REPLICATION EXERCISE 5: KARLAN AND ZINMAN (2009)

VOJTĚCH BARTOŠ

In this exercise we try to get a better understanding and to replicate the paper by Karlan and Zinman (2009) on observing unobservables, namely on disentangling information asymmetries in credit markets. Read the paper.

1. Joint classroom discussion

(1) Why should we be interested in understanding which type of information asymmetries is more important in the real life?
(2) Most of the academic work on information asymmetries has been theoretical. Why is it difficult to identify existence of information asymmetries, let alone to distinguish between different types?
(3) This paper provides an interesting method of documenting information asymmetry and also aims to separate between the two of its main sources. Previewing the results, what type of information asymmetry seems to be the main driver of defaults in this study?
(4) Who is the partner organization and what services do they offer?
(5) To get a better picture about the rates offered in the market of the study, what is the difference between the rates to high and low risk borrowers in South Africa? What are the typical rates offered by the partner organization, absent the experiment?
(6) What are the typical default rates among the approved clients of the partner organization? Separate by first-time and repeated borrowers.
(7) Describe the sample of the experiment. Who are the individuals studied here? How did the experimenters reach them? Do you see any potential problem with the sample used? Be specific.
(8) How does the present experiment allow us to identify the specific types of information asymmetries? Describe the experimental design by answering the questions that follow:
   (a) What is the difference between r^o, r^c, and r^f in the paper?
   (b) It is important that the individuals do not respond to r^c when they are making the decision whether to apply or not? Why? What test do the authors provide to confirm that the individuals indeed do not respond to r_c, but rather only to r^o?
   (c) How many individuals applied for the offered loan? How many of those individuals passed the initial screening (i.e. those who form the sample of main interest)?

(d) How was the r^c presented to the applicants by the loan officers? Why is this important?
(e) Why were the applicants given a chance to readjust the loan size and maturity after r^c was revealed to them? Why could this potentially be a problem? How do the authors convince us that it is actually not an issue?
(f) How is r^f assigned using the experimental design presented here? Do you see potential issues with this? Why do we believe that offering r^f is attractive for the clients?
(g) Now we should describe the identification of the presence of different types of information asymmetries using the experimental design. Use Figure 1 and section 3.3 to describe the identification strategy.

(9) At this point, we can also turn to the formal model presented in section 4 for a better understanding of what the empirical strategy using the experimental design described above can say about the types of asymmetric information. Due to the randomization on the side of the lender (interest rates are randomly assigned), we model only the behavior of a borrower.
(a) First, what phases does the model assume? Describe them briefly.
(b) What solution method (recall our lecture on games) does the borrower use when making his or her decision?
(c) What determines the probability of success of the individual project? What happens to the loan repayment when the project succeeds (see the assumptions).
(d) How does r^c affect the choice of effort? To answer this, differentiate the equation below with respect to e (assume it is a continuous variable for simplicity) and use assumption 2 (hint: it has to do with the curvature of the probability function):

$$\max_{e \in e_l, e_H} \pi(\theta_i, e)[Y(\theta_i) - 1 - r^c + C_b(r^f)] - e - C_b(r^f) \tag{1}$$

(e) How does r_f affect the choice of effort (keeping r^c fixed)? Use the differentiated equation from above, recall that the cost of default has the following property: $\frac{\partial C_b(r^f)}{\partial r^f} < 0$, and use assumptions 1 and 5 to determine the sign of the denominator in the expression for $\frac{\partial \pi(\theta_i, e)}{\partial e}$ derived above. Then use the assumption 2.
(f) Link what we have just learned from the model predictions derived in the previous two steps to the expectations about the effect of r^c and r^f on default rates, keeping r^o fixed.
(g) For the sake of simplicity and time, let's just take the second part of assumption 4 as given (i.e. that borrowers with higher θ_i choose a lower effort level).[1]

[1]But if you want, just take a derivative of the identity in assumption 4 with respect to e, plug this into the derivateve of Equation 1 with respect to e, rearrange and use the fact implied by assumption 4 that optimal effort $\hat{e}$ is a function of r^c, $C_b(r^f)$, and θ_i.

Given this, the individual chooses whether to borrow based on comparing the expected net returns when borrowing with an outside option:

$$\underbrace{\pi(\theta_i, \hat{e})\Big(Y(\theta_i) - 1 - r^o\Big)}_{\substack{\text{project success,} \\ \text{loan repaid}}} - \underbrace{\hat{e}}_{\substack{\text{cost} \\ \text{of effort}}} - \underbrace{(1 - \pi(\theta_i, e))\Big(C_b(r)\Big)}_{\substack{\text{project fails,} \\ \text{cost of default "paid"}}} \geq \underbrace{0}_{\substack{\text{Outside} \\ \text{option}}} \tag{2}$$

(i) Use the assumption 4 to rewrite Equation 2.
(ii) Then take a derivative of the left-hand side of the rewritten equation with respect to θ.
(iii) Then assume that $C_b(r) < 1 + r^o$ and recall assumption 3. You should immediately see from the resulting formula that the derivative is positive. In other words, the expected return from the project is increasing in θ. This also implies that, depending on the actual size of $C_b(r)$, the cost of effort, $\hat{e}$, and of course given r_o, Equation 2 will definitely be positive either for all risk types, or at least for some higher θ_i types above a threshold, the paper defines it as $\underline{\theta_i}$. This potentially creates the separation of the borrower pool based on risk.[2]

(h) Keeping $C_b(r)$ and $\hat{e}$ constant, the paper further shows that the higher the r^o, the riskier the pool of clients becomes. What conclusion about the link between r^o and default rates does this lead to?

(10) What proxies for the cost of default does the paper use? Why is it difficult to estimate the true cost of default to the lender?

(11) What limitation does the paper have in isolating the classic adverse selection effect? See the discussion on potential offsetting effects in the model part discussing "Hidden Information Effect".

(12) This provides us with all we need for the empirical analysis. We will reconstruct the following regression model (see section 5 for details):

$$Y_i = \alpha + \beta_o r^o + \beta_c r^c + \beta_b C + X_i + \varepsilon_o \tag{3}$$

Briefly comment on how to read the coefficients in terms of the model / identification strategy. What signs of coefficients do we expect?

2. Group work

Now we are in good shape to open the data.[3]

(1) Before we replicate Table 1, we should first familiarize ourselves with the data a bit. Open the dataset, browse the data and see the variables manager for what is available.

[2] The paper also examines a situation in which $C_b(r) > 1 + r^o$. In this case, the situation reverses and we can see that it is rather the safer borrowers who end up borrowing. This is rather an empirical question.

[3] You might wonder that there are not many variables. Unfortunately, the authors do not provide us with the data from the survey that was administered after the loan contracts were signed.

(2) How many solicitation letters were sent in total?
(3) What was the average offer rate, r^o, that the individuals were offered? Show this by credit risk classification by the organization.
(4) Can the data tell you how many clients came to the branch and applied for a loan?
(5) How many clients who applied were approved by the branch officials and took the loan?
(6) How many clients out of those approved received a surprise contract rate (i.e. $r^c < r^o$)?
(7) How many clients out of those approved received a dynamic incentive?
(8) We should check whether the randomization was successful.
    (a) Create a dummy variable for whether an individual, conditional on getting a loan, was offered a surprise contract rate (i.e. $r^c < r^o$).
    (b) Create dummy variables for each of the respective credit risk categories as classified by the organization.
    (c) Run series of regressions (clustering at a branch level) with the pre-treatment variables (female, mailer wave, dummies created one step above) on the left hand side and the dummy created two steps above and the *yearlong* variable, and their interaction on the right hand side. Store the results using `outreg2`[4] and comment.
(9) Replicating Table 1
    (a) Reconstruct the respective regressions in Table 1. Cluster at the branch level. Include branch fixed effects (i.e., dummies for all branches, use the `xi:` option at the beginning of the line and then include `i.branchuse` as a variable, it creates the dummies for you), mailer wave fixed effects, as well as the credit risk category fixed effects. The dependent variables are *pstdue_perc_average* in columns (1) and (2), *cdl_pos_average* in columns (3) and (4), *badacct_last* in columns (5) and (6). Store the results using `outreg2`.[5]
    (b) Intentionally, I dropped the standardized index. However, you can create it yourself quite easily: Take the standardized values (i.e. mean 0, SD=1) of all the three dependent variables and create their average.[6] Then run the regression specification used one step above using the newly created variable as a depentent variable. This should give you the results in columns (7) and (8). Store the results.
    (c) Comment on the results.

[4] I am using the `symbol(***, **, *) bdec(3) sdec(3) se replace excel` option.

[5] Here, I am using the `symbol(***, **, *) bdec(3) sdec(3) se keep(offer4 final4 yearlong) replace excel` option.

[6] For this, you will want to use the `egen` function and you will want to read about `std` and `rowmean` options.