at commercial banks, and $d = 5$ are deposits at category 15 deposit institutions.[13]

C.2 Constructing Borrowing Variables

{asec:borrvars

C.2.1 Interest Rate on Individual Loans

{sec:loanrate}

See File: `\data_log\DataD_Borr_initVars_PRTMGR`

Each loan is taken out starting at a particular month, with a specific loan size, an interest rate, and loan length. Each month after a loan is taken out, we can see how much repayment is made for interest and for principle. For each loan, we have potentially both the initially stated interest rate and the actual interest rate repayment.

Given our goal of annualizing loans for each household, we compute individual loan interest rates in two ways. We calculate a reported rate and also an actual repayment rate. The goal is to find out, if 1 baht was borrowed, how much should be paid back. In the simplest case, we observe a household that borrowed 100 Baht from a lender at an annual rate of 6 percent. The reported rate was 6 percent. Then we observe that after one year, indeed 106 Baht was paid back, which means the actual repayment rate is also 6 percent.

We first calculate reported interest rates on different loans. Reported rate comes from two sets of variables. One set of variable explicitly asks households if the loan they have taken on has interest rate or not and what type of rate it is (annual, monthly etc), and another asks how much have to be paid back in Baht including repayment in kinds.

Some loans have reported rate, some do not. For those that do not, we set the interest rate to 1.00, which means households still have to pay back the loans, but they do not have to pay interest rates.[14] For these loans with 1.00 reported loan rate, we look at how much has been paid back during the span of the loan.

13. Do not know what category 15 is.
14. We do not consider transfers here

If more than 100 percent interest has been paid back, then we use the actual repayment interest rate as the loan interest rate.

Loan j Level Variables We observe these variables for each loan j, from househld i, taken out in month m:

$$\underbrace{P_j}_{\text{Principle}}, \overbrace{R_j^{R1}, T_j, M_j, G_j}^{\text{Initial Rate Info}}, \underbrace{R_j^A}_{\text{Repay Rate}} \tag{10}$$ {eq:binitvars}

where P_j is the principle amount of the loan, R_j^{R1} is the initially recorded interest rate for the loan. R_j^{R1} could be annual or monthly interest rate: T_j is a categorical variable for the type of loan $T_j \in (\text{Annual}, \text{Monthly}, \text{Other})$, M_j is the number of month after which the loan is due. G_j is the total amount to be repaid in both money and kind, it is initially asked of households when loans are acquired and could be higher then the repayment implied by (R_j^{R1}, T_j, M_j) if a household did not consider repayment in kinds when reporting the rate earlier. R_j^A is the interest rate that the household has paid by the end of the loan's existance in the survey or by the final survey month. Some loans are not due yet, which means this rate would be lower than initially reported rate. In cases where the repayment rate is lower than initially reported rate, this could represent partial non-repayment.

Finding loan rate R_j, our best estimate of loan rate The loan specific interest rate that we will use is R_j, this records the total repayment for the loan divided by the principle of the loan. To arrive at R_j, we follow these equations, and generate two intermediate interest rate variables $\hat{R_j}^{R1}$ and R_j^{R2}. We find $\hat{R_j}^{R1}$ following this:

$$\hat{R_j}^{R1} = \begin{cases} R_j^{R1} \cdot M_j, & \text{if } T_j = \text{Month} \\ R_j^{R1} \cdot \frac{1}{12} \cdot M_j, & \text{if } T_j = \text{Annual} \\ \text{NA}, & \text{otherwise} \end{cases} \tag{11}$$ {eq:bRj1}

We take the max of the two reported interest rate[15]:

$$R_j^R = \max\{\hat{R_j}^{R1}, R_j^{R2} = \frac{G_j}{P_j}\} \tag{12}$$ {eq:bRj2}

Then we take the max betweeen the reported and actual interest rate[16]:

$$R_j = \begin{cases} R_j^R \text{ if } R_j^R \text{ exists} \\ R_j^A \text{ if } R_j^R \text{ is null} \end{cases} \tag{13}$$ {eq:bRj3}

This means we end up with three interest rate variables:

$$\underbrace{R_j}_{\substack{\text{Estimated Rate} \\ \text{max of} \\ \text{R and A}}}, \quad \underbrace{R_j^R}_{\substack{\text{Initial} \\ \text{Reported} \\ \text{Rate}}}, \quad \underbrace{R_j^A}_{\substack{\text{Repay Rate} \\ \text{after} \\ \text{14 months}}} \tag{14}$$ {eq:bRj4}

C.2.2 Annualize Loan Information {sec:annualrat

Individual loan principle and interest rates are P_{iylj} and R_{iylj}: loan j from household i in year y from lender type l. We describe how we arrive at individual loan interest rate, R_{iylj}, the in Section C.2.1.

In the span of a year, a household could have borrowed multiple times from a particular type of lender. We sum up total borrowing from each lender in each year. To arrive at annualized interest rate for each lender, we take the average of interest rates for loans weighted by the loan sizes from each lender type:

$$P_{iyl} = \sum_j P_{iylj} \tag{15}$$ {eq:bpiyl}

$$R_{iyl} = \sum_j R_{iylj} \frac{P_{nylj}}{P_{iyl}} \tag{16}$$ {eq:bRiyl}

We categorize loans into three types of lenders, Village Fund loans, traditional formal loans, and informal loans. This means we for each household i in

15. It is possible that the the reported rate that considers goods and money repayment is slightly higher because repayment in goods was not considered in the other variable.

16. It is possible that some loans due to mis-measurement did not report an interest rate initially, but later repayment measurement better captured repayment.

year y, we have 3 pairs of principle and interest rates:

$$\left\{ \left\{ P_{iyl}, R_{iyl} \right\}_{l \in \left(\text{inf, baac+, vf}\right)} \right\}_{iy} \tag{17}$$

Village Fund (vf) loans come from the village fund, traditional formal loans (baac+) consists mainly of BAAC loans, but also includes loans from agricultural cooperatives.[17] Informal loans (inf) consists of loans from non-institutional sources including friends, neighbors, relatives and local money-lenders.[18] The key feature of the informal loans, as we try to capture in our model, is that they could be borrowed at smaller and continuous amounts, rather than at discretized larger amounts as is the case with BAAC loans.

We also keep a record of principle and interest rate variables for smaller loan categories, $\hat{l}$:

$$\left\{ \left\{ P_{iy\hat{l}}, R_{iy\hat{l}} \right\}_{\hat{l} \in (1,..,8)} \right\}_{iy} \tag{18}$$

where $\hat{l} = 1$[19] are loans from friends neighbors and relatives, $\hat{l} = 2$ are loans from money lenders, $\hat{l} = 3$ are loans from PCG groups, $\hat{l} = 4$ are loans from BAAC, $\hat{l} = 5$ are loans from agricultural cooperatives, $\hat{l} = 6$ are loans from village fund, $\hat{l} = 7$ are loans from category 21[20], and $\hat{l} = 8$ are loans from commerical banks (the smallest category).

Compare loan sizes and rates for actual loans and annualized loans Discuss in this section how many loans from each group of lender a household has in a particular year. If have one loan from each source always per year, then annualization is not an issue. Given that sometimes multiple loans are obtained from the same lender, do the annualized loans capture original characteristics for each loan? In particular, when we take weighted average of interest rate, what was the rate distribution at the loan level, what is the rate distribution with annualized loans.

17. Agricultural cooperative loans are generally similar to BAAC loans in terms of loan sizes and rate–the agricultural cooperative loan menu overlaps with the BAAC loan menu.

18. We include PCG loans within the informal loan category. PCG is normally considered formal, but their loan size and rates are similar to informal loans

19. lenderType 1, 2 and 3

20. There is loan category = 21 which I have been unable to identify, I categorize loan category = 21 that comes from within Tambon as informal loans, and those from outside of the villages as traditional formal loans. Checked with Lek once about category 21, could not figure out what it is, someone probably knows.

PCG and Informal Loans Show similarity between PCG and Informal Loans in rate and size, and why it is OK to combine them in the same category.

Agricultural Cooperative loans and BAAC Show similarity between agricultural cooperative and BAAC loans in rate and size, and why it is OK to combine them in the same category.