A firm level R&D analysis of Indian High-tech manufacturing : An empirical estimation through Heckman's two-step method using stata (11).

Sunil Kumar, Research scholar, IIT Indore
Dr.Ruchi Sharma, Assistant Professor,IIT Delhi

1.Introduction

- Deals with panel data problem of :
  - **Selection bias**
  - Heterogeneity
  - Endogeneity

1.1 .Selection problem

- Problem of sample selection arises either because of self selection by the individual or sample selection decision made by the data analysts.

$$y_i = x\beta + \upsilon_i \quad if \quad y_i \rangle 0 \quad (1)$$

$$y_i = Not\,observed \quad if \quad y_i \leq 0 \quad (2)$$

$$y_{2i} = z_i\delta + \upsilon_{2i} \quad (3)$$

$$D_i = 1 \quad if \quad y_i \rangle 0 \quad (3.1)$$

$$D_i = 0 \quad if \quad y_i \leq 0 \quad (3.2)$$

Selection problem continues

- The estimates of β is unbiased if the errors in these equation are independent.
- i.e

$$E[\upsilon_i|\upsilon_{2i}]=0 \qquad (4)$$

- So the data are missing 'randomly' and selection process is ignorable.
- Assumes that υ_i and υ_{2i} are jointly distributed: According to Baye's Rule

$$E[\upsilon_i|\upsilon_{2i}]=\lambda(Z\delta;\theta) \qquad (5)$$

Selection problem continues

- Therefore ,the probability of yi observed will be a sum of linear (1) and nonlinear (5) equation

- **Implications**

- 1. The estimate intercept will be biased because the mean of the disturbance is not zero
- 2. The estimate slope coefficient will be biased because Xs and Zs are not completely independent. A variable $\lambda(\mathrm{Z}\delta;\theta)$ is missing

The omitted variable case.

Selection problem continues

- See the Works of:-

- Gronau (1974) :- Price of time of house wife
- Lewis (1974) :- Measures the biases to which mean market wage of a person in group overstate their common wage offer distribution
- Heckman (1974) :- Estimates the female labour supply and wages

- *All of them discussed the participation decision*

Selection problem continues

- ***All the above works are relate to the cross sectional data.***

- Maddala(1983) : labour supply equation in the context of a panel data.
- Tobit model would the best method
- The model does not consider the fixed effect
- Honore (1992) : semi parametric estimator for fixed effect; but the unconditional fixed effects are biased

1.2. Unobserved heterogeneity

- Individual specific effect are unobservable because of their qualitative nature
- Hausman and Taylor (1981) : Transformation of data in to deviation from individual means

- Problems:

  1. All time invariant variables are eliminated
  2. Inefficient estimator due to variation across individuals are ignored

Unobserved heterogeneity continues

- **Simultaneous presence of selection bias and heterogeneity**
  - Nijamn and verbeek (1992) and Wooldridge (2010): Method of testing and correcting selection bias and unobservable effect(errors are normally distributed)
  - Kyriazidou (1997): Left the assumption of distribution of errors unspecified

1.3 Endogeneity

- Problem arises when individual unobserved effects are correlate with explanatory variables (idiosyncratic errors).
- Therefore, we concentrate on the three issues in a panel data frame work.

2. Variables

	Category	Variables
Dependent variables		RDD,RDI and RDS
Independent variables	Firm specific	AGE,SIZE and FOS
	Industry specific	HHI and ADVI
	Technology related	CI,SPILL and FLP
	Institutional factors	PATPOL,FTM,GID and TAR
	Demand and supply side	EXPI,PBTI and MGR

3. Data Sources

- All firms of High-tech and Medium-high-tech manufacturing (1995-2010)
- CMIE-prowess
- Department of Science and Industrial Research
- UNCTAD TRAINS database

4.Empirical Strategy

- Heckman's Two-step procedure
- There would be a selection equation and a primary equation
- Estimation with simultaneous presence of (Heterogeneity, Endogeneity and Selection bias)

Empirical strategy continues

$$y_{it} = x_{it}\beta + \upsilon_{it} + \varepsilon_{it} \qquad (6)$$

$$s_{it} = z_{it}\gamma + \upsilon_{it} + \varepsilon_{it} \qquad (7)$$

$$s_{it} = \begin{cases} 1 = if\, y_{it} \rangle 1, \\ 0 \;\; otherwise \end{cases} \qquad (7.1)$$

Υ_i

Empirical strategy continues

- Heckman's two step method Logit or probit model to estimate ϒ of equation2
- OLS to estimate β
- It was insufficient to deal with all these problem in a panel data frame work

- Kyriazidou (1997) : First differencing would eliminate the unobserved effect
- The problem of endogenity persist
- Instrumental variable approach

Empirical strategy continues

- A vector of instrument
- All exogenous variables plus vectors of instrument
- Semykina and Wooldridge (2010) look in to the three aspects
- Simple variable addition test
- Two-estimators with endogenous regressors
- Pooled 2SLS : Test for selection bias
- FE-2SLS : Correction for selection bias

4.1. Endogeneity test

Variable	OLS	Htaylor	Variable	OLS	Htaylor
AGE	-0.022(-0.87)	-0.022(-0.86)	*PATPOL*	0.015(0.34)	0.016(0.35)
FOS	0.000(-0.01)	0.000(-0.02)	*TAR*	0.216(1.74)	0.217(1.75)
PBTI	***-0.010(-1.15)***		*MGR*	0.000(-0.49)	0.000(-0.51)
SIZE	***-0.049(-4.89)***		*LARGED*	-0.013(-0.79)	
CI	0.109(2.47)	0.109(2.46)	*Constant*	0.024(0.68)	0.024(.66)
SPILL	-0.037(-0.91)	-0.036(-0.91)	***TV. Endogenous***		
FTM	0.006(0.6)	0.006(0.6)	*PBTI*		***-0.010(-1.18)***
HHI	-0.043(-1.66)	-0.043(-1.66)	*SIZE*		***-0.047(-4.62)***
ADVI	0.006(0.57)	0.006(0.58)	***TI. Exogenous***		
EXPI	0.004(0.54)	0.004(0.53)	*LARGED*		-0.013(-0.8)
GID	-0.042(-2.07)	-0.042(-2.07)			
Observation	8310	8310	*Observation*	8310	8310

4.2. Estimation with exogenous variable

- **Step 1.**Estimating the equation number (2) with a probit equation
  - Estimate the inverse mill's ratio
  - $$\lambda_{it} = \frac{\phi(z_{it}\Upsilon_i)}{\Phi(z_{it}\Upsilon_i)}$$
- **Step 2.** Add the IMR to the equation number (1) and estimate the coefficient by OLS

5. Results and Discussion

4.1. *Relationship between market concentration and R&D intensity*

		HHI	
		Low	**High**
RDI	**High**	Pharmaceuticals	Radio, T.V and communication equipment(RTC); Motor vehicles, trailers and semi-trailers(MOTOR)
	Low	Rail road equipment and transport equipment (RTE) ; Machinery and equipment (ME)	Office, accounting and computing machinery (OAC); Medical precision and optical instruments (MPO); Electrical (ELE)

5.1.Empirical results on R&D activity

- We have two part:
- The probit model that explains the probability of R&D decision and
- The OLS regression that explains the determinants of R&D investment.
- Introduce time and industry dummies
- As an alternative R&D stock as a dependent variable

- Negatively significant lambda value.

5.2 Probit estimation

Dependent variable : Dummy variable for R&D expenditure	
Positively Significant	**Negatively significant**
AGE, **FOS,** **GID ,** **PATPOL**	--------

5.3. OLS estimation

Result with R&D intensity

Positively significant	Negatively significant
CI ADVI	HHI SIZE

6. Findings and conclusion

- Improvement in appropriability condition enhances firms decision to invest in R&D expenditure.
- Government should continue its policy towards the innovativeness of the firms
- In a concentrated market firms investment is low
- Product differentiation is an influential factor in R&D

Thank you

Heckman selection model -- two-step estimates Number of obs = 8310
(regression model with sample selection) Censored obs = 4094
Uncensored obs = 4216

Wald chi2(2) = 1.17
Prob > chi2 = 0.5571

	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
dlrdi						
lgage	.4372694	.4045212	1.08	0.280	-.3555777	1.230116
ford	.1184523	.1148752	1.03	0.302	-.1066991	.3436036
_cons	-1.267669	1.102926	-1.15	0.250	-3.429364	.8940262
dtinv						
lgage	.9381254	.0444927	21.08	0.000	.8509213	1.025329
ford	.2918569	.0368944	7.91	0.000	.2195452	.3641685
dlpbti	-.0045727	.015424	-0.30	0.767	-.0348032	.0256577
_cons	-1.327991	.0625262	-21.24	0.000	-1.450541	-1.205442
mills						
lambda	.6335266	.6584919	0.96	0.336	-.6570939	1.924147

rho 0.64284
sigma .98551798
lambda .63352659 .6584919