Decision evaluation in a selection process

Gianfranco Atzeni

Laboratorio di valutazione delle decisioni di investimento

Introduction

- In real decision making processes manager take decisions selecting their preferred options.
- We want to evaluate these decisions in which manager self-select into some option.
- Self-selection models: appropriate estimation tools to estimate how observables determinants affect the final outcome, taking into account the possible bias in parameters estimation.
- In addition, they represent a way to incorporate and controlling for **unobservable private information** possessed by firms.

Self-selection: the statistical issue

$$y_i = x_i'\beta + u_i \tag{1}$$

where $\varepsilon_i | x_i \sim N[0, \sigma^2]$

In equation 1 the dependent variable is typically an outcome such profitability, return, amount financed etc.

Suppose the outcome is referred to firms that self-select into choice E: for this subsample equation 1 can be written as:

$$y_i | E = x_i'\beta + u_i | E \tag{2}$$

If self-selecting firms are not random subset of the population the usual OLS or GLS don't give consistent estimator of β

Accounting for self-selection

It is in two steps:

- Step 1: specifies a model of self-selection. Using a theory to model why some firms select *E* and other *NE*

This step is important because the underlining assumptions affect what econometric model should be used.

- Step 2: links the random variables that drive self-selection to the outcome variable *y*.

Baseline Heckman selection model

- The regression of equation 1 must be estimated using a sub-sample of firms that select into choice E.
- Selection is specified using a probit model in which firm i chooses E when the net benefit for this choice (W_i) is positive:

$$W_i = Z_i\gamma + \varepsilon_i > 0 \Rightarrow C = E \tag{3}$$

$$W_i = Z_i\gamma + \varepsilon_i \leq 0 \Rightarrow C = NE \tag{4}$$

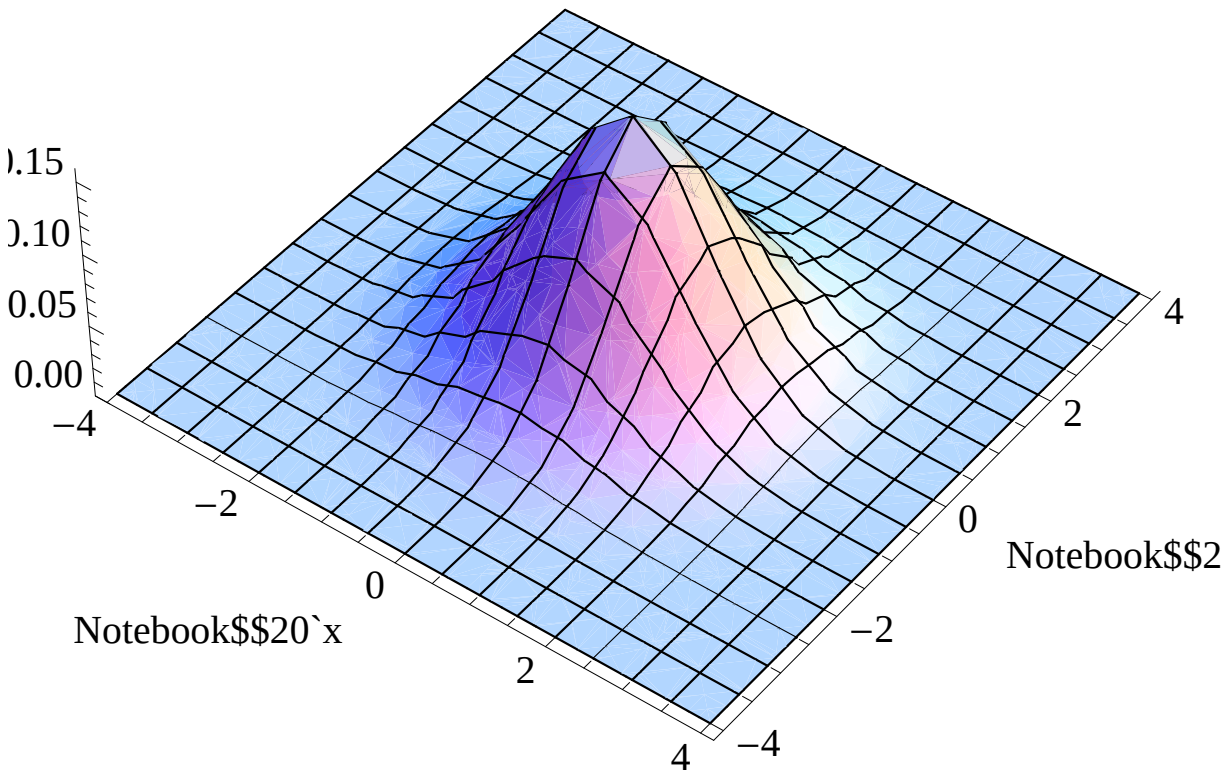
where: Z_i denotes publicy known information influencing a firm choice; γ is a vector of probit coefficient ε_i is the error term orthogonal to public variables Z_i

$$y_i = x_i'\beta + u_i \tag{5}$$

Assuming that u_i and ε_i are bivariate normal it is possible to derive the likelihood function and ML estimator

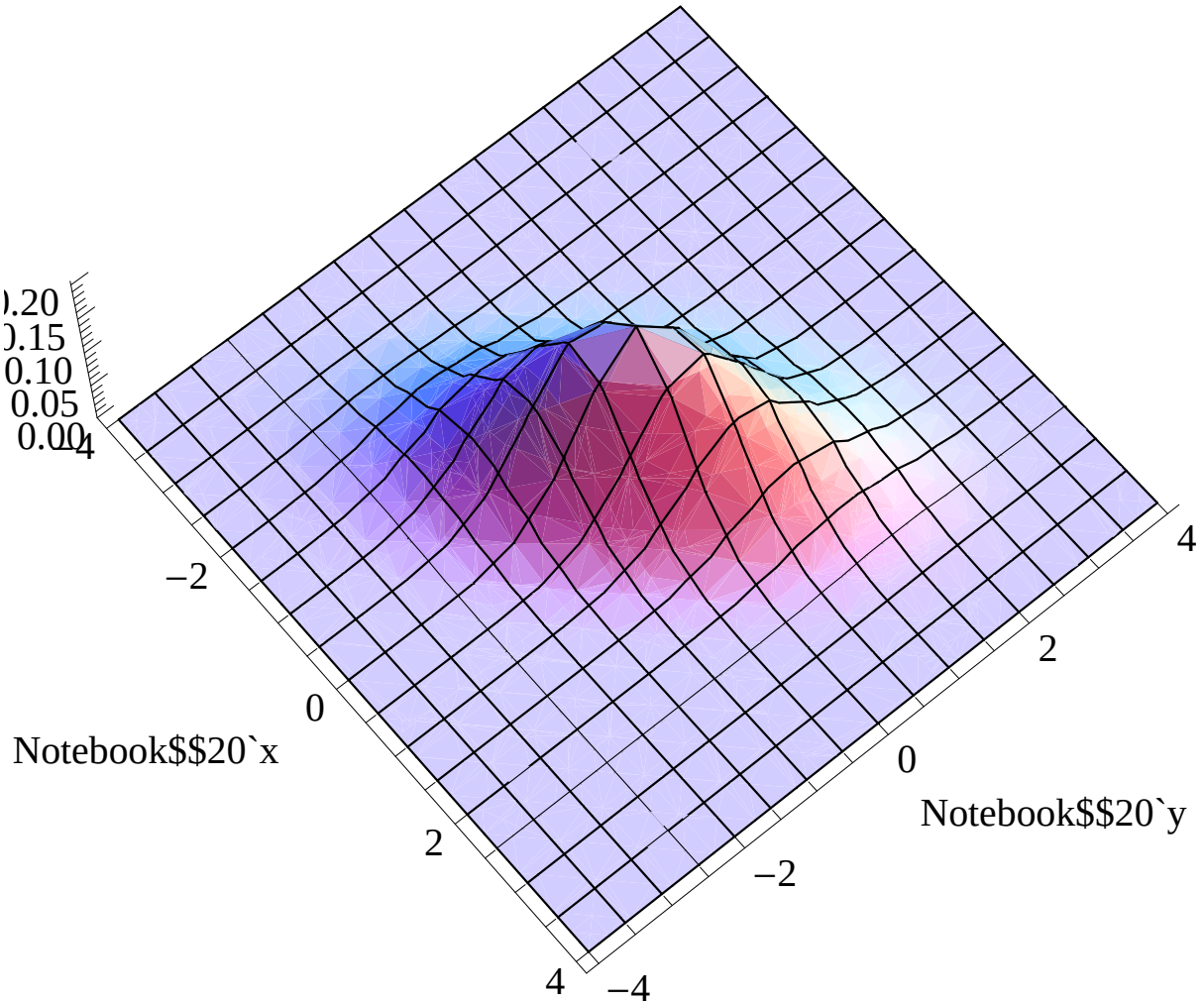
Bivariate Normal distribution.

$\rho = 0$



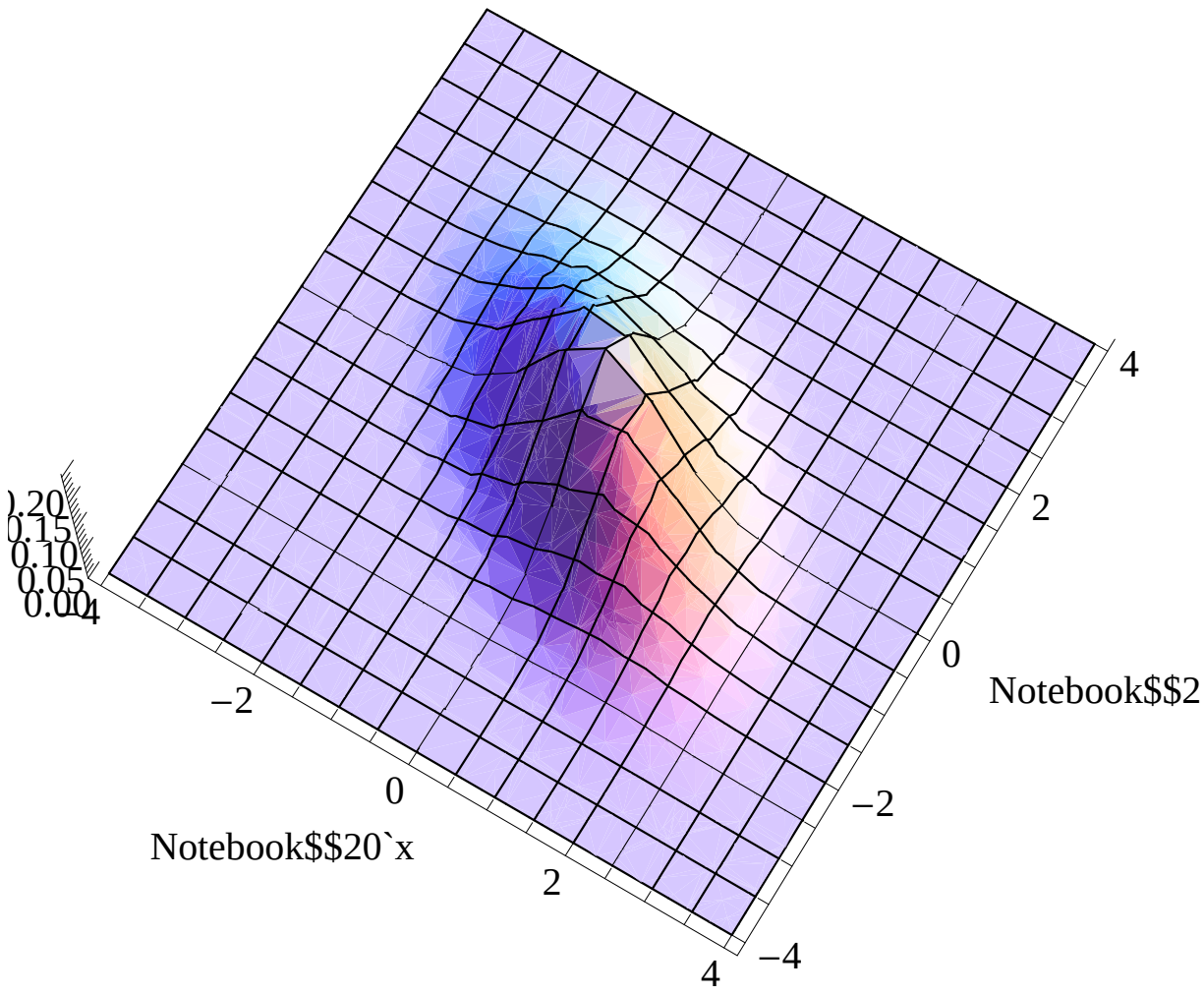
Bivariate Normal distribution.

$\rho > 0$



Bivariate Normal distribution.

$\rho < 0$



Self-selection model

Suppose that firm i self-selects choice E, equation 5 becomes:

$$\begin{aligned} y_i|E &= x_i'\beta + (u_i|Z_i\gamma + \varepsilon_i > 0) \\ &= x_i'\beta + (u_i|\varepsilon_i > -Z_i\gamma) \end{aligned} \tag{6}$$

Taking expectation of equation 6 we obtain the regression:

$$\begin{aligned} E(y_i|E) &= x_i'\beta + E(u_i|\varepsilon_i > -Z_i\gamma) \\ &= x_i'\beta + \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon}\frac{\phi(-Z_i\gamma)}{1-\Phi(-Z_i\gamma)} \\ &= x_i'\beta + \rho_{\varepsilon u}\sigma_u\frac{\phi(-Z_i\gamma)}{1-\Phi(-Z_i\gamma)} \end{aligned} \tag{7}$$

This result follows from the expected value of the truncated distribution from below. $\sigma_{\varepsilon u}$ is the covariance between ε and u, σ_ε is the standard deviation of ε, $\rho_{\varepsilon,u}$ is the correlation between u_i and ε_i and σ_u is the standard deviation of u

Self-selection as an omitted variable problem.

In case of choice *NE*

$$\begin{aligned} E(y_i|NE) &= x_i'\beta + E(u_i|\varepsilon_i \leq -Z_i\gamma) \\ &= x_i'\beta - \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon}\frac{\phi(-Z_i\gamma)}{\Phi(-Z_i\gamma)} \\ &= x_i'\beta - \rho_{\sigma\varepsilon}\sigma_u\frac{\phi(-Z_i\gamma)}{\Phi(-Z_i\gamma)} \end{aligned} \tag{8}$$

This result follows from the expected value of the truncated distribution from above.

Both equation can be written in a compact way

$$E(y_i|C) = x_i'\beta + \rho_{\sigma\varepsilon}\sigma_u\lambda_C(-Z_i\gamma) \tag{9}$$

where $C \in \{E, NE\}$ and $\lambda_C(-Z_i\gamma)$ is the conditional expectation of u_i given C and $\lambda_E(.) = \frac{\phi(-Z_i\gamma)}{1-\Phi(-Z_i\gamma)}$ $\lambda_{NE}(.) = -\frac{\phi(-Z_i\gamma)}{\Phi(-Z_i\gamma)}$

The omitted variable as private information (1)

- Comparing equation 1 and equation 9 it is clear that self-selection is an omitted variable problem.
- However in the probit model of equations 3 and 4 ε_i is the part of W_i not explained by public variables Z_i.
- Thus, ε_i can be viewed as the private information driving the corporate financing decision.
- The ex-ante expectation of ε_i is zero, given that it is the error term in the probit model

The omitted variable as private information (2)

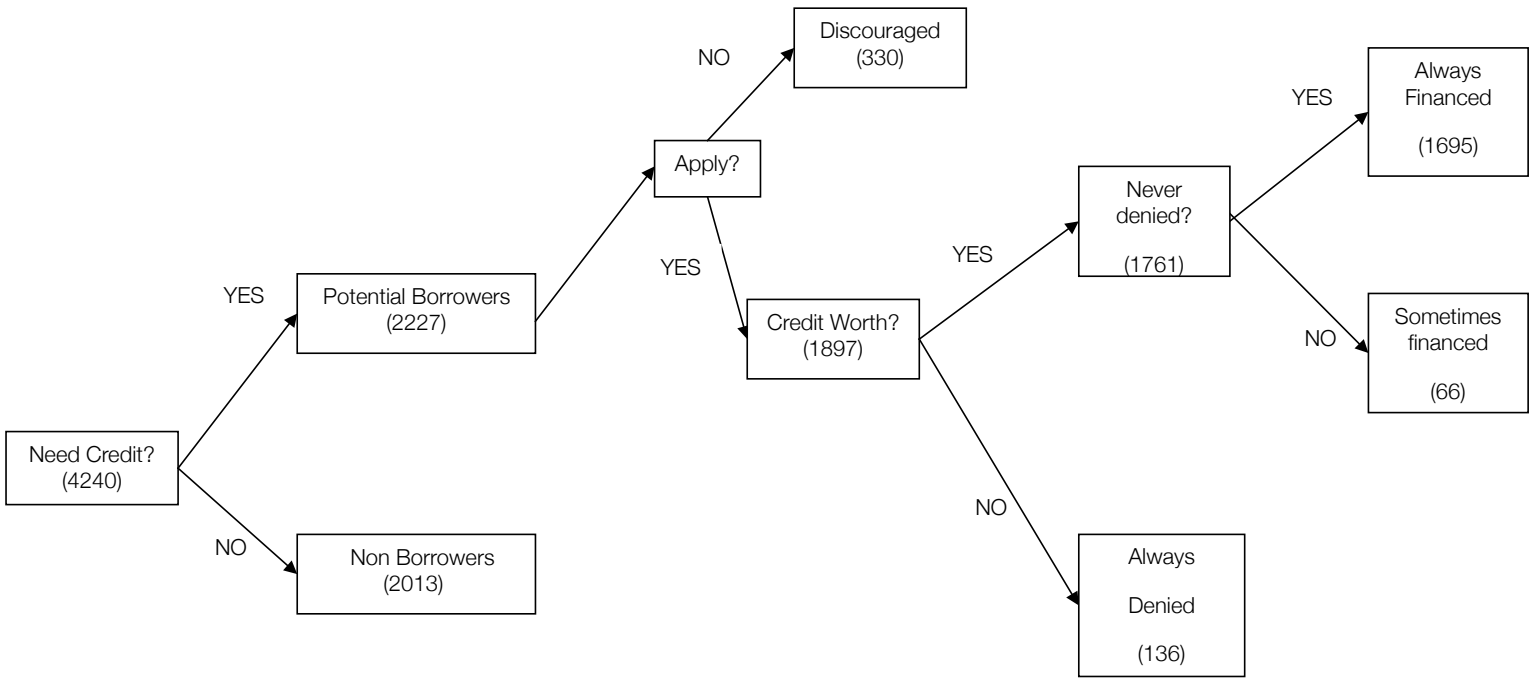
- Ex-post after that the firm i select $C \in \{E, NE\}$, the expectation of ε_i can be updated.
- The revised expectation $E(\varepsilon_i|C)$ is an updated estimate of firm's private information.
- If we want to test whether the private information in the firm choice affected post-choice outcome we would regress y on $E(\varepsilon_i|C)$
- But $E(\varepsilon_i|C) = \lambda_C(-Z_i\gamma)$ is the inverse Mills ratio that we add to adjust for selection
- The IMR is an estimate of the private information underlining a firm's choice and testing its significance is a test of whether private information explains ex-post outcomes.

Specification issues: exclusion restriction

- In estimating selection and outcome equation we have to specify two sets of variables: those determining selection (Z_i) and those determining the outcomes (x_i).
- At least one variable in Z_i must be different from those included in x_i (**exclusion restriction**)
- however very often variable that explain the selection also affect the outcome
- Actually, the Heckman selection model does not require exclusion restrictions because is identified by non linearity of Inverse Mills's Ratio. Under the assumption of bivariate normal errors IMR is a non linear function. However, in practice IMR may have little variation with respect to the other remaining variables included in the equation, leading to near-collinearity in the estimation.

Self-selection and access to credit (1)

Access to credit can be viewed as a sequence of decisions.



Self-selection in access to credit (2)

Access to credit. In the estimation of the probability to be financed we may observe two selection effects:

- higher returns increase the borrower willingness to give guarantees. Other things equal (e.g. wealth) more high borrowers are in the pool of applicants. The decision of the bank is taken on a sample of selected high return firms. Although the sample is non random this fact by itself does not bias the estimation. [See Achen, 1986]
- More importantly, some low return borrowers are in the pool of applicants. They decide to apply because they may have high probability of success, which is an unobservable variable. Thus, some observation in the financing equation that have low value of return have large error terms.

Self selection and access to credit (3)

- It is not important that probability of success and return are correlated in the population, they are correlated in the selected sample.
- If we assume that the probability of success does not lead to higher probability to be financed (banks are concerned by expected value), we will underestimate the effect of returns on the probability to be financed because borrower with low return are unusually safe.

Exclusion restriction

- Need for an exclusion restriction: Achen, 1986, p 99. “The researcher must know that some factor influencing selection [application for a loan] makes no difference in outcomes [financing decision by the bank]”.
- Difficult to find, at least for three reasons:
  1. the application for a loan and the financing by the bank involve similar decisions;
  2. the determinants of the application may be the same of those that induce the bank to finance or turn down;
  3. the self-selection decision may depend on the anticipated outcome from making that decision. In other words, given that borrowers are able to observe the bank financing model, they try to anticipate it in the decision to apply.

Endogeneity

- The issue in point 3 is important because firms want to anticipate the contract the bank will offer them. If the contract involve a positive probability to be rationed the question become crucial in presence of some positive cost of application.
- $\Rightarrow$ Selection and outcome equations may be simultaneous

Summary

Summing up, there are two linked issues suggested by the theory in the case of access to credit:

1. exclusion restrictions are difficult to identify;
2. selection and outcome equations may be simultaneous.

- In both cases we face the problem of finding exogenous instruments.
- A possible solution to point 1 is given by the Sartori estimator, which uses identical explanatory variables, assuming as identification instrument that errors for the same observation are the same in the two equations.
- A way to tackle the problem in point 2 is to use structural self-selection models (for example the Roy model).

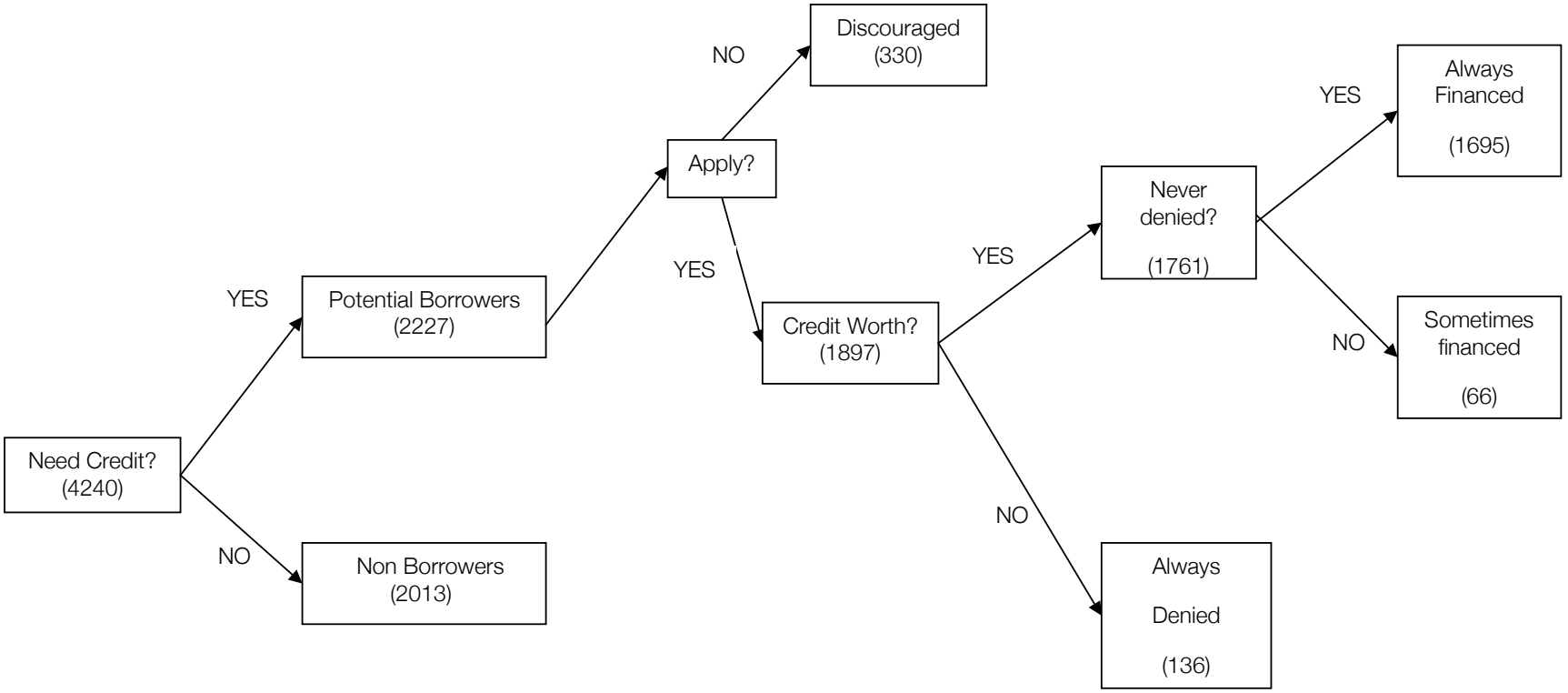
- Estimation methods for self-selection models applied to access to credit
- Examples of bivariate probit, bivariate probit with selection, Sartori's estimator, probit least square using STATA
- Data from the survey on small business finances (SSBF) of the Board of Governors of the US Federal Reserve Bank

Decisions trees

- Data are generated through a process involving subsequent decisions that lead to a wide range of outcomes.
- The decisions can be also simultaneous but can be represented with a decision tree.
- The nodes of these decision trees imply a selection into a certain choice such that the data after the node are a non random sample
- The determinants of all choices can be observable or not. Some private information can drive choices

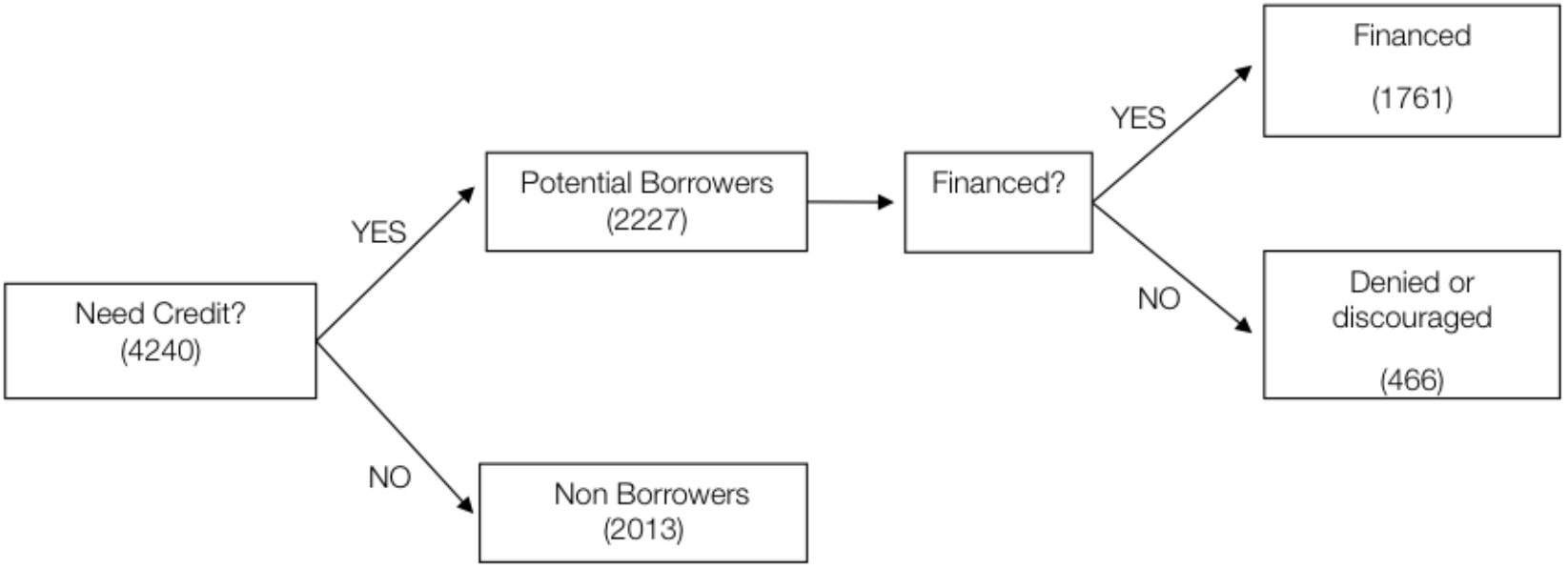
Decision tree: example one

Need, apply for, get credit



Decision tree: second example

Need, get credit

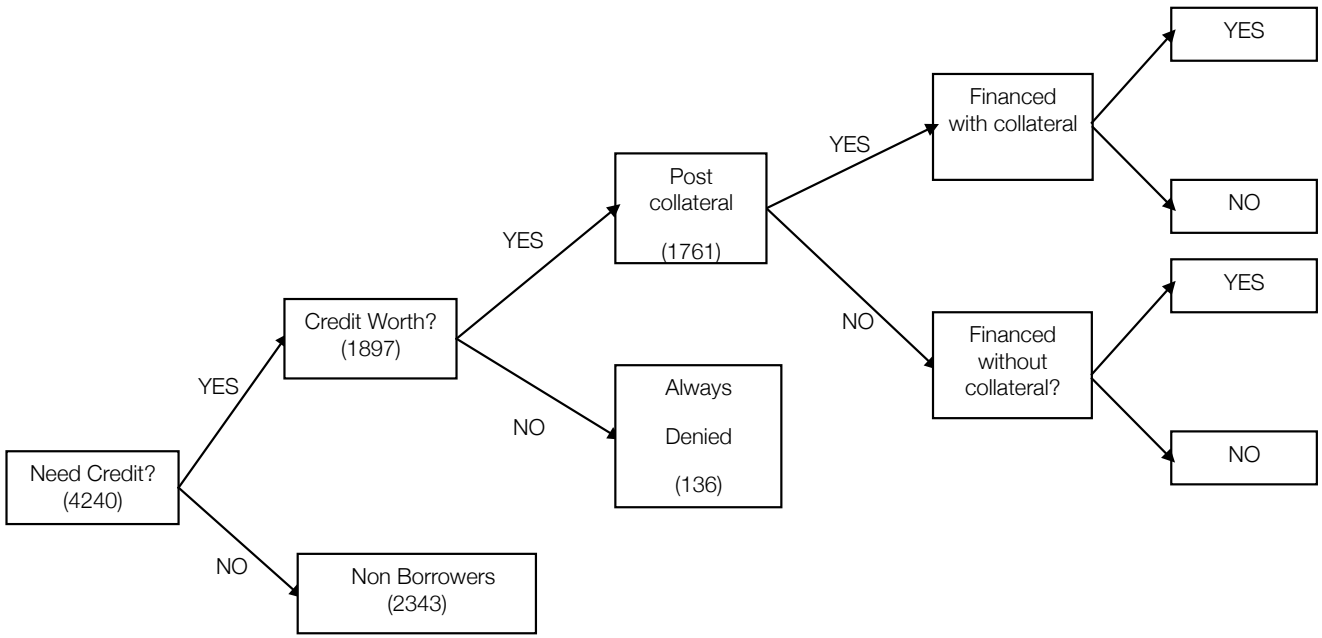


Depending on how the selection is specified the estimation highlights some aspects of the process. In the first tree the decision to apply is intermediate between the need of credit and being credit worthy. In the second tree Discouraged borrowers are pooled with Denied and no emphasis is put on the decision to apply

Decision tree: third example

Need credit, post collateral, get credit

In this example the emphasis is on the decision to post collateral.



Bivariate probit

In the third example the decision to post collateral (y_1) by the firm and that of the bank to grant or deny credit (y_2) can be modeled as a bivariate probit:

$$
\begin{aligned}
y_1^* = x_1'\beta_1 + \varepsilon_1 \quad y_1 &= \begin{cases} 1 \;\; \textit{if} \;\; y_1^* > 0, \\ 0 \;\; \textit{otherwise} \end{cases} \\
y_2^* = x_2'\beta_2 + \varepsilon_2 \quad y_2 &= \begin{cases} 1 \;\; \textit{if} \;\; y_2^* > 0, \\ 0 \;\; \textit{otherwise} \end{cases}
\end{aligned}
\tag{10}
$$

$E[\varepsilon_1|x_1, x_2] = E[\varepsilon_2|x_1, x_2] = 0 \;\; Var[\varepsilon_1|x_1, x_2] = Var[\varepsilon_2|x_1, x_2] = 1$

$Cov[\varepsilon_1, \varepsilon_2|x_1, x_2] = \rho$

Sample selection

There are nodes in the above decision trees in which the observed variables in the bivariate probit model are censored in one way or another. For example in the first tree the decision of the bank to grant a loan or to turn down the firm is observed for any given firm if the firm apply for credit.

$y_1 = 1$ if firm apply for a loan, 0 otherwise

$y_2 = 1$ if firm is granted a loan, 0 otherwise

Thus, there are three type of observation in the sample, with unconditional probabilities:

$$\begin{aligned} y_1 = 0 &: Prob(y_1 = 0|x_1, x_2) = 1 - \Phi(x_1'\beta_1) \\ y_2 = 0, y_1 = 1 &: Prob(y_2 = 0, y_1 = 1|x_1, x_2) = \Phi_2(x_1'\beta_1, -x_2'\beta_2, -\rho) \\ y_2 = 1, y_1 = 1 &: Prob(y_2 = 1, y_1 = 1|x_1, x_2) = \Phi_2(x_1'\beta_1, x_2'\beta_2, \rho) \end{aligned} \tag{11}$$

where Φ_2 denotes the cdf of the bivariate normal distribution.

Estimation of a Bivariate probit

Post collateral, get credit

In the dataset SSBF consider the following variables:

collatgiven	=1 if a collateral was required to secure the MRL, 0 otherwise
financed	=1 if the firm has been always financed, 0 otherwise

In terms of the third tree the variable Financed pools together the answers

YES to the questions Financed with collateral and Financed without collateral

IN STATA

```
 biprobit (financed=debtOnEq deliqobbl company firmage
minority owntotpw owndelinq fem) (collatgiven newline
hcredscor maturity grantOnTapp HHI company somtrat fem
lenghtrel own_family, noconstant)
```

Bivariate probit: Prob(post collateral) and Prob(get credit)

Seemingly unrelated bivariate probit — Number of obs = 8019
Wald chi2(18) = 754.46
Log likelihood = -6411.1975 — Prob > chi2 = 0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
financed						
debtOnEq	-.0041319	.0019707	-2.10	0.036	-.0079943	-.0002695
deliqobbl	-.1383178	.0242626	-5.70	0.000	-.1858716	-.0907641
company	.2485061	.0638232	3.89	0.000	.123415	.3735973
firmage	.0175843	.0027913	6.30	0.000	.0121135	.0230552
minority	-.342344	.0811259	-4.22	0.000	-.5013479	-.1833401
owntotpw	1.19e-08	7.92e-09	1.50	0.134	-3.64e-09	2.74e-08
owndelinq	-.5889388	.0836346	-7.04	0.000	-.7528596	-.425018
fem	.0153444	.0736895	0.21	0.835	-.1290844	.1597732
_cons	1.638356	.0767864	21.34	0.000	1.487857	1.788855
collatgiven						
newline	-.025466	.0394949	-0.64	0.519	-.1028744	.0519425
hcredscor	-.0890935	.0294603	-3.02	0.002	-.1468347	-.0313523
maturity	.005035	.0002815	17.89	0.000	.0044834	.0055867
grantOnTapp	-.1672034	.0272006	-6.15	0.000	-.2205156	-.1138911
HHI	.1135662	.0282421	4.02	0.000	.0582127	.1689196
company	.3040494	.0301615	10.08	0.000	.2449339	.3631649
somtrat	-.3537986	.3272726	-1.08	0.280	-.9952411	.2876439
fem	-.2218132	.0402749	-5.51	0.000	-.3007505	-.1428758
lenghtrel	-.0039549	.0013382	-2.96	0.003	-.0065778	-.001332
own_family	-.0787972	.0316073	-2.49	0.013	-.1407463	-.0168481
/athrho	-.2479885	.1582716	-1.57	0.117	-.5581952	.0622182
rho	-.2430269	.1489238			-.5066371	.062138

Likelihood-ratio test of rho=0: chi2(1) = 2.31884 Prob > chi2 = 0.1278

Estimation of a Bivariate probit with selection: be creditworth, get credit

In the dataset SSBF consider the following variables:

creditworth	=1 if the firm has been financed always or sometimes, 0 otherwise
financed	=1 if the firm has been always financed, 0 otherwise

We estimate the probability to get credit or to be rationed given in the selected sample of those firms that have been considered credit worthy duirng the last three years.

In STATA

```
 heckprob financed maturity loanrate lenghtrel deliqobbl
grantOnTapp debtOnEq company, select (credworth=
deliqobbl firmage minority debtOnEq owntotpw owndelinq
fem)
```

Bivariate probit with selection: Prob(get credit | creditworth=1)

Probit model with sample selection — Number of obs = 8679; Censored obs = 660; Uncensored obs = 8019

Log likelihood = -3075.76 — Wald chi2(7) = 204.00; Prob > chi2 = 0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
financed						
maturity	-.0017838	.0003816	-4.67	0.000	-.0025317	-.0010359
loanrate	-.0594472	.0087782	-6.77	0.000	-.0766523	-.0422422
lenghtrel	.0147091	.0032506	4.53	0.000	.008338	.0210802
deliqobbl	-.1768013	.0211031	-8.38	0.000	-.2181626	-.13544
grantOnTapp	.882111	.1424139	6.19	0.000	.6029849	1.161237
debtOnEq	-.0042723	.0019366	-2.21	0.027	-.0080679	-.0004768
company	.2595003	.0617165	4.20	0.000	.1385382	.3804624
_cons	1.383296	.1710936	8.09	0.000	1.047958	1.718633
credworth						
deliqobbl	-.1046446	.0189386	-5.53	0.000	-.1417635	-.0675256
firmage	.0102718	.0021952	4.68	0.000	.0059694	.0145743
minority	-.5957542	.0598595	-9.95	0.000	-.7130768	-.4784317
debtOnEq	-.0031336	.0014007	-2.24	0.025	-.005879	-.0003882
owntotpw	2.72e-07	2.63e-08	10.32	0.000	2.20e-07	3.24e-07
owndelinq	-.7292143	.0631054	-11.56	0.000	-.8528987	-.60553
fem	-.1471496	.0526441	-2.80	0.005	-.2503302	-.043969
_cons	1.397495	.0530942	26.32	0.000	1.293432	1.501558
/athrho	-.9086326	.4260499	-2.13	0.033	-1.743675	-.0735901
rho	-.7204753	.204894			-.9406514	-.0734575

LR test of indep. eqns. (rho = 0): chi2(1) = 23.70 Prob > chi2 = 0.0000

Private information (1)

- In a model of loan rate determination we want to evaluate the presence of private information.
- In the decision of posting collateral there might be some private information
- Calculate the Inverse Mills Ratio from a probit of the decision of posting collateral:

In STATA

```
probit collatgiven newline hcredscor maturity grantOnTapp HHI company somtrat fem lenghtrel
own_family, noconstant

**** this calculate Inverse Mills ratio
predict lp1, xb
gen imrcoll=normalden(lp1) / normal(lp1)
```

Private information (2)

```
reg loanrate collatgiven ratetype newline mort hcredscor numapp size HHI
ownersExper fem black othmino lenghtrel dist debtonass own_family imrcoll
if debtonass>=0
```

Source	SS	df	MS		
Model	8712.1087	17	512.476982		
Residual	44926.2707	7961	5.6432949		
Total	53638.3794	7978	6.72328646		

Number of obs	=	7979		
F(17, 7961)	=	90.81		
Prob > F	=	0.0000		
R-squared	=	0.1624		
Adj R-squared	=	0.1606		
Root MSE	=	2.3756		

loanrate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
collatgiven	-.3218323	.0568908	-5.66	0.000	-.4333532	-.2103114
ratetype	.9264006	.0592438	15.64	0.000	.8102673	1.042534
newline	-.2551645	.077267	-3.30	0.001	-.406628	-.1037011
mort	.2734013	.0989827	2.76	0.006	.0793692	.4674333
hcredscor	-.1546335	.0562865	-2.75	0.006	-.2649698	-.0442972
numapp	.0184061	.0106212	1.73	0.083	-.0024142	.0392265
size	-.3105727	.0156748	-19.81	0.000	-.3412995	-.2798459
HHI	.2383649	.0543703	4.38	0.000	.1317849	.3449449
ownersExper	-.0159079	.0027069	-5.88	0.000	-.0212142	-.0106015
fem	-.0579022	.0771734	-0.75	0.453	-.2091823	.0933779
black	.8945756	.2201961	4.06	0.000	.4629335	1.326218
othminority	.6620408	.1087385	6.09	0.000	.4488849	.8751967
lenghtrel	-.0094654	.0026163	-3.62	0.000	-.0145941	-.0043366
dist	.0006629	.0004277	1.55	0.121	-.0001755	.0015012
debtonass	.0278915	.0069939	3.99	0.000	.0141815	.0416014
own_family	-.2896242	.0669828	-4.32	0.000	-.420928	-.1583204
imrcoll	.3338205	.1207937	2.76	0.006	.0970333	.5706078
_cons	10.29503	.2636681	39.05	0.000	9.778171	10.81189