Online Appendix

Learning by Exposure: Evidence from Foreign Direct Investment in Viet Nam

Nguyen, Ngoc Phuong Linh

(Draft)

May, 2025

1 Appendix A - Data

A.1 - Viet Nam Enterprise Survey - VNES

Data obtained from Mr. Nguyen Viet Phong, General Statistic Office of Viet Nam. The survey is conducted every year by government officials, with rules change from to year. While there is no one official record of the detailed rules for all years, I have managed to extract information from the Vietnamese government's official websites and related official guidelines. The information to each year's data collection method from 2010 up to 2025 are in the replication package (project/data/original_paperworks file). The general rules are:

- Time of survey is around March to May each following year of the year surveyed (for example, 2024 data is recorded in April-May of 2025)
- Most registered firms are surveyed and reported, except the following three sectors: sector O (Communist-related entities), sector U (national defense-related entities), and sector T (maid services, household activities to provide for their own use)
- Choice of sample changes from year to year, with different guidance administered by the government annually. For example, 2015 requires all firms with physical presence on Vietnamese land to fill out the survey, while 2020 commits to all big firms and foreign-

related firms but only sampled small domestic firms (firms owned by Vietnamese with less than 100 workers).

- There has been a transition of who conduct the survey. Originally, the Ministry of Planning and Investment handled the survey up until the GSO was established and took over the work around 2010. This might explain the hardship in obtaining pre-2011 related documents.

A.2 - Other datasets

Area code

- Area codes are not consistent over time due to a major code change in 2004, and minor merging/separating of areas throughout the 2000-2018 period. In my dataset, pre-2004 code is changed to be consistent with post-2004 code. Minor changes are not reflected in the dataset: provinces are kept inconsistent if it was merged or separated.

- Provincial code, district code, and ward code (post-2004) are provided by the courtesy of Mr. Nguyen Viet Phong from the Statistic Office of Viet Nam. The file can be accessed using the file *project/data/macro_data/population.xlxs/district_id* in the replication package.

- Provincial code (pre- and post-2004) can be publicly accessed by the General Statistic of Viet Nam (General Statistic Office, n.d.), which is only available in Vietnamese, but the interface is fairly simple.

Sector code

- Sector code underwent a major change in 2007.

- Sector code post-2007 are extracted from the government's legal document ID 10/2007/$ND{-}CP$, which is provided in *project/data/original_paperworks* in the replication package.

- Sector code pre-2007 are provided by the courtesy of Dr. Brian McCaig, Professor in Economics at Wilfrid Laurier University. A version of it can be found in *project/data/macro_data/sector.xlsx/sec_div_old* in the replication package.

Inflation and exchange rate

- Inflation data is obtained from GDP deflator[]

- Annual exchange rate (VND/USD) is publicly available at World Bank, n.d.-a.

- Vietnamese population is publicly available at World Bank, n.d.-b

Province's educational level, population, urbanization, sector code (Viet Nam Standard Industrial Classification 2007), GDP deflator are obtained from the website of General Statistic Office of Viet Nam. This data is public and updated annually.
GDP deflator is public through World Bank data.
Input-output table in public by the ADB data library.

1.1 Legal information

The FDI-related law are public (in Vietnamese language), obtained from legal documents from [].
Accounting rules are public (in Vietnamese language), obtained from the Ministry of Finance.

A.3 - Data choice for empirical observations

The data choice for the empirical observations only include 2009-2015 instead of the full-length dataset. This is because of the couple of reasons:

- I wanted to calculate the productivity using both inputs and investments. For inputs (*cogs* variable), the data was not collected before 2009. For investment (*investment* variable), it is obtained from the original value of fixed capital at the end and beginning of the year; however, after 2015, that data is either missing or at 0 for more than half of the dataset. Since the calculation of productivity needs accurate and consistent data collecting, such crop in data is important.

- Another reason is the input-output table. The IO data is only available in 2000 and from 2007 on.

- Only manufacturing and service are examined. There is no significant difference in the productivity distribution in agriculture. Furthermore, most agriculture "firms" are family-owned and not surveyed, which means the agriculture data is majorly missing.

Appendix B - Calculating TFPR

Each of these values are deflated using the GDP deflator. This is not an accurate evaluation of fixed asset value, since they are reported in original value; therefore, deflating them tends to underestimate their value. However, if we assume that the productivity of these fixed asset erodes overtime, part of the underestimation should be accounted for. A more accurate estimation should be a serial discounting of each year's fixed asset investment, yet this tends to overestimate the productivity of fixed assets.

Appendix C

Appendix C.1

Proposition: "If $\sum_{i=1}^{I} M_i > 0$, $\{g_i, f_i, f_{ei}\}_I$ satisfy the Assumptions 1-4, $f_1 = f_i$ for all i, and $M_1 > 0$, then there exists $i > 0$ such that $M_i > 0$."

Proof: We will use the free-entry and zero-profit cutoff conditions, in particular

$$\frac{1}{\epsilon} r(\varphi_i^*) - w f_i = 0, \forall i = 1, 2, ..., I \tag{1}$$

$$M_{ei} \begin{cases} > 0, & \frac{1}{\epsilon} r(\tilde{\varphi}_i) = \frac{w \delta_i f_{ei}}{1 - G_i(\varphi_i^*)} + w f_i \\ = 0, & \frac{1}{\epsilon} r(\tilde{\varphi}_i) < \frac{w \delta_i f_{ei}}{1 - G_i(\varphi_i^*)} + w f_i \end{cases} \tag{2}$$

Assume, by contradiction, no other type exists, then $\frac{1}{\epsilon} r(\tilde{\varphi}_i) < \frac{w \delta_i f_{ei}}{1 - G_i(\varphi_i^*)} + w f_i$ for all $i \neq 1$, and $g_i(x) = g_1(x)$ for all i and x on the domain of each type.

Applying (1), we have $\frac{f_1}{f_i} = \frac{r(\varphi_1^*)}{r(\varphi_i^*)} = \left(\frac{\varphi_1^*}{\varphi_i^*}\right)^{\epsilon - 1}$. Since $f_1 = f_i$ for all i, we have

$$\varphi_1^* = \varphi_i^* \text{ for all } i$$

This means all firm types have identical distribution $\gamma(.)$ to firm type 1.

Since $M_1 > 0$ and $M_i = 0$ for all $i \neq 1$, apply (2) for type 1 and type i then divide them against each other to get:

$$\left(\frac{\tilde{\varphi}_1}{\tilde{\varphi}_i}\right)^{\epsilon-1} = \frac{\frac{1}{\epsilon}r(\tilde{\varphi}_1)}{\frac{1}{\epsilon}r(\tilde{\varphi}_i)} > \frac{\frac{w\delta_1 f_{e1}}{1-G_1(\varphi_1^*)} + wf_1}{\frac{w\delta_i f_{ei}}{1-G_i(\varphi_i^*)} + wf_i} = \frac{\frac{\delta_1 f_{e1}}{1-G_1(\varphi_1^*)} + f_1}{\frac{\delta_i f_{ei}}{1-G_1(\varphi_1^*)} + f_1}$$

The RHS of the inequality is greater than 1 by Assumption 4 on fixed costs. The LHS of the inequality is equal to 1 as all types now have the same distribution. This is a contradiction. Therefore, at least another type besides type 1 exists.

Appendix C.2

Assume $f : (-\infty, \infty) \to (-\infty, \infty)$ such that for all $x_1, x_2 \in (-\infty, \infty)$,

$$f(e^{x_1}) - f(e^{x_2}) = (\epsilon - 1)(x_1 - x_2) \Rightarrow x_1 = [f(e^{x_1}) - f(1)]/(\epsilon - 1)$$

When X follows a normal distribution $N(\mu, \sigma)$. We have

$$\Phi\left(\frac{x-\mu}{\sigma}\right) = Pr(X \le x) = Pr\left(\frac{f(e^X) - f(1)}{\epsilon - 1} \le x\right) = Pr(f(e^X) \le x(\epsilon - 1) + f(1))$$

This is equivalent to

$$Pr(f(e^X) \le t) = \Phi\left(\frac{t - f(1) - \mu(\epsilon - 1)}{\sigma(\epsilon - 1)}\right)$$

which mean $f(e^X)$ follows a normal distribution with mean $\mu(\epsilon - 1) + f(1)$ and standard deviation $\sigma(\epsilon - 1)$.

In our model, $\frac{r(\varphi)}{r(\varphi')} = (\frac{\varphi}{\varphi'})^{\epsilon-1}$, with φ follows a log-normal distribution with parameters μ and σ (which are the mean and standard deviation of $\ln(X)$). Take log both sides, let $f(\varphi) = \ln(r(\varphi))$ and $x = \ln(\varphi)$, we have $\ln(r)$ also follow a normal distribution with mean $\mu(\epsilon - 1) + \ln(r(1))$ and standard deviation $\sigma(\epsilon - 1)$.

Appendix C.4

To find the solution for our competitive general equilibrium, I will first find $\{\varphi_i^*\}_{i=1,2,...}$ for types i that exists, then M_1/M, which then pins down the whole distribution for each type and their statistics.

C.4.1. Solution for unique φ_1^*:

Assuming that fixed entrance costs for type 1 is sufficiently low such that firm type 1 exists, by Proposition **??**, both types exist. Combining ZPC (1), FE (2), and the revenue ratio in (**??**), we can solve for the unique cutoff φ_1^* as

$$\frac{\delta_1 f_{e1}}{f_1[1 - G_1(\varphi_1^*)]} - 1 = \left(\frac{\tilde{\varphi}_1}{\varphi_1^*}\right)^{\epsilon-1} \tag{3}$$

This pins down φ_1^* uniquely, as G_1 is known and not dependent on M_i distribution. The proof for a general g distribution is in Melitz, 2003, appendix B.1. In our case, by fully expanding $\tilde{\varphi}_1$ and G_1, then set $t = \tilde{\varphi}_1/\varphi_1^*$, we can prove that (3) is equivalent to

$$\frac{\delta_1 f_{e1}}{f_1} = \int_1^\infty (t^{\epsilon-1} - f_1)\frac{1}{t\sigma_1\sqrt{2\pi}} \exp\left[-\frac{(\ln(t) + \ln(\varphi_1^*) - \mu_1)^2}{2\sigma_1^2}\right] dt \tag{4}$$

From here, we can solve for φ_1^*.

C.4.2. Solution for unique $\{\varphi_i^*\}_{i\neq 1}$:

For other types, if type i exists, then it must be true from the zero-profit condition that

$$\frac{f_1}{f_2} = \left(\frac{\varphi_1^*}{\varphi_2^*}\right)^{\epsilon-1} \tag{5}$$

and the second one pins down φ_i^* uniquely. Here, we can see that neither technological differences ($\sigma_1 - \sigma_2$ and $\mu_1 - \mu_2$) nor the high type's penetration $\left(\frac{M_1}{M}\right)$ affects the cutoff of either low or high type. This means that, assuming type 1 is foreign firms and type 2 is domestic firms, having fewer foreign firms will not induce the entrance of domestic firms with very low productivity.

C.4.2. Solution for $r(.)$:

Using φ_1^*, we get a unique $\tilde{\varphi}_1$ by definition because type 1's distribution is fixed. Using the ZPC and FE conditions, we can also identify $r(\varphi_1^*)$ and $r(\tilde{\varphi}_1)$ uniquely as

$$r(\varphi_i^*) = \epsilon w f_1$$

$$r(\tilde{\varphi}_1) = \frac{\epsilon w \delta_1 f_{e1}}{1 - G_1(\varphi_1^*)} + \epsilon w f_1$$

Proof for the independence of $r(\tilde{\varphi}_1)$ with respect to μ_1

From equation (4), we have $\ln(\varphi_1^*) - \mu_1$ is dependent on $\delta_1, f_{e1}, f_1, \sigma_1$. Now, by expanding $1 - G_1(\varphi_1^*)$ using its definition, we have

$$\begin{aligned} 1 - G_1(\varphi_1^*) &= \int_{\varphi_1^*}^{\infty} \frac{1}{\varphi \sigma_1 \sqrt{2\pi}} \exp\left[-\frac{(\ln(\varphi) - \mu_1)^2}{2\sigma_1^2}\right] d\varphi \\ &= \int_{1}^{\infty} \frac{1}{t \sigma_1 \sqrt{2\pi}} \exp\left[-\frac{(t + \ln(\varphi_1^*) - \mu_1)^2}{2\sigma_1^2}\right] dt \end{aligned}$$

which is also dependent on $\delta_1, f_{e1}, f_1, \sigma_1$ and not μ_1. This means $r(\tilde{\varphi}_1)$ is also independent of μ_1 by the free-entry condition.

C.4.2. Solution for unique $\frac{M_1}{M}$:

For any type i that exists, using the free-entry condition for type i,

$$\tilde{\varphi}_i^{\epsilon-1} = r(\tilde{\varphi}_i) r(\tilde{\varphi}_1)^{-1} \tilde{\varphi}_1^{\epsilon-1} = \left(\frac{\epsilon w \delta_i f_{ei}}{1 - G_i(\varphi_i^*)} + \epsilon w f_i\right) r(\tilde{\varphi}_1)^{-1} \tilde{\varphi}_1^{\epsilon-1} \quad (6)$$

which is a function of known variables and $\frac{M_1}{M}$ (embedded in G_i). We also know from definition

$$\tilde{\varphi}_i^{\epsilon-1} = \frac{1}{1 - G_i(\varphi_i^*)} \int_{\varphi_i^*}^{+\infty} \varphi^{\epsilon-1} g_i(\varphi) d\varphi \quad (7)$$

which is another function of known variables and $\frac{M_1}{M}$ (embedded in G_i and g_i). Setting (6) equals to (7), we can solve for M_1/M. The uniqueness of M_1/M comes from the fact that in (6), $\tilde{\varphi}_i$ is strictly decreasing in M_1/M; while in (7), it is strictly increasing in M_1/M. elaborate more

In the two-type case, the problem is simplified that there is only one other type $i = 2$. However, when $i > 2$, applying the above method to each type i can yield a different solution to M_1/M. This is because such solution is only correct if type i exists. However, there is only one solution for M_1/M that satisfies the free-entry condition for all types, in which the expected average profit for each type is non-positive. need to be a little more technical on this.

Solution for other variables