# Reassessing Selection Effects on Sectoral Productivity Gaps: Evidence from Indonesia

Isabella Germinario

September 30, 2025

**Abstract**

This paper examines the role of individual sorting in sectoral productivity gaps in Indonesia between 1993 and 2000, using the Indonesia Family Life Survey (IFLS). The analysis shows that sector-wide technology differences account for most of the observed gap, while individual sorting based on unobserved comparative advantages plays only a minor role—in contrast to prior studies that report large selection effects. Three key contributions of this paper are: First, it demonstrates that the aggregate effect of sorting consists of two components—the extra return to unobserved comparative advantage and the difference in mean latent abilities across sectors—whereas earlier studies emphasize only the former. Second, it applies a correlated random coefficient framework to study productivity gaps, estimating the return component without imposing distributional assumptions. Third, it maps the resulting selection terms to the classical Roy model, linking alternative approaches. The findings suggest that improving agricultural technology and infrastructure, rather than labor reallocation policies such as vocational training, is key to narrowing productivity gaps in 1990s Indonesia.

## 1 Introduction

Agricultural productivity in low-income countries lags significantly further behind that of high-income countries, compared to their non-agricultural sectors. To make matters worse,

the majority of people in low-income countries work in this less productive sector. Hence, agricultural productivity plays a crucial role in explaining the massive cross-country income gaps (Gollin et al., 2002; Caselli, 2005; Chanda and Dalgaard, 2008; Restuccia et al., 2008). Poor countries exhibit a pronounced and persistent productivity disparity between the agriculture and non-agriculture sectors, posing a fundamental development challenge (McMillan and Rodrik, 2014). Implementing effective pro-growth policies requires a deep understanding of the mechanisms underlying the substantial productivity gap between the agricultural and non-agricultural sectors in developing countries.

The structural transformation literature, pioneered by Lewis' seminal analysis (1954) of the dual economy and inter-sectoral labour movements, has since adopted the term, Agricultural Productivity Gap (APG)—the ratio of value-added per worker in non-agriculture relative to agriculture, to quantify sectoral productivity gap (Gollin et al., 2014). In what follows, I use the terms "sectoral productivity gap" and "APG" interchangeably.

In this body of literature, two prominent potential explanations for substantial APG in developing countries are: One is due to a misallocation of resources, such as land, labour, capital, or investments (Restuccia et al., 2008; Bryan et al., 2014; Munshi and Rosenzweig, 2016; Alvarez-Cuadrado et al., 2017; Lagakos, 2020). The alternative hypothesis posits that potential causes involve agents sorting into sectors based on their comparative advantages. Under the condition that comparative and absolute advantages are positively correlated, individual sorting could amplify the sectoral productivity gaps in poor countries (Lagakos and Waugh, 2013; Alvarez-Cuadrado et al., 2020).

This paper examines how much individual selection based on unobserved comparative advantage contributes to sectoral productivity gaps in Indonesia, a low-income country that underwent considerable structural transformation between 1993 and 2014. The answer to this question has far-reaching policy implications. If most of the observed gap reflects efficient sorting, then the disparities largely mirror workers' latent abilities and policy levers are limited. But if sector-wide technology is the primary driver, then improvements in seeds, irrigation, or infrastructure—as demonstrated by the Green Revolution in Asia during the 1960s–1980s—can dramatically narrow productivity gaps without changing who works in agriculture (Evenson and Gollin, 2003; David and Otsuka, 1994). Conversely, misdiagnosing

the role of selection risks wasted effort: in Sub-Saharan Africa, large-scale vocational training and microenterprise programs in the 1990s–2000s often assumed farmers could seamlessly transition into non-farm jobs and succeed, yet many fell short, one possible reason being that skill differences between sectors were underestimated (Blattman and Ralston, 2015; McKenzie, 2017). Understanding how much individual sorting contributes to APG is therefore central to designing effective strategies for structural transformation and sustainable growth in low-income countries.

A large literature, rooted in the Roy model, studies how occupational choice affects individual earnings and defines the selection effect as the average extra return to unobserved abilities. This paper follows that convention. My research question, however, is different: how much does individual sorting contribute to the sectoral productivity gap? When the outcome of interest shifts from individual to sectoral earnings, aggregating to the sector level introduces an additional component. At the aggregate level, the impact of sorting has two parts: (1) the selection effect—the extra return to unobserved abilities—and (2) the average difference in unobserved abilities between workers across sectors. Only when the two groups are alike on average does the selection effect alone capture the full contribution of sorting to the APG. This paper builds a framework to quantify both the selection effect and the group differences in latent abilities, answering the research question posited.

Having laid out the two components of sorting's sectoral contribution, I now examine the first component, selection effect. Since Lagakos and Waugh's (2013) self-selection hypothesis, a growing literature has sought to measure how much this effect explains productivity gaps in developing countries (Lagakos and Waugh, 2013; Pulido and Świecki, 2019; Alvarez, 2020; Lagakos et al., 2020; Alvarez-Cuadrado et al., 2020; Adamopoulos et al., 2022). These studies consistently report that selection is important, but the estimated magnitudes vary widely. This variation reflects the difficulty of the task: estimates depend crucially on how heterogeneity in latent abilities is modeled and how the endogeneity of sector choice is addressed. The stakes are high, because mismeasuring selection risks misdirecting policy—leading governments to focus on moving workers across sectors when the real constraint may be technology, or vice versa.

In the APG literature, two prevailing approaches stand out when measuring selection on

sectoral productivity gap. The first applies two-way fixed effects to panel data, attributing the gap reduction after controlling for individual effects to selection. But this masks comparative advantage, since all time-invariant traits—including those irrelevant to sector choice—are swept into the fixed effect, leading to upward bias. The second imposes a distribution for unobserved abilities (e.g. joint normal or Fréchet) and recovers a closed-form parameter in the spirit of Roy model (1951) framework. While tractable, these strong assumptions rarely have empirical support (Heckman and Honore, 1990), and they heavily influence estimated magnitudes.

To address these limitations, this paper adopts a Correlated Random Coefficient (CRC) framework, following Suri (2011), and applies it to the context of sectoral choice and productivity gaps. The CRC approach models unobserved abilities as time-invariant and sector-specific, which distinguishes individual fixed effects that matter for sectoral choice from those that do not. By exploiting the information embedded in individuals' sectoral choice histories, this method reveals their latent comparative advantages and recovers the selection effect without relying on strong distributional assumptions about unobserved heterogeneity.

The estimation proceeds in two stages. First, I estimate the reduced-form empirical model using Seemingly Unrelated Regressions (SUR). Second, I recover the structural parameters with a Minimum Distance Estimator (MDE). Both steps are implemented in the STATA package **randcoef** (Cabanillas et al., 2018).

Most of the APG literature treats the selection effect as if it fully captured the role of sorting in productivity gaps. In practice, however, it reflects only one component: the extra earnings workers receive when their unobserved skills are more highly rewarded in one sector than the other. Once individual earnings are aggregated to the sector level, a second component arises—the average difference in unobserved skills between agricultural and non-agricultural workers. This difference shapes the observed APG unless farmers and non-farmers are very similar on average. In that special case, the mean difference is negligible, and the selection effect alone accounts for sorting's contribution. But when the groups differ, focusing only on the first component gives an incomplete and potentially misleading picture. My framework measures both components to assess how individual sorting contributes to APG.

To examine these two components empirically, I use the Indonesia Family Life Survey (IFLS) (Frankerberg et al., 1995), a nationally representative panel dataset spanning five waves from 1993 to 2014 and covering about 80 percent of the population. During this period, Indonesia moved from low- to lower-middle-income status, with per capita incomes roughly doubling and the share of agricultural employment falling by about 12 percent. The IFLS is particularly well suited for studying sorting and productivity gaps because of its rich panel structure, which tracks both sector choices and earnings. In this paper, I focus on the first three waves (1993–2000), before Indonesia's sweeping political and institutional changes after 2000, to provide a clean setting for analyzing sectoral productivity gaps.

Applying this framework to the first three waves of the IFLS (1993–2000), I find that Indonesia exhibited large sectoral productivity gaps, but individual sorting played only a limited role in explaining them. Roughly 80 percent of workers remained in their initial sectors, and among those who switched, the estimated selection effect was statistically insignificant. The average difference in unobserved abilities between farmers and non-farmers accounted for only about 2 percent of the gap, while sector-wide technology differences explained most of the disparity. These findings stand in contrast to other studies using the same data that report large selection effects, underscoring how methodological choices shape conclusions about the sources of APG.

This study advances the APG literature in three ways. First, it establishes that the sectoral impact of individual sorting has two components: the extra return to unobserved comparative advantage and the average difference in latent skills between workers across sectors. Earlier studies treat only the former as the effect of sorting on APG, overlooking the latter. Second, it adapts Suri's (2011) correlated random coefficient framework to the study of sectoral productivity gaps by redefining sector-specific abilities so the model can be applied in this context. In doing so, the framework separates abilities that influence sectoral choice from those that do not. By exploiting workers' sectoral choice histories, it then recovers the selection effect without relying on restrictive distributional assumptions. Third, it maps the resulting terms back to the classic Roy model, making the connection between alternative approaches explicit and situating the findings within the broader self-selection literature.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature on sectoral productivity gaps and self-selection. Section 3 defines the measurement of APG and develops the empirical framework based on Suri's (2011) correlated random coefficient model, linking it to the classic Roy framework. Section 4 sets out the identification strategy and estimation procedures. Section 5 introduces the dataset, provides descriptive evidence, and presents baseline results for the first three IFLS waves. Section 6 discusses the findings in light of the literature. Section 7 concludes.

# 2 Literature Review

This section situates the paper within the literature on sectoral productivity gaps (APG), with particular attention to studies examining the role of individual self-selection. Two dominant hypotheses explain persistent productivity gaps in low-income countries: misallocation of resources and sorting based on comparative advantage. While not mutually exclusive, distinguishing their relative importance is critical for interpreting observed gaps. If misallocation dominates, then the APG signals inefficiencies that policy can alleviate. If sorting dominates, then much of the gap reflects workers' latent abilities, and welfare-enhancing interventions play a smaller role.

The review proceeds in three steps. First, it discusses the origins of the APG literature and the theoretical motivation for focusing on selection. Second, it synthesizes empirical evidence on the magnitude of selection effects, highlighting the wide range of reported estimates. Third, it identifies the limitations of prevailing empirical approaches, motivating the need for the alternative framework developed in this paper.

## 2.1 APG and the Selection Hypothesis

The study of sectoral productivity gaps is rooted in classic theories of structural transformation and growth (Lewis, 1954; Kuznets, 1971). A consistent empirical finding is that agricultural productivity lags far behind non-agricultural productivity in poor countries, with gaps far larger than those observed in rich economies (Gollin et al., 2002; Restuccia et al., 2008; Gollin et al., 2014; McMillan and Rodrik, 2014). Because low-income countries

employ a high share of workers in agriculture, this productivity disadvantage contributes directly to cross-country income inequality.

Early critics questioned whether such large gaps merely reflected measurement error in national accounts. Gollin, Lagakos, and Waugh (2014), however, demonstrated using household survey data that the APG remains large even with microdata, dispelling this concern. Herrendorf and Schoellman (2018) further showed, across 42 censuses in 13 countries over seven decades, that poor countries consistently exhibit large wage differences between sectors. Together, these studies confirmed that APGs are real, persistent, and central to understanding income disparities.

Two mechanisms have been emphasized. Misallocation arises when barriers to mobility or frictions in input use prevent efficient allocation of workers and resources (Restuccia et al., 2008; Munshi and Rosenzweig, 2016; Bryan et al., 2014). Selection, by contrast, highlights that individuals choose sectors based on comparative advantage. In Roy's (1951) framework, Lagakos and Waugh (2013) formalized this hypothesis by combining comparative advantage with subsistence food requirements: in low-productivity economies, many workers must remain in agriculture, creating large dispersion in productivity among farmers, while in high-productivity economies only highly skilled farmers remain, raising average productivity. When comparative and absolute advantage are positively correlated, sorting amplifies the APG. Later work noted that selection and misallocation likely coexist, and that the correlation between comparative and absolute advantage may be weak or even negative in poor countries (Alvarez-Cuadrado et al., 2020; Lagakos, 2020).

Thus, the selection hypothesis provides a compelling microfoundation for observed APGs, but the empirical challenge lies in measuring its magnitude.

## 2.2 Empirical Evidence on Selection and APG

Since Lagakos and Waugh, a growing literature has attempted to quantify how much of the APG is attributable to individual sorting. Reported magnitudes, however, vary widely—from one-third of the observed gap to nearly all of it—depending on the empirical strategy and country context.

Structural calibration approaches. Several studies estimate reduced-form relationships

in microdata and use them to calibrate macro models. Lagakos and Waugh (2013) assumed Fréchet-distributed abilities and argued that selection strongly amplifies gaps, even in the absence of frictions. Pulido and Swiecki (2019), using IFLS data, imposed joint normality and concluded that selection explains 45–70 percent of the APG, depending on substitution elasticities. Both emphasize the puzzle of two-way sector transitions despite large observed wage gaps, interpreting it as evidence of misallocation.

Herrendorf and Schoellman (2018) and Alvarez (2020) relaxed functional form assumptions. Herrendorf and Schoellman used multi-country census data to model wage returns and found sorting important but not dominant relative to frictions. Alvarez, using Brazilian panel data, compared wages of switchers and multi-sector workers: average gaps were nearly 50 log points, but within-individual wage gains from switching were small (4–9 log points), implying limited misallocation and a large role for sorting.

Gai et al. (2021) and Lagakos et al. (2020) examined rural–urban migration. Both concluded that migration costs matter, but disagreed on magnitudes: in some contexts misallocation dominated, in others selection explained more. Adamopoulos et al. (2022), studying rural China, emphasized land and capital misallocation, but found that sorting further amplified its effects.

Reduced-form panel approaches. Hamory et al. (2021) applied two-way fixed effects to panel data in Kenya and Indonesia, reporting that 67–92 percent of observed APGs disappear once individual effects are controlled for. Their interpretation is that most of the APG reflects sorting.

Taken together, this empirical literature reaches a consistent qualitative conclusion—that selection matters—but reported magnitudes differ enormously, from 32 percent to over 90 percent. This divergence reflects the deep methodological challenges of modelling heterogeneity and endogeneity in sectoral choice.

## 2.3 Limitations of Prevailing Approaches

In the APG literature, the "selection effect" has generally been treated as if it were identical to the classical Roy model's notion of selection on individual earnings. This framing overlooks an important distinction: in the Roy setting, the outcome of interest is an individual's wage,

whereas in the APG contex,t the outcome is the sectoral earnings gap. By simply equating the two, existing studies implicitly treat the individual-level return component as if it fully captures sorting's role in sectoral gaps, thereby overlooking the possibility that group-average differences in latent abilities also matter.

Regarding the selection effect, the empirical strategies used to estimate it fall into two dominant approaches. The first is to apply two-way fixed effects (TWFE) to panel data, interpreting the gap reduction after controlling for individual effects as the effect of sorting. While appealing for its simplicity, this method masks comparative advantage: all time-invariant traits—including those irrelevant to sector choice—are absorbed into the fixed effect, leading to upwardly biased estimates of selection.

The second is to impose a distributional form on unobserved abilities, most often assuming joint normality or Fréchet, and then recover a closed-form selection parameter in the spirit of Roy. This structural approach is tractable but fragile. Its results depend heavily on functional-form assumptions that lack strong empirical support in either labour or development economics, as noted by Heckman and Honore (1990).

Together, these limitations leave the magnitude of the selection effect highly sensitive to methodological choices. In practice, existing studies report wide-ranging estimates, all significant but inconsistent in size, precisely because they conflate individual-level and aggregate concepts of selection and lean on restrictive assumptions to handle heterogeneity and endogeneity. To address these challenges, this paper adapts the correlated random coefficient framework of Suri (2011), which avoids parametric assumptions about unobserved heterogeneity and better isolates the sector-relevant component of individual abilities.

A fuller description of Suri's original CRC framework and its application to hybrid seed adoption in Kenya is provided in Appendix A, for readers seeking additional background before turning to this paper's adaptation to the APG setting.

# 3 Empirical Model

This section outlines the empirical model of this paper by extending the approach in Suri (2011) to the APG literature. I begin by modelling individual earnings by using the Mince-

rian representation for human capital. Building on this foundation, I follow the method in Lemieux (1998) and Suri (2011) to incorporate heterogeneous latent skills across sectors into this model, where absolute and comparative advantages are defined. Next, I put together the main empirical model that enables the measurement of how self-selection affects individual earnings. Finally, I illustrate how to draw the impact of individual selection on the sectoral productivity gap through aggregation.

To anchor the selection effect of my model, I further map it to the types of selection proposed by Borjas (1987) in the classic Roy's model framework and discuss the selection effect in the adopted model in Appendix B.

## 3.1 Model Setup

As individual sorting is based on comparing the potential earnings in each sector, the starting point is to model how the choice of sector affects an individual's potential earnings. In an economy, an individual $i$ at time $t$ can choose to work in one of two sectors: $j \in \{n, a\}$, where $n$ refers to the non-agricultural sector and $a$ to the agricultural sector. An individual's potential earnings in each sector at each period are determined by the sector productivity, $A_t^j$, and her own human capital, $h_{it}^j$, as expressed in equations (1) and (2). In these two equations, $P_t^j$ represents the price level at each sector $j$, which is the returns to sector technology level.

$$W_{it}^n = P_t^n A_t^n h_{it}^n \tag{1}$$

$$W_{it}^a = P_t^a A_t^a h_{it}^a \tag{2}$$

Following Mincer regression, the human capital $h_{it}^j$ can be expressed as an exponential function of observed and unobserved characteristics, see equations (3) and (4), where $X_{it}$ is a vector of observed characteristics endowed by individual $i$ at time $t$, and $U_{it}^j$ is a vector of individual $i$'s unobservable in sector $j$ at time $t$.

$$h_{it}^n = exp(X_{it}\gamma^n + U_{it}^n) \tag{3}$$

$$h_{it}^a = exp(X_{it}\gamma^a + U_{it}^a) \tag{4}$$

Substitute equations (3) and (4) into (1) and (2), the individual potential earnings in each sector $j$ and time $t$ can be represented as equations (5) and (6).

$$W_{it}^n = P_t^n A_t^n exp(X_{it}\gamma^n + U_{it}^n) \tag{5}$$

$$W_{it}^a = P_t^a A_t^a exp(X_{it}\gamma^a + U_{it}^a) \tag{6}$$

Taking logs, I can obtain equations (7) and (8), where the lower cases represent the logarithmic terms.

$$w_{it}^n = \underbrace{p_t^n + a_t^n}_{=\delta_t^n} + X_{it}\gamma^n + U_{it}^n \tag{7}$$

$$w_{it}^a = \underbrace{p_t^a + a_t^a}_{=\delta_t^a} + X_{it}\gamma^a + U_{it}^a \tag{8}$$

In equations (7) and (8), $p_t^j$ represents the price for each sector $j$ at time $t$, and $a_t^j$ stands for average productivity for sector $j$ at time $t$. Together, they represent the returns from the sector-wide productivity at time $t$ for sector $j$, denoted as $\delta_t^j$. Hence, rewrite potential earnings for each sector as equations (9) and (10).

$$w_{it}^n = \delta_t^n + X_{it}\gamma^n + U_{it}^n \tag{9}$$

$$w_{it}^a = \delta_t^a + X_{it}\gamma^a + U_{it}^a \tag{10}$$

## 3.2 Absolute and Comparative Advantages

Comparative advantage is the difference in an individual's abilities between sectors. It is the differential returns based on this comparative advantage that affect sector decisions for economic agents. This comparative advantage comprises two parts: observed characteristics, such as education, and unobserved abilities, such as detail-oriented attributes. The returns to the observed characteristics are seen in the data and collected in the vector of $X$'s. On the other hand, the unobserved comparative advantages, the differences in individual unobserved abilities between the two sectors, are not recorded in the data. However, individuals know their own latent skills and internalize this information when deciding which sector to enter. Thus, selection bias arises if this unobserved comparative advantage is not taken into account.

Therefore, I will introduce more structure to the unobserved error terms, $U_{it}^j$, in equations (9) and (10), reflecting the source of selection within the unobserved ability terms $U_{it}^j$. Following Lemieux (1998) and Suri (2011), decompose as in equation (11) and (12).

$$U_{it}^n = \theta_i^n + \xi_{it}^n \tag{11}$$

$$U_{it}^a = \theta_i^a + \xi_{it}^a \tag{12}$$

where $\theta_i^j$ is a time-invariant, sector-specific ability (permanent) and $\xi_{it}^j$ is an idiosyncratic shock, which is unknown at choice, but follows a zero conditional mean (transitory). Hence, sorting depends on the permanent vector $(\theta_i^n, \theta_i^a)$, not on $\xi_{it}^j$, shown in equation (13):

$$E(U_{it}^n - U_{it}^a) = \theta_i^n - \theta_i^a \tag{13}$$

Simply put, this transitory component of the error terms does not affect individuals' sector choices. Thus, when choosing a sector, the differences in individuals' sector-specific abilities will play a crucial role. Note that it is impossible to identify absolute advantages, $\theta_i^n$ and $\theta_i^a$, separately. Nor is it needed, as it is the difference in absolute advantages that matters. To separate abilities that influence sector choice from those that do not, linearly project $\theta_i^j$ on the "difference of latent abilities across sector", $\theta_i^n - \theta_i^a$, shown as equations (14) and (15):

$$\theta_i^n = b_n(\theta_i^n - \theta_i^a) + \tau_i \tag{14}$$

$$\theta_i^a = b_a(\theta_i^n - \theta_i^a) + \tau_i \tag{15}$$

where $\tau_i$ is a common component (orthogonal to $\theta_i^n - \theta_i^a$) that moves productivity in both sectors equally (e.g., work ethic) and thus does *not* affect sector choice. The coefficients $(b_n, b_a)$ are projection coefficients determined by the variance–covariance matrix of $(\theta_i^n, \theta_i^a)$.[1]

Define the individual's *comparative advantage* as equation (16)

$$\theta_i \equiv b_a(\theta_i^n - \theta_i^a), \tag{16}$$

and the *selection effect* parameter as equation (17)

$$\beta \equiv \frac{b_n}{b_a} - 1. \tag{17}$$

[1]Closed-form expressions: $b_n = \frac{\sigma_n^2 - \sigma_{na}}{\sigma_n^2 + \sigma_a^2 - 2\sigma_{na}}$ and $b_a = \frac{\sigma_{na} - \sigma_n^2}{\sigma_n^2 + \sigma_a^2 - 2\sigma_{na}}$, where $\sigma_n = Var(\theta_i^n)$, $\sigma_a = Var(\theta_i^a)$, and $\sigma_{na} = COV(\theta_i^n, \theta_i^a)$.

Then sector-specific absolution advantages can be expressed as a function of comparative advantage and selection effect, as shown in equations (18) and (19):

$$\theta_i^n = (1+\beta)\theta_i + \tau_i \tag{18}$$

$$\theta_i^a = \theta_i + \tau_i \tag{19}$$

Equations (18) and (19) are the key decompositions: $(\theta_i, \beta)$ describe the part of unobserved ability that *drives sorting* (comparative advantage and its sectoral loading), while $\tau_i$ is the sector-irrelevant component. This is precisely where TWFE regressions confound selection: they absorb *both* $\theta_i$ and $\tau_i$ into a single fixed effect, attributing common, sector-irrelevant traits to selection, such as hard work.

Hence, conditional on observables, sector choice is governed by the differential return to $\theta_i$ (the comparative advantage component), with loading $(1+\beta)$ in non-agriculture and 1 in agriculture. The common component $\tau_i$ cancels in the choice comparison and is irrelevant for sorting.

In this formulation, the structural parameter $\beta$ summarizes how strongly unobserved comparative advantage is rewarded *differentially* across sectors after netting out the correlation between sectoral abilities. Positive $\beta$ indicates that the nonagricultural sector loads more on the relevant ability dispersion than agriculture; negative $\beta$ indicates the opposite.

In the Appendix B, I formally map the selection effect $\beta$ to types of selection in the classic Roy model. When $\beta > 0$, individuals are positively selected—drawn from the upper tail in agriculture and landing in the upper tail of non-agriculture. When $-1 < \beta < 0$, the sorting is negative, as workers come from the lower tail in both sectors. When $\beta < -1$, a 'refugee' case arises: workers below average in agriculture sort into non-agriculture but earn above its mean. Finally, when $\beta = -1$, switchers earn exactly the sectoral mean in non-agriculture, a case absent in Borjas. Full algebra, the formal conditions, and the side-by-side comparison with Borjas's are in Appendix C.

Essentially, the selection effect and unobserved comparative advantages formulated by Lemieux (1998) capture the equivalent selection effect from unobserved heterogeneity that affects individuals' choices, as modelled in the classic Roy choice framework. The difference is that this formulation allows for the flexibility to estimate underlying unobserved comparative

advantages avoiding distributional assumption.

### 3.3 Main Empirical Model

After the discussion on $\beta$, let's go back to the empirical model used in this paper. Building on equations (18)-(19), which decompose latent ability into comparative advantage and selection effect, I substitute into the potential earnings framework (equations (9) and (10)). I can rewrite the individual's log potential earnings at time $t$ for each sector $j$ in equations (20) and (21).

$$w_{it}^n = \delta_t^n + (1+\beta)\theta_i + \tau_i + X_{it}\gamma^n + \xi_{it}^n \tag{20}$$

$$w_{it}^a = \delta_t^a + \theta_i + \tau_i + X_{it}\gamma^a + \xi_{it}^a \tag{21}$$

Let $D_{it}$ be a dummy variable, taking the value one if an individual $i$ chooses the primary job in the nonagricultural sector at time $t$ and zero otherwise. Then, I can write the individual's log earnings as equation (22).

$$w_{it} = D_{it}w_{it}^n + (1 - D_{it})w_{it}^a \tag{22}$$

where

$$D_{it} = \begin{cases} 1 & \text{non-agricultural sector} \\ 0 & \text{agricultural sector} \end{cases}$$

Then, substitute equations (20) and (21) in (22) to obtain equation (23).

$$\begin{aligned} w_{it} = {} & \delta_t^a + (\delta_t^n - \delta_t^a)D_{it} \\ & + \theta_i + \beta\theta_i D_{it} + X_{it}\gamma^a + X_{it}(\gamma^n - \gamma^a)D_{it} + \tau_i + \epsilon_{it} \end{aligned} \tag{23}$$

where

$$\epsilon_{it} = D_{it}\xi_{it}^n + (1 - D_{it})\xi_{it}^a \tag{24}$$

Equation (23) is the main empirical model in this paper. I am interested in estimating the parameter $\beta$, the selection effect of comparative advantages, and recovering the distribution of comparative advantages $\theta_i$. As the fourth term's coefficient ($\beta\theta_i$) in equation (23) contains the unobserved random variable $\theta_i$; moreover, $\theta_i$ is correlated to the sectoral choices $D_{it}$,

hence, this is a correlated random coefficient (CRC) model. In this formulation, individual comparative advantage ($\theta_i$) is explicitly defined as the individual deviation from the average sector productivity. If an individual works in the agricultural sector, sector-wide productivity is $\delta_t^a$, and each individual's comparative advantage, $\theta_i$, expresses how each individual is more productive compared to the average productivity in the sector. If the nonagricultural sector is chosen, the sector-wide productivity is represented by $\delta_t^a + (\delta_t^n - \delta_t^a)$, and each individual's productivity deviation from the sector mean is $(1 + \beta)\theta_i$.

By assuming time-invariant sector-specific unobserved abilities, this formulation combines Leumieux's decomposition of absolute advantages. It allows for the modelling of comparative advantages as the difference between absolute advantages across sectors, scaled by the covariance-adjusted spread, while distinguishing which unobserved abilities are relevant to sector choices. This empirical framework effectively addresses the two core challenges identified in the Literature Review section— heterogeneity and endogeneity. Moreover, this formulation of comparative advantages, aligning with Roy's model choice setting, is better equipped to estimate individual sorting than a TWFE estimator or one that controls for fixed effects, as reviewed in the previous section, and allows for the estimation of selection effects without imposing functional assumptions on latent abilities.

## 3.4 Individual Sorting and Agricultural Productivity Gap (APG)

The objective of the main empirical model, as described in equation (23), is to estimate the structural parameter, the selection effect $\beta$, without imposing any distributional assumptions on latent skills. In the estimation section, I will describe how to obtain $\beta$ without distributional assumptions. Now, suppose $\beta$ is estimated from the data. What $\beta$ measures is the extent to which individual sorting based on comparative advantages affects their earnings. Therefore, an aggregation is necessary to provide an answer to the research question posed in this paper: how much individual sorting affects sectoral productivity gaps?

Consider an economy with two sectors: agriculture and non-agriculture, both of which

are perfectly competitive. The output in each sector is given by equations (25) and (26).

$$Y_n = A_n H_n \tag{25}$$

$$Y_a = A_a H_a \tag{26}$$

where $Y_j$ represents the sector aggregate output, $A_j$ is the sector-specific efficiency, $H_j$ is the efficient labour in each sector. Furthermore, the efficient labour $H_j$ is the product of sector-specific human capital, $h_j$, and the total number of workers $L_j$ in each sector $j$, represented by equations (27) and (28).

$$H_n = h_n L_n \tag{27}$$

$$H_a = h_a L_a \tag{28}$$

Following the agricultural productivity gap (APG) defined by Gollin et al. (2014), under the assumption of a perfectly competitive labour and goods market in both sectors, the wage in each sector equals the marginal value product of labour, and it also equals the average value of output per labour in each sector at equilibrium. Let $W_j$ be the wage in the sector $j$, and $P_j$ be the price of good $j$. In equilibrium, APG–the ratio of value-added (VA) per worker between sectors $j$—can be expressed as the wage ratio at the sector level. $\frac{W_n}{W_a}$. See equation (29).

$$\underbrace{\frac{P_n Y_n / L_n}{P_a Y_a / L_a}}_{= \frac{VA_n/L_n}{VA_a/L_a} \equiv APG} = \frac{W_n}{W_a} \tag{29}$$

In the data, the difference in log earnings between sectors is APG, expressed in equation (30). This APG can be directly calculated by differencing the mean log earnings between two sectors, using the data.

$$log\left(\frac{W_n}{W_a}\right) = log(W_n) - log(W_a) = APG \tag{30}$$

Furthermore, the APG can be decomposed into three components, shown in equation (31): (1) APG from observed characteristics, $\overline{X}(\hat{\gamma^n} - \hat{\gamma^a})$; (2) APG from sector-wide productivity gap, $\delta^n - \delta^a$; (3) APG from individual sorting based on the unobserved comparative

advantages, denoted as $S_\theta$, see equation (32).

$$APG = \underbrace{\overline{X}(\hat{\gamma^n} - \hat{\gamma^a})}_{APG_{observed}} + \underbrace{\delta^n - \delta^a}_{APG_\delta} + S_\theta \tag{31}$$

$$S_\theta = \underbrace{\beta E[\theta_i|D=1]}_{\text{extra returns in nonag}} + \underbrace{(E[\theta_i|D=1] - E[\theta_i|D=0])}_{\text{mean diff in comparative advantages}} \tag{32}$$

$$S_\theta = APG - APG_{observed} - APG_\delta \tag{33}$$

Since $\theta_i$ is unobserved, it may not be possible to attain the mean of $\theta_i$. Equation (33) provides an alternative way to get the selection effect component in APG. As a result, the impact of individual sorting based on comparative advantages is the share of APG from the unobserved component, $\frac{S_\theta}{APG}$.

To recap, the main empirical model in this section explicitly models individual unobserved comparative advantage $\theta_i$ as a deviation from the average productivity for each sector, addressing unobserved heterogeneity. Moreover, it formulates the absolute and comparative advantages to allow for the estimation of the selection effect $\beta$ without imposing a functional form on latent abilities. In addition, the types of selection in this main model can be mapped to the selection types in the classic Roy model framework.

I aim first to recover the structural parameters $\beta$ and $\theta_i$, and then aggregate to assess the importance of individual selection based on the observed APG. Before getting into estimation strategies, I will first discuss the identification of this CRC model.

# 4 Identification

A strict exogeneity of the composite error term, expressed in equation (34), delivers the identification for the main estimation equation (23).

$$E(\tau_i + \epsilon_{it}|\theta_i, D_{i1}...D_{iT}, X_{i1}...X_{iT}) = 0 \tag{34}$$

As $\tau_i$ represents individual $i$'s unobserved abilities, regardless of sector choices, this strict exogeneity assumption is not overly restrictive for $\tau_i$. Mathematically, equations (18) and (19) indicate that $\tau_i$ is removed from the comparative advantages, $\theta_i$; hence, $\tau_i$ does not affect the sectoral choices and other observed regressors in (23).

The concern mainly resides in the transitory error term, $\epsilon_{it}$. Under the assumption of the mean independent transitory error term ($\epsilon_{it}$), equations (20) and (21) indicate that the sector-specific transitory shocks $\xi_{it}^n$ and $\xi_{it}^a$ do not affect the individual's sector choices. I am now examining whether such an assumption is plausible in this setting. Importantly, the sectoral choices occur before most transitory shocks hit the relevant sector. Assuming that individuals are risk-neutral, in expectation, a risk-neutral agent does not consider the unobserved sector-specific transitory shocks when choosing a sector, i.e. $E(\epsilon_{it}|D_{it}) = 0$.

However, some transitory shocks could impact individuals' realized earnings, such as crop failure due to extreme weather conditions, which could raise concerns about the identification. This concern arises because individuals are likely to take measures to maintain their normal income level during adverse shocks, such as switching sectors or increasing the number of hours worked. Notably, individuals selected for their primary sector based on comparative advantage are unlikely to respond to short-lived, sector-specific shocks by changing sectors. This tendency to remain in the original sector is plausible for two reasons. First, the shocks are transitory and do not justify abandoning a productivity-maximizing allocation. Second, without retraining or skill upgrading, individuals are unlikely to achieve higher earnings in an alternative sector that lacks a comparative advantage.

Empirical evidence from the Indonesia Family Life Survey (IFLS) supports this reasoning. In the first wave, respondents were asked whether they had experienced major economic shocks in the previous five years—including events such as crop failure, business loss, household member illness, and income loss (Figure 1)—and how they coped with them (Figure 2). The most frequently reported coping strategies were taking on additional jobs, borrowing from or receiving transfers from relatives, using savings, and reducing expenditures. Importantly, switching sectors does not appear among the recorded responses, reinforcing the notion that individuals tend to remain in their chosen sector and absorb short-term shocks through secondary adjustments rather than by changing their primary occupation.

If individuals choose to change their hours worked in response to temporary shocks, that could raise concerns about identification because hours worked is one of the regressors in the estimation equation, included in $X_{it}$. As the IFLS survey provides information on major transitory shocks that could affect potential earnings at the household level, it is plausible to

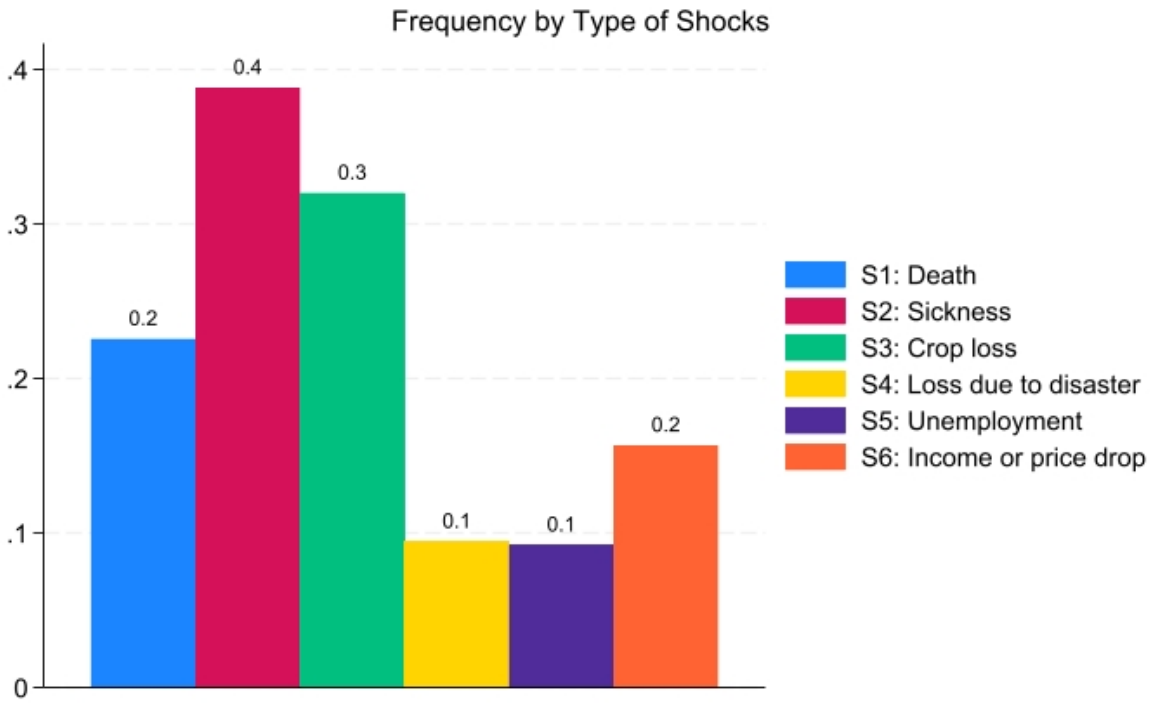


Figure 1

assume strict exogeneity of $\epsilon_{it}$ in equation (34) after the control for such transitory shocks. Regarding measures to cope with shocks, this information was only collected in the first round of surveys and is not available for subsequent waves. However, the information in the first wave provides clear empirical evidence that transitory shocks do not affect sector switching in the IFLS dataset. Hence, the exogenous shocks that cause individuals to switch sectors must be cost-related, such as making the pursuit of the alternative sector cheaper. For instance, as more factories are established in villages, the cost of securing a stable wage in the manufacturing sector is substantially reduced, and some people may be persuaded to switch sectors.

In sum, strict exogeneity of the composite error term is justified here on two grounds: (i) $\tau_i$ is orthogonal to choices and regressors by construction, and (ii) $\epsilon_{it}$ is plausibly mean independent of choices once household shocks are controlled for. Both the theoretical structure of the model and empirical evidence from IFLS support the claim that individuals do not switch sectors in response to transitory shocks, ensuring that equation (34) delivers valid identification.

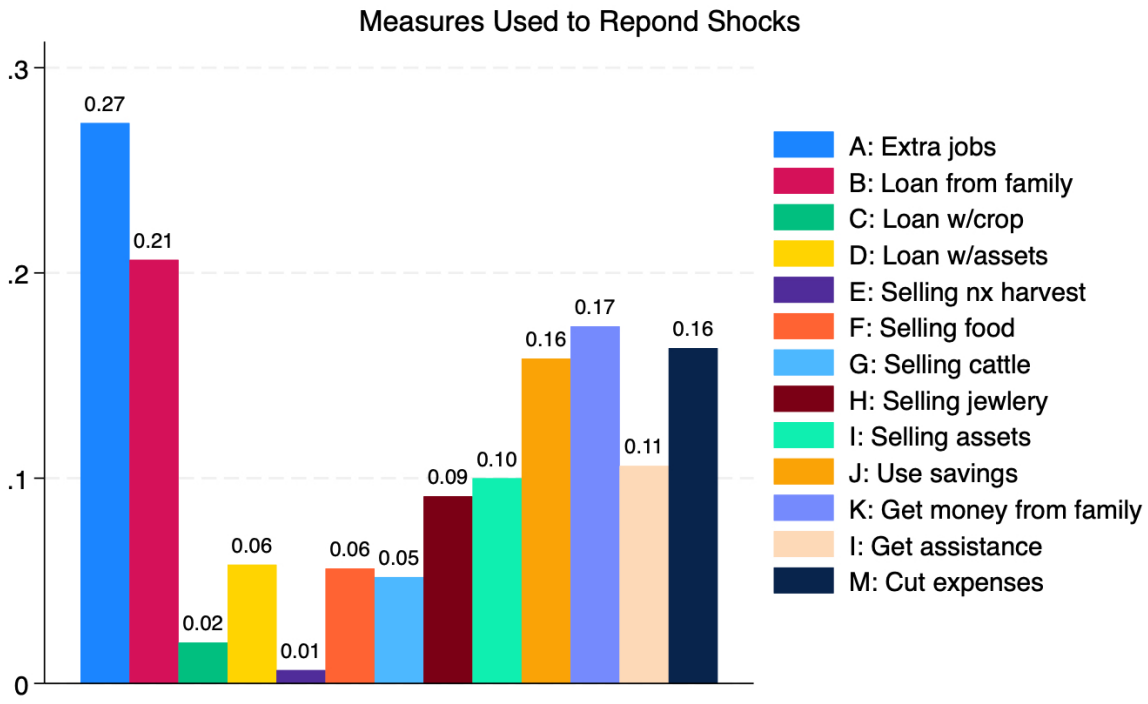


Figure 2

Because $\theta_i$ is unobserved, recovering it requires exploiting individuals' sectoral choice histories. Following the projection approach of Chamberlain (1982) and Suri (2011), I use observed choices over time to infer comparative advantage without imposing functional forms on unobserved heterogeneity. The central idea is that choices reveal which sector yields higher returns for each individual, thereby uncovering the latent abilities $\theta_i$ that govern sectoral allocation. In practice, substituting the choice trajectory into the model yields reduced-form regressions whose coefficients can be written as functions of the underlying structural parameters. Appendix E illustrates this mapping in a simple two-period case, showing explicitly how reduced-form coefficients recover $\theta_i$. By doing so, I estimate comparative advantage without assuming any specific distribution for latent abilities—unlike Roy-type models, which hinge on joint normality. This revealed comparative advantage interpretation is what connects the structural framework to observable data. Appendix F further demonstrates how these recovered $\theta_i$ values capture economically meaningful variation in individual sorting.

Together, this section shows that the selection effect $\beta$ is identified under the plausibility of the strict exogeneity assumption in 1990s Indonesia; moreover, $\theta_i$ is predicted from indi-

vidual choice trajectories, which I refer to as revealed unobserved comparative advantages.

The empirical framework developed in Sections 3 and 4 combines several features: it models comparative advantage explicitly, defines it as individual deviation from sector averages while allowing it to drive sectoral choice, and recovers structural parameters without imposing distributional assumptions on latent heterogeneity.

Relative to the existing literature, this framework contributes by (i) redefining absolute advantage in line with the APG literature, (ii) linking the selection effect $\beta$ to Borjas's (1987) Roy-model selection types, (iii) introducing the notion of revealed comparative advantage to add economic interpretation to the projection method, and (iv) decomposing the APG into observed, productivity, and sorting components.

The next section applies this empirical model to data from the first three waves of the IFLS.

# 5 Estimate Selection on APG

In the previous section, I have described the empirical methodology based on Suri's (2011) approach, extending it to measure the impact of individual sorting on sectoral productivity gaps. This section will implement this empirical model on the first three waves of the IFLS survey data (1993-2000). To implement Suri's method, I use the STATA package, **randcoef**, developed by Cabanillas, Michler, Michuda and Tjernström (2018). To implement **randcoef**, it is required to install the **tuple** package first. STATA provides download links to both packages via Stata Community-Contributed programs.

In this section, I will first illustrate descriptive statistics for the first three waves of the IFLS survey. Then, I will show the reduced-form estimation for each wave, using Seemingly Unrelated Regression (SUR) for the first stage. Next, I will present the results at the second stage, which involves recovering the structure parameters of interest, including the selection effect $\beta$, and the distribution of comparative advantage $\theta_i$. The first and second stage estimations are both accomplished via **randcoef** in STATA. Finally, I will calculate the impact of individual sorting on APG.

## 5.1 Descriptive Statistics

The Indonesia Family Life Survey (IFLS) is a rich, longitudinal household survey that spans five waves, conducted in 1993, 1997, 2000, 2007, and 2014, representing approximately 80% of the Indonesian population. The survey covers 13 out of the then 27 provinces in Indonesia in 1993, selected to reflect the country's socioeconomic status (Strauss et al., 2016). During this period, Indonesia experienced a significant economic transformation, transitioning from a low-income to a lower-middle-income country. According to World Bank estimates (2025), Indonesian real GDP per capita in constant 2015 U.S. dollars rose from $1,693 in 1993 to $3,171 in 2014, going through the Asian financial crisis to a resilient post-crisis recovery and then steady economic growth. The IFLS provides detailed, repeated observations on individuals' employment, sectoral choices, earnings, and household information, making it particularly well-suited for investigating the impact of individual sorting on sectoral productivity gaps.

During the IFLS study period (1993-2014), Indonesia went through a significant political and institutional transformation. Indonesia's authoritarian New Order collapsed with the ouster of President Suharto on 21 May 1998, initiating the Reformasi transition to democratic rule (Freedom House, 1998). In this transition, Parliament passed Laws 22/1999 and 25/1999, mandating a sudden, wide transfer of authority and revenues to the regional governments, effective 2001, which is known as "Big Bang" decentralization (Hofman and Kaiser, 2002). Given this sudden structural break in the institutions and its significant economic implications, this research project naturally breaks the IFLS survey into two separate periods, 1993-2000 and 2000-2014, divided by the 2001 "Big Bang" decentralization. This paper examines the first period, and a subsequent paper will further investigate the later waves of the IFLS survey.

To study APG needs the data on earnings; moreover, unobserved comparative advantage utilizes information from individuals' sectoral choices. Therefore, when implementing the empirical approach described in the previous section, I only include individuals with information on both earnings and sector choice in each period. Hence, the panel used in this analysis comprises 4,615 individuals across three waves, totalling 13,845 individual-wave

observations.

Tables 3 and 4 report summary statistics at the individual level. On average, individuals in the balanced panel were 40.4 years old in 1993, with mean age increasing steadily across waves as expected. Roughly 44% of the people resided in urban areas, a proportion that remained stable across survey rounds. This relatively stable rural-urban location likely reflects the nature of this balanced sample, which comprises individuals with complete earnings and sector information across all three waves. Those individuals with stable incomes are less likely to migrate between urban and rural areas than the broader population. The sample also exhibits a disproportionately male population, with 73% of individuals being male. This gender imbalance does not accurately reflect the gender composition of the full IFLS sample, but instead arises from sample restrictions: many women in the broader population work without pay in family enterprises or are out of the labour force as homemakers, and thus are excluded from the earnings-based analysis.

The share of individuals in non-agricultural work is relatively stable, declining slightly from 65.5% in 1993 to 63.0% in 2000. Approximately 45% to 40% report waged employment across waves, implying a gradual rise in informal or self-employment within this sample - from about 54% in 1993 to 60% in 2000. The average monthly nominal income increased substantially, from IDR 126,195 in 1993 to IDR 432,583 in 2000, reflecting both real income growth and the effects of high inflation surrounding the Asian financial crisis during this study period. This is corroborated by the rise in the Consumer Price Index (CPI), which climbed from 145.2 to 209.4 over the same period. Meanwhile, the average number of hours worked per month remained stable, hovering around 174 hours, with a slight decrease in working hours over time.

Table 5 reports household-level characteristics for the same balanced sample. Average household size increased modestly over time, from 4.78 members in 1993 to 5.80 in 2000. Family business ownership rose sharply—from 68% in the first wave to 80% by the third period. The share of households operating farm businesses remained relatively stable (rising from about 43% to 48%), while the share engaged in non-farm businesses increased significantly, from 37% to 52% over the three periods. Nominal asset values in both categories rose over time: farm business assets grew from IDR 7.1 million in 1993 to IDR 26.6 million

in 2000, while non-farm business assets increased from IDR 6.4 million to IDR 9.5 million. Notably, household assets not tied to any business, primarily in the form of real estate, were substantially higher in value, rising from IDR 11.4 million to IDR 32.1 million over time. This reference to the property values suggests that most family businesses are relatively small in scale compared to households' overall wealth holdings.

The IFLS also provides detailed information on economic shocks experienced by households in the five years preceding each survey. In 1993, 31.2% of households reported at least one shock, with an average of 0.39 shocks per household. This economic shock rose to 40% of households, with an average of 0.57 shocks reported in 1997, coinciding with the onset of the Asian financial crisis. By 2000, both the incidence (35%) and intensity (0.44 shocks on average) of reported shocks had declined.

The descriptive figures further illuminate sectoral choices and income dynamics in the dataset. Figure 3 displays the distribution of sectoral transition patterns across the three survey rounds. A majority of individuals exhibit persistent sectoral attachment: 55% remain in the non-agriculture sector across all waves (Nonag–Nonag–Nonag), and 25% stay continuously in agriculture. In contrast, sector switchers constitute a relatively small share of the sample, with each specific transition path accounting for no greater than 5%. This high degree of persistence suggests that individuals make their sectoral choices early on, likely based on their comparative advantages, and tend to remain in those sectors over time. This pattern lends empirical support to the assumption that unobserved sector-specific abilities—key to sorting—are time-invariant over the study period. Moreover, it provides empirical evidence that individuals are unlikely to switch sectors in response to transitory shocks. Combining the information on the economic shocks collected from each wave, these sectoral transition patterns provide additional support to the model identification illustrated in the previous section.

Complementing the evidence on sectoral persistence, Figure 6 presents the share of individuals engaged in non-agricultural employment by urban and rural location. Participation in non-agriculture consistently exceeds 85% in urban areas, compared to only 45–50% in rural areas. These spatial differences are remarkably stable across survey rounds, indicating a strong geographic pattern in sectoral employment. In addition, the high prevalence of non-

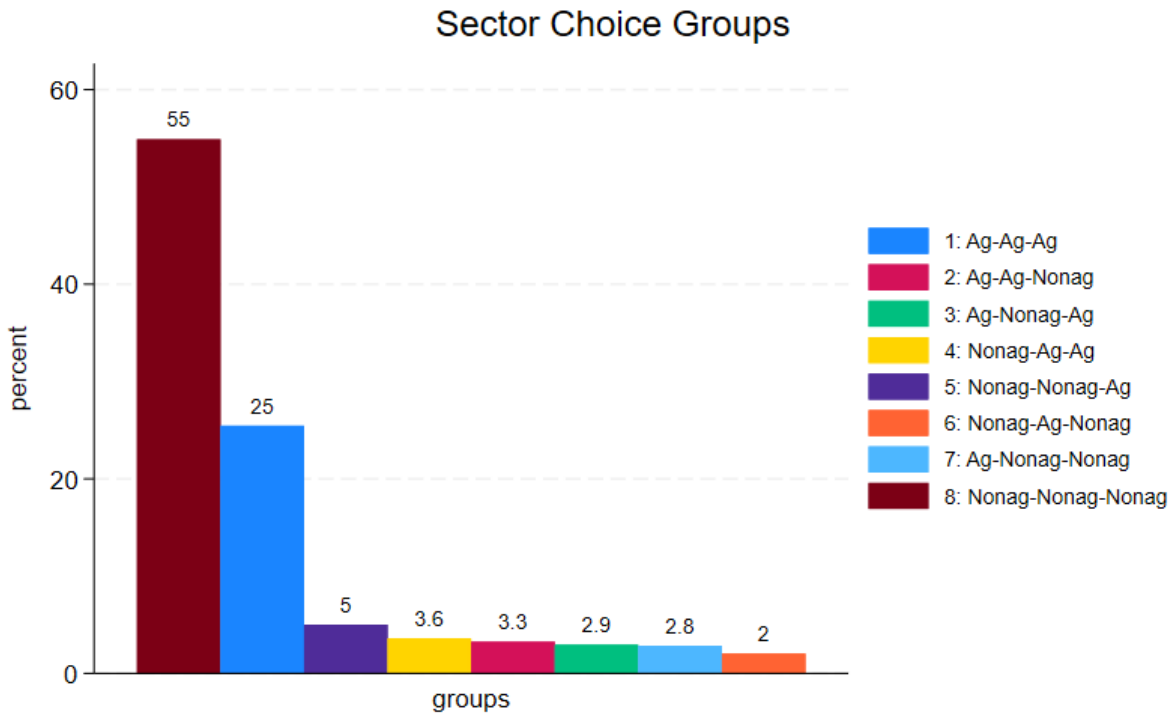


Figure 3: Choice Trajectories over Three Waves

agricultural work, even in rural areas, suggests that individuals can switch sectors without relocating. This sizable share of non-agriculture employment in rural areas challenges the common strategy of using rural–urban migration as a proxy for sectoral transitions in the Indonesian context.

Figure 7 displays the distribution of monthly income (in logarithmic terms) by wave and sector. Across the three rounds, individuals in the non-agricultural sector consistently earn more on average than those in the agricultural sector. In agriculture, outliers tend to be on the lower side in every wave: many observations fall well below the lower whisker, including a handful near zero, while high-side outliers are sparse. On the other hand, the non-agricultural sector has relatively few low-side outliers, with a recurrent upper tail of high earners protruding above the upper whisker in each wave. This figure also reflects substantial within-sector earnings heterogeneity, which further corroborates the summary statistics in Table 3, showing that the standard deviation of earnings exceeds the mean in each wave. The distributions in Figure 7 confirm that the dispersions in the earnings are substantial throughout the study period.

Figures 7 and 8 illustrate the substantial and persistent earnings differences between the agricultural and non-agricultural sectors over time. In Figure 7, the distributions of log monthly earnings differ markedly by sector and remain consistently distinct across survey waves. Figure 8 further shows that raw income ratios between non-agriculture and agriculture, i.e., the observed APG, remain sizable but exhibit a declining trend over time.

In urban areas, the ratio of average earnings between the non-agricultural and agricultural sectors was approximately 1.83 in 1993, peaked at 2.16 in 1997 during the Asian Financial Crisis, and declined to 1.34 by 2000. In rural areas, the corresponding ratios were 1.95, 1.66, and 1.70, respectively. Notably, the sectoral productivity gap is consistently larger in rural than in urban areas before and after the crisis. However, this pattern temporarily reversed in 1997, when the urban APG sharply exceeded that in the rural areas—likely reflecting the differential impact of the crisis on agricultural versus non-agricultural earnings during the Asian financial crisis. Overall, the sectoral productivity gap remains large throughout the period but exhibits a clear downward trend, suggesting convergence in sectoral earnings and raising important questions about the evolving role of sector-specific productivity and selection.

Figure 9 presents the average log earnings across sectors between formal and informal workers, where formal workers refer to waged workers in this context. In both sectors, formal workers, on average, earn higher than informal ones, while non-agricultural workers enjoy much higher earnings than those in agriculture.

Together, these descriptive statistics highlight five key empirical patterns in the balanced panel from the first three IFLS waves: (1) roughly 80% of workers remain in their initial sector, with 55% staying in non-agriculture and 25% in agriculture, while only 20% switch sectors; (2) income gaps between sectors are large but decline over time, with distinct heterogeneous patterns in urban and rural areas; (3) earnings dispersion is substantial in both sectors, skewed toward low outliers in agriculture and high earners in non-agriculture; (4) sectoral composition differs sharply between urban and rural areas, with little change across waves; and (5) informal workers earn markedly less than formal workers, a pattern that persists both within and across sectors.

These regularities underscore the empirical motivation for modelling sectoral sorting

through time-invariant unobserved comparative advantages, while also highlighting the need to account for geographic location and macroeconomic conditions in explaining earnings gaps. In sum, the descriptive evidence is consistent with the core assumptions of the identification strategy. The following subsection turns to preliminary estimation results and the recovery of the structural parameters of interest.

## 5.2 Estimates Reduced-Form Coefficients

The reduced-form specification corresponds to equations (60) and (61) in Section 4.2, which illustrate the estimation procedure using a simplified two-period model without covariates. In the actual implementation presented here, the first-stage regression extends this procedure to a three-period panel and controls for a suite of observed characteristics. The dependent variable is log earnings from the individual's primary job at each wave.

Tables 6, 7, and 8 report the estimation results from the first stage, corresponding to the 1993, 1997, and 2000 survey rounds, respectively. Each table presents four model specifications: Column (1) includes sectoral choices and their interactions across waves, without any covariates. Column (2) adds the controls for locations: urban/rural residence and province. Column (3) further controls for hours worked, waged work, age, gender, marital status, religion, and education in addition to the model in the preceding specification. Column (4) adds two more variables: the log of province-level CPI (collected from the Indonesian Statistical Bureau (Indonesa, 2019)) and a binary indicator for whether the household experienced any economic shock in the past five years. As the CPI is at the provincial level, the categorical variable, province, is dropped in the complete specification.

Coefficients are in the first row for each regressor, and the corresponding standard errors are beneath them. A double-asterisk, **, denotes statistical significance at the 5% level, and a single-asterisk, *, indicates the 10% level. The end of each column reports the number of observations ($N$) and the R-squared value ($R^2$) for each regression. Five key patterns emerge from the reduced-form regressions:

1. Initial sectoral choice as a persistent predictor: The sector of employment in the first wave significantly influences earnings across all periods. In contrast, sectoral choices

in later waves exhibit little effect. This persistence supports the assumption of time-invariant unobserved comparative advantage, consistent with the observed pattern of a relatively high tendency to stay in the initially chosen sector. Additionally, the contemporaneous sector choice indicator also exhibits significant explanatory power for earnings in the respective period.

2. The impact of urban-rural residence is more important than provincial locations in terms of explaining earnings. Column (2) adds two spatial controls: urban and province. Urban is a dummy variable with a value of 1 if an individual resides in an urban area. The province is a categorical variable, with 16 provinces: 13 initial ones from the survey and three additional provinces due to households' relocation. The share of families that have moved across provinces is less than 0.1% in the balance sample. Once control for urban-rural locations, the provinces do not have significant explanatory power for earnings.

3. Urban-rural location in the initial period has significant explanatory power for earnings, but this explanatory power diminishes in the third period. This substantial impact of initial location is consistent with the low observed urban-rural mobility in the balanced sample, where the proportions are as follows: 1,877 always-urban, 2,453 always-rural, and 285 switchers.

4. Hours worked exhibit a strong positive correlation with the earnings in the same period. Moreover, the dummy variable "wagedwork" captures the types of work for each individual, with value 1 for wage-paid jobs and 0 for self-employment. The types of employment have a significant impact on the earnings, and the formal employment pays higher, which aligns with the descriptive graph shown in the previous section.

5. CPI fluctuations align with economic shocks: The inclusion of log province-level CPI captures geographic price variation and some time-varying transitory shocks. After controlling for CPI, the economic shocks are not significant in the first period; however, they exhibit more explanatory power for earnings in the subsequent waves, suggesting that CPI may also absorb some of the economic shocks.

Finally, the covariates that commonly explain earnings behave as expected: age and education are consistently strong predictors of earnings. These results establish a credible reduced-form foundation for identifying structural parameters in the second-stage estimation. In addition, the point estimates are stable across the four specifications as the controls gradually increase. Next, I will show the results of recovered structural parameters.

## 5.3 Recovering Structural Parameters

The second-stage estimation recovers structural parameters by solving a system of linear equations corresponding to the estimated reduced-form coefficients. Specifically, after controlling for the observables, the coefficients of sector choices and their interactions contain information about the underlying structural parameters. This section retrieves two key parameters of interest: the selection effect ($\beta$) and the distribution of comparative advantage ($\theta_i$), utilizing the **randcoef** package in STATA. The complete mathematical formulation of the second-stage estimation refers to the paper by Cabanillas, Michler, Michuda, and Tjernström (2018), who developed this STATA package.

Table 1 presents structural parameters $\lambda_1$–$\lambda_7$, $\alpha$ for sector-wide productivity gap, and $\beta$, selection effect. Each column in Table 1 corresponds to the four specifications in the first-stage estimation, presented in Tables 6, 7, and 8: Column (1) excludes all covariates; Column (2) adds urban-rural residence and province; Column (3) further includes log hours worked, waged work, age, gender, education, religion and marital status; and Column (4) adds province-level log CPI and household-level economic shocks. The structural parameter estimates exhibit three key insights:

1. The selection effect ($\beta$) is positive but *not* statistically significant. The point estimate of $\beta$ captures the extent to which sectoral income differentials stem from the individual sorting based on unobserved comparative advantages. Its values increase from 0.154 log points in the baseline specification to 0.39 under the full controls. In all the cases, $\beta > 0$ suggests positive selection: individuals with a stronger comparative advantage in non-agriculture are more likely to enter that sector. Moreover, the people who choose the non-agricultural jobs are those who are better farmers, i.e. earning higher than the

average farmers. However, the standard errors are substantially large, ranging from 0.368 to 0.613, thereby rendering the estimates statistically insignificant.

This insignificant result aligns with the substantial variation in income observed in the data. As shown in Table 3, the pooled average monthly income from the primary job is IDR 268,329, while the standard deviation is IDR 1,711,989, more than six times larger than the mean. Moreover, this pattern of widespread monthly earnings is persistent in each wave. This high dispersion in earnings likely implies high dispersion in latent abilities. Such dispersion means that even if selection is strong at the individual level for some, it may not translate into a substantial sectoral wage gap after considering the distribution of comparative advantages. Here is where the distribution assumption matters to the estimation results.

This finding highlights the critical role of the underlying distribution of comparative advantages. A more concentrated distribution could yield a more significant aggregate selection effect, whereas a widely dispersed distribution, as observed here, dilutes its statistical significance. Hence, distributional assumptions on unobserved comparative advantages raise concerns for the empirical estimation of the selection effect, even though it can offer profound theoretical insights.

2. The estimated distribution of comparative advantages ($\theta_i$) deviates substantially from normality. As specified in this model, Equation (35) expresses the unobserved comparative advantages $\theta_i$ in a three-period model. After obtaining estimates of $\lambda_1$–$\lambda_7$, normalize $\theta_i$ such that $\sum \theta_i = 0$. Then, I can obtain $\lambda_0$ by calculating the intercept using Equation (36).

$$\begin{aligned}\theta_i &= \lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_3 D_{i3} + \lambda_4 D_{i1} D_{i2} \\ &\quad + \lambda_5 D_{i1} D_{i3} + \lambda_6 D_{i2} D_{i3} + \lambda_7 D_{i1} D_{i2} D_{i3} + \nu_i \qquad (35)\\ \lambda_0 &= -\lambda_1 \overline{D_{i1}} - \lambda_2 \overline{D_{i2}} - \lambda_3 \overline{D_{i3}} \\ &\quad - \lambda_4 \overline{D_{i1} D_{i2}} - \lambda_5 \overline{D_{i1} D_{i3}} - \lambda_6 \overline{D_{i2} D_{i3}} - \lambda_7 \overline{D_{i1} D_{i2} D_{i3}} \qquad (36)\end{aligned}$$

Once all the $\lambda$'s are available, using linear prediction recovers the empirical distribution of revealed comparative advantages $\hat{\theta}_i$. Figure 4 displays the estimated distribution for

specification (4) in Table 1, where the red line is the empirical distribution using the default Epanechnikov kernel bandwidth in STATA, determined by the sample standard deviation, inter-quartile range, and sample size (StataCorp, 2025). The STATA default bandwidth for $\hat{\theta}_i \approx 0.0199$, represented by the red line. The various kernel bandwidths provide references about how the distribution shape changes with the bandwidths. The key takeaway of Figure 4 is that the empirical distribution of revealed comparative advantages $\hat{\theta}_i$ is not close to normality.

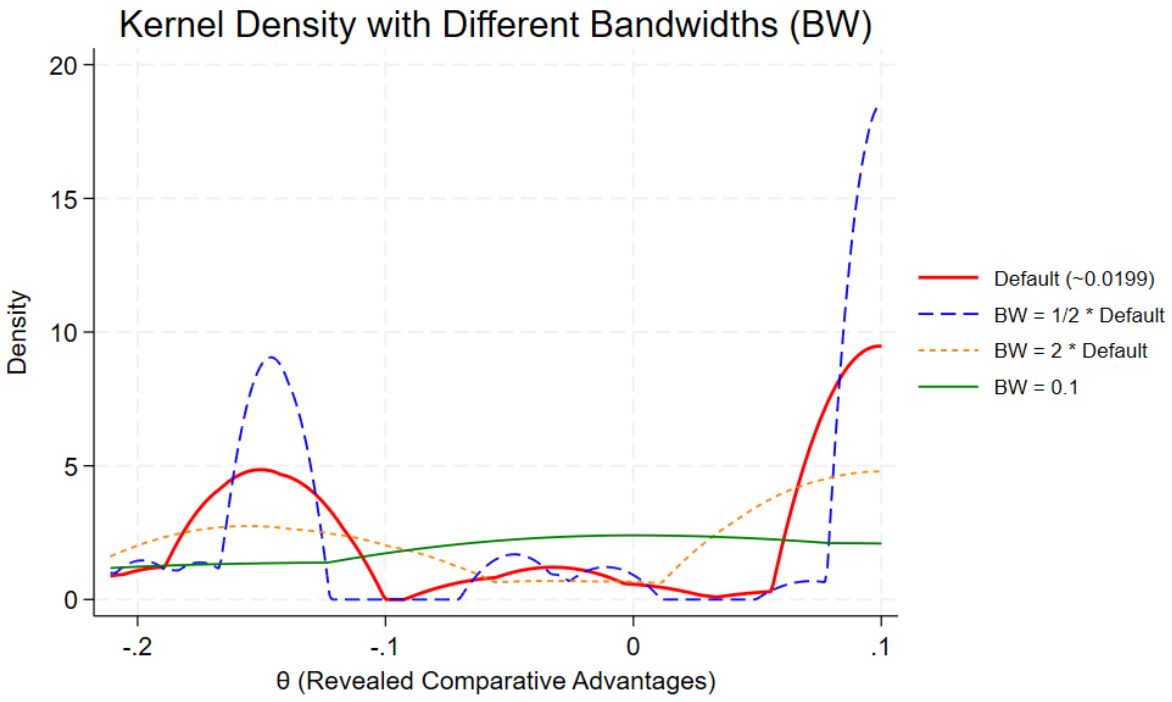


Figure 4: Recovered Distribution of Comparative Advantage for Specification (4)

This finding on the empirical distribution of comparative advantages aligns with extensive studies in the labour economics literature that question the conventional assumption of log-normality in the Roy model selection framework (Heckman and Honore, 1990). The two peaks located on both tails and the broad dispersion of $\theta_i$ call into question the empirical estimation approach that relies on strong parametric assumptions about unobserved heterogeneity in the Indonesian context in this study period.

Importantly, for individual selection to meaningfully influence sectoral productivity

gaps, a sufficiently large share of the population must sort into sectors based on comparative advantages in the same direction, i.e the same sign in $\beta$. Without such alignment, even substantial individual-level sorting may fail to generate aggregate effects large enough to shift the observed APG.

3. Sectoral productivity differences ($\alpha$) explain a substantial and statistically significant share of the observed earnings gap. The estimated sectoral premium in the baseline specification without controls is 0.509 log points, which declines slightly to 0.42 in the complete specification with all the controls. Importantly, as the average values of $\alpha$ decrease, their associated standard errors also contract, enhancing the precision and statistical significance of the estimates. As shown in Table 1, the estimated $\alpha$ values across columns (1) to (4) are 0.509, 0.491, 0.413, and 0.420, with corresponding standard errors of 0.090, 0.082, 0.066, and 0.061. This consistent precision suggests that sector-wide productivity differences remain a robust and one of the primary factors in explaining the APG, even after controlling for a rich set of covariates.

Table 1: Structural Parameters

| **Structural Parameters** | **(1)** | **(2)** | **(3)** | **(4)** |
|---|---|---|---|---|
| $\lambda_1$ | 0.222 ** | 0.135 ** | 0.127 ** | 0.134 ** |
| | (0.063) | (0.063) | (0.058) | (0.056) |
| $\lambda_2$ | -0.037 | -0.063 | -0.097 | -0.066 |
| | (0.072) | (0.071) | (0.067) | (0.064) |
| $\lambda_3$ | -0.055 | -0.073 | -0.054 | -0.042 |
| | (0.063) | (0.062) | (0.054) | (0.053) |
| $\lambda_4$ | -0.031 | -0.009 | 0.028 | 0.029 |
| | (0.093) | (0.091) | (0.081) | (0.077) |
| $\lambda_5$ | 0.308 ** | 0.295 ** | 0.155 | 0.125 |
| | (0.124) | (0.124) | (0.099) | (0.093) |
| $\lambda_6$ | 0.193 * | 0.185 * | 0.118 | 0.092 |
| | (0.106) | (0.106) | (0.091) | (0.085) |
| $\lambda_7$ | 0.015 | -0.063 | -0.024 | -0.026 |
| | (0.153) | (0.148) | (0.127) | (0.112) |
| $\alpha$ | 0.509 ** | 0.491 ** | 0.413 ** | 0.420 ** |
| | (0.090) | (0.082) | (0.066) | (0.061) |
| $\beta$ | 0.154 | 0.119 | 0.180 | 0.390 |
| | (0.368) | (0.444) | (0.508) | (0.613) |

*Notes:* Numbers in parentheses are standard errors. Asterisks indicate significance: * $p < 0.10$, ** $p < 0.05$.

In contrast, the selection effect ($\beta$) increases in magnitude across specifications but remains statistically insignificant throughout. The estimated values of $\beta$ rise from 0.154

in column (1) to 0.39 in column (4), yet the standard errors increase even more sharply, from 0.368 to 0.613, resulting in wide confidence intervals and imprecise inference. This pattern suggests that while selection may play a role in shaping the sectoral choices at the individual level, its aggregate contribution to explaining the APG is weak and highly uncertain in this specific context.

Hence, the recovered structural parameters suggest that sector-wide productivity differences are the primary driver of the observed APG in this setting. As more covariates increase, the explanatory power of $\alpha$ becomes more precise and consistent, while the contribution of $\beta$ becomes increasingly diffuse and empirically fragile. This fragility of individual sorting at the sector level originates from the empirical distribution of unobserved comparative advantages $\theta_i$.

## 5.4 Individual Selection on APG

As discussed in the previous subsection, the selection effect $\beta$ is not statistically significant in shifting the observed earnings at the sector level. However, this extra reward from the individuals who work in the non-agricultural sector only captures the partial selection effect from unobserved comparative advantages at the aggregate level. As described by the Equation 32 in Section 3.5, the impact of individual sorting based on latent abilities at the sector level comprises two components: one is the additional returns from the average workers by choosing the non-agricultural vs. agricultural sector, and the other is the average differences of the unobserved comparative advantages between the two sectors. $\beta$ only reflects the selection effect from the former.

To infer the second component of the sorting at the sector level, I need to compute further the difference of the conditional means of $\theta_i$ for the workers between the non-agricultural and agricultural sectors. Since the estimated $\hat{\lambda}_0$ is not at level, I cannot calculate $E[\theta_i|D=1]$ and $E[\theta_i|D=0]$ separately. However, the difference in these two conditional means cancels the intercept $\lambda_0$ shown in the Equation 35. Therefore, I can sum the estimated $\hat{\theta}_i$ values in each sector at each wave, and then take the mean difference of these two estimated $\hat{\theta}_i$'s in the respective sectors. At each period, the mean difference between the estimated $\hat{\theta}_i$'s between

the two sectors yields the conditional mean difference from the unobserved comparative advantage, which is the second component in the Equation 32 and refers to how different the latent abilities are between the farmers and non-farmers on average at a given wave.

Table 2 shows the share of this conditional mean difference in APG at each period and the average over the waves. The observed log earnings differences at the sector level, which is APG, are in the second column, right next to the column for recording the waves. The difference in the means of latent abilities between farmers and non-farmers is in the third column, and the share of this conditional mean difference in the APG is in the far right column. The APG from the data are 1.0655, 1.0208, and 1.1224 log points in waves one to three, respectively, and 1.0696 log points on average. The latent skills between non-farmers and farmers are slightly negative, -0.0063, -0.0361, and -0.0285 log points in the sequential waves, and -0.0236 log points on average. Overall, the impact of this latent ability's conditional mean difference between non-agricultural and agricultural workers accounts for only 2.22% of the APG, with a direction opposite to that of the observed earnings gap.

Table 2: Conditional Mean Difference in APG

| **Wave** | $APG$ | $E[\theta_i\|D=1] - E[\theta_i\|D=0]$ | $\frac{\Delta E[\theta_i]}{APG}$ |
|---|---|---|---|
| 1 | 1.0655 | -0.0063 | -0.0059 |
| 2 | 1.0208 | -0.0361 | -0.0354 |
| 3 | 1.1224 | -0.0285 | -0.0254 |
| Average | 1.0696 | -0.0236 | -0.0222 |

Notes: This table reports the estimated conditional mean difference in unobserved comparative advantage between non-agriculture ($D = 1$) and agriculture ($D = 0$) groups, corresponding to the second component in equation 32.

This finding suggests that the latent abilities are not substantially different between non-agricultural and agricultural workers at each period after averaging over the samples. This result does not indicate that farmers and non-farmers are alike in terms of their unobserved comparative advantages at the individual level; instead, the heterogeneity of the latent abilities is not distinct after averaging over the workers in respective sectors. This weak difference

of latent abilities between nonagricultural and agricultural workers at the sector level reconciles with the estimation on the structural parameter selection effect $\beta$ with a large point estimate but highly dispersed.

Figure 5 plots the values of the estimated revealed comparative advantages $\hat{\theta}_i$ after centring the $\sum \hat{\theta}_i = 0$, with the normalized value of $\hat{\theta}_i$ labelled on the horizontal axis and the group share in the sample on the vertical axis. Although the gap of earnings between sectors is substantial (1.0696 log points on average), the average latent skills among groups fall within the range of -0.2 to 0.1 log points. This moderate unobserved heterogeneity among the different choice trajectory groups likely stems from their high dispersion within the sector. Moreover, the groups closer to zero have more observations than those located far away. Together, the average difference of unobserved heterogeneity between agricultural and non-agricultural workers plays a little role in explaining observed APG.

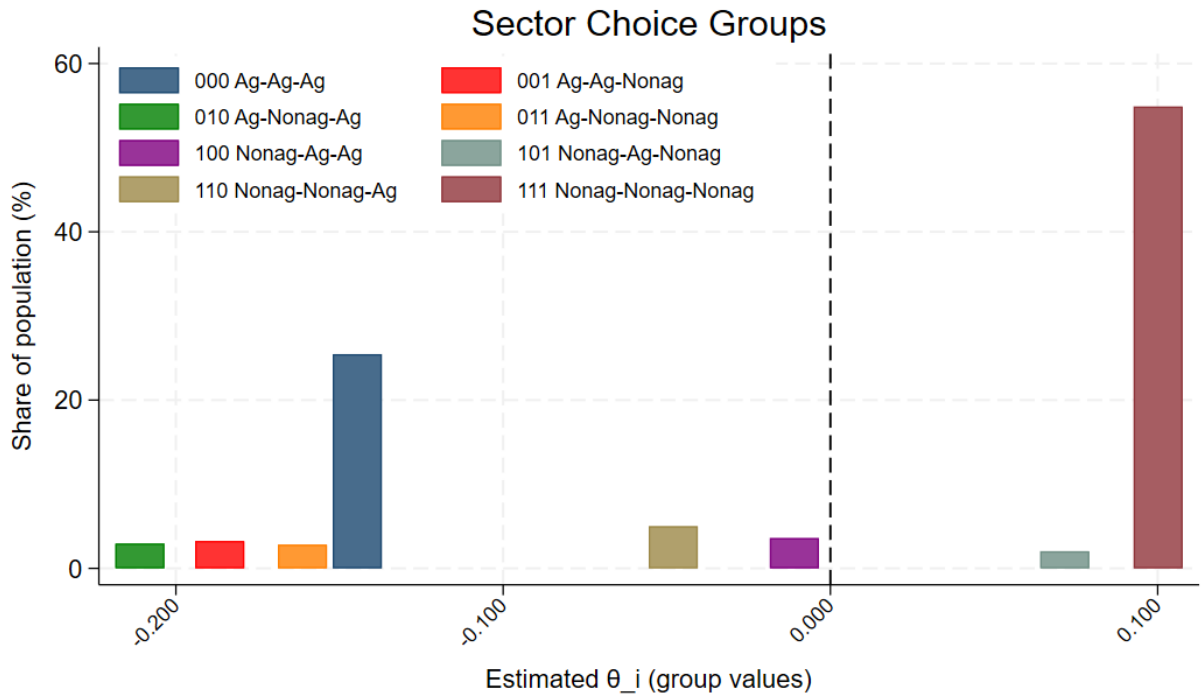


Figure 5: Recovered Distribution of Comparative Advantage for Specification (4)

In sum, the selection from the unobserved comparative advantage does not have a significant impact on the sectoral productivity gap in this setting. The selection at the sector level comprises two components: the extra returns for the average non-agricultural workers,

and the mean differences between non-agricultural and non-agriculture workers. The structural parameter, selection effect ($\beta$), suggests an insignificant impact of individual sorting on the sectoral productivity gap. In addition, the mean difference between farmers and non-farmers explains about 2.2% of the APG, which is insignificant compared to the sector-wide productivity gap and selection on observed characteristics.

This paper analyzes the first three waves of the IFLS survey and finds three key points. First, the sector-wide productivity gap drives the APG, suggesting that sectoral structure and technology are vital in explaining the productivity disparities at the aggregate level. This finding suggests a crucial role for sector-level allocation efficiency and technology improvement in reducing the APG. Second, the selection effect is positive in magnitude but statistically insignificant due to high dispersion in latent abilities, suggesting widespread heterogeneity in individual comparative advantages. However, this heterogeneity in latent skills at the individual level plays a relatively minor role in explaining the APG, after averaging across sectors. More importantly, a high average selection effect at the personal level may not significantly impact sectoral impact, as it also critically depends on the distribution of the unobserved heterogeneity. Third, the estimated empirical comparative advantages features two peaks and does not exhibit normality, indicating that imposing restrictive distributional assumptions may misrepresent underlying heterogeneity and selection effects.

# 6 Discussion

Contrary to the prevailing consensus in the APG literature, which attributes a substantial share of sectoral productivity gaps to self-selection, this paper finds that individual sorting contributes minimally to the average earnings disparities between the agricultural and non-agricultural sectors. Moreover, this study identifies a persistent and substantial sector-wide productivity gap as a crucial driver of the APG. Hence, the combination of the sector-wide technology difference and individual observed heterogeneity explains the majority of the sectoral productivity gaps when using a balanced panel in the first three waves of the IFLS data.

To reconcile my findings with the consensus in the current literature, I conduct a series

of comparison exercises using the same balanced panel data in this study. First, I estimate the selection effect using a two-way fixed effects (TWFE) model and calculate the difference between the sectoral productivity gaps with and without controlling for individual fixed effects to infer the selection effect. This method yields a significant selection effect on APG. Next, I apply the canonical Heckman two-step estimator, which assumes that the distribution of the unobserved components is jointly normal. The results from both pooled and panel data exhibit significant selection on the sectoral productivity gaps. Hence, both the TWFE and distributional assumption methods produce estimates that align with the relevant findings in the APG literature.

Finally, I implement the selection bias correction procedure for panel data developed by Wooldridge (1995), which relaxes the joint normality assumption on individual latent abilities and applies a control function to correct for selection bias in panel settings. On the same dataset, once I relax the distributional assumptions on unobserved abilities, the estimated selection effect becomes statistically insignificant, which reconciles with my findings. The key takeaway from these comparison exercises shows that the estimation of the selection effect depends critically on the choice of the method. Therefore, it is essential to recognize the limitations of various approaches and choose the most suitable one to address the research question at hand.

## 6.1 Estimating Selection Using a TWFE Approach

Several studies in the APG literature rely on TWFE models to estimate or infer the impact of latent skills on the sectoral earnings gap, concluding that there is a large and significant selection effect on APG. This subsection shows that using the TWFE on the first three-wave IFLS dataset also yields a significant selection effect.

Table 9 reports pooled OLS for three IFLS waves with log primary-job earnings as the outcome. Adding covariates and year fixed effects reduces the raw APG, with the fully specified model yielding an observed gap of 0.651 log points. Table 10 estimates the same specifications in panel form; even-numbered columns include individual and year fixed effects (TWFE), odd-numbered columns omit individual fixed effects. Standard errors are clustered at the person level. Introducing individual fixed effects substantially lowers the

agriculture–nonagriculture gap across specifications. Table 11 quantifies these reductions: pooled OLS vs. TWFE differs by 0.556–0.213 log points (Panels b–d), and adding individual fixed effects within the panel reduces the gap by 0.451–0.191 log points (Panels c–d). All differences are statistically significant at the 1% level,consistent with the claim in the literature that selection effects are large.

While these reductions are statistically significant and align with previous works, TWFE's interpretation as "selection on comparative advantage" is problematic for two reasons. First, identification is local to switchers. With roughly 20% switching sectors in these waves, the fixed-effects estimates reflect within-person changes for a small, possibly non-representative subset, rather than the population-level selection relevant for APG. Second, TWFE absorbs *all* time-invariant heterogeneity. In the main model (eqs. (18)–(19)), permanent unobserved ability decomposes into a sector-relevant component, $\theta_i$ (comparative advantage), and a sector-irrelevant component, $\tau_i$. The earnings equation (eq. (23)) shows that TWFE differences out both $\theta_i$ and $\tau_i$. Because TWFE cannot separate sector-specific ability from common ability, it risks attributing earnings changes due to $\tau_i$—e.g., general work ethic or family networks that help in either sector—to sector choices. In short, TWFE could conflate general individual heterogeneity unrelated to sectoral choice with the actual sorting based on the latent skills, thereby overstating the contribution of the selection effect to the observed APG.

Unlike the fixed-effect method, the CRC approach models sector-specific latent abilities and exploits choice histories, isolating the comparative-advantage component $\theta_i$ rather than incorporating all time-invariant traits into a single undifferentiated fixed effect for each individual.

## 6.2 Empirical Consequences of Distributional Assumptions

After evaluating the limitations of the TWFE method in modelling selection based on latent skills, I turn to a second dominant strategy in the APG literature: parametric selection corrections rooted in Roy's (1951) framework. This approach assumes individuals choose sectors by comparing potential earnings, but only the chosen earnings are observed. While this setup mirrors the structure developed in Section 3, the literature typically departs in one

critical respect: it imposes strong distributional assumptions on unobserved heterogeneity, most commonly joint normality of sector-specific abilities.

Why does this matter? Because the parametric assumption, rather than the data alone, often drives the size and significance of the estimated selection effect. For example, several studies—including Pulido and Świecki (2019)—extend the Roy model with selection and mobility frictions and estimate their frameworks on the IFLS. Using indirect inference (Gouriéroux et al., 1993), Pulido and Świecki match wage regressions and mobility patterns under the joint normality of latent abilities and idiosyncratic shocks. Within this structure, they find that selection accounts for 45–70% of the APG, depending on substitution elasticities. These results illustrate how significant selection effects can emerge under parametric assumptions, rather than directly from the data-generating process.

To assess how these assumptions shape empirical results in my context, I re-estimate the selection effect using the first three waves of IFLS under three standard parametric corrections: the canonical Heckman two-step (Heckman, 1979), the panel MLE estimator **xtheckman**, and Wooldridge's (1995) control function approach. For clarity, the full model setup for each method, as well as the discussion of exclusion restrictions, are provided in Appendix G; here I focus on comparative interpretation. These exercises demonstrate how imposing or relaxing distributional assumptions on latent abilities changes the magnitude, and even the sign, of the estimated selection effect.

Table 14 reports the Heckman two-step results for the pooled data. The IMR coefficients are statistically significant in both sectors, implying systematic sectoral selection. Specifically, the estimates indicate negative selection into non-agriculture (IMR coefficient of around $-0.14$) and positive selection into agriculture (IMR coefficient ranging from 0.20 to 0.25). To bolster credibility, I impose exclusion restrictions: variables such as age and non-farm business are used for non-agriculture, while rural-born, marital status, and farm business are used for agriculture. As Tables 12–13 show, these instruments strongly predict sector choice but are not significant determinants of sectoral wages, supporting their validity.

Table 15 summarizes the implied contribution of selection to the observed APG. Depending on specification, selection into non-agriculture explains 21–22% of the earnings gap, while selection into agriculture accounts for 31–38%. These magnitudes align with the find-

ings of Pulido and Świecki (2019), underscoring the strong influence of the joint normality assumption on estimates of selection.

To account for the panel nature of the data, I next apply `xtheckman`, a maximum likelihood estimator that extends Heckman's framework to panel settings. Rather than computing an IMR for each observation, this method directly estimates the correlation between unobserved sectoral fixed effects and time-varying error terms. Convergence proved challenging, with none of the specifications reaching full convergence despite extensive iterations. Nonetheless, where results are available (Table 16), the correlations between unobservables across sectors are large and statistically significant. This suggests that, under the joint normality assumption, selection bias remains substantial even in the panel framework. These results broadly align with Pulido and Świecki's (2019) indirect inference estimates, highlighting that, regardless of estimation technique, assuming bivariate normality produces significant selection effects. Appendix G.2 provides further details.

Finally, I turn to Wooldridge's (1995) control function approach, which relaxes the joint normality assumption. Unlike Heckman or `xtheckman`, it does not hinge on any parametric assumption about the joint distribution of unobserved sectoral abilities. This method employs panel differencing to eliminate time-invariant heterogeneity and incorporates generalized residuals from a first-stage probit regression as control functions in the wage equation. Applied to the same dataset, the estimated coefficients on the control functions are statistically insignificant across specifications and both sectors (Table 19). Appendix G.3 provides the full model exposition and implementation details.

Taken together, these results demonstrate that distributional assumptions, rather than the data alone, drive the size and even the sign of the estimated selection effect. Under joint normality (as in Heckman or `xtheckman`), selection appears large and significant, explaining a non-trivial share of the APG. Under weaker assumptions (as in Wooldridge), the selection effect disappears. This sensitivity highlights the methodological risk of interpreting parametric Roy-model estimates as structural facts about the labour market: what appears to be strong evidence of self-selection may instead be an artifact of functional form assumptions imposed on unobserved heterogeneity. This contrast reinforces the value of my CRC framework, which avoids imposing such assumptions while directly modelling sector-specific

comparative advantage.

In sum, the fixed-effect approach is not well-suited for studying self-selection based on latent skills because unobserved comparative advantages necessitate distinguishing individual fixed effects at the sector level, which the fixed-effect method cannot achieve. On the other hand, imposing distributional assumptions on latent skills across sectors often has a consequential impact on the estimation results of the selection effect. In this section, I demonstrate that the selection on APG would be significant if I were to implement either a fixed-effects or distributional assumption. This illustration reconciles the findings in my paper with the relevant studies in the APG literature. However, using the different empirical method by Suri (2011), I find that individual sorting is insignificant in terms of sectoral productivity gaps in Indonesia.

This paper estimates sector-specific comparative advantage without imposing any distributional assumptions. This method treats sectoral choices over time as informative signals of latent comparative advantage. It exploits the panel structure of the data to recover structural parameters non-parametrically. However, this method also faces limitations. As Tjernström et al. (2023) emphasize, identification requires variation in choice trajectories and earnings over time. When incomes are very similar across different choice groups or choice transitions are incomplete, the system of equations may become weakly identified or even break down. Moreover, the method requires a balanced panel and assumes that unobserved comparative advantages are time-invariant.

Despite these caveats, the Suri-inspired approach offers a valuable alternative to existing methods by avoiding functional form assumptions and directly modelling unobserved comparative advantages. Given the stark contrast in results across the three parametric approaches examined in this section, and the fragility of distributional assumptions in this context, the Suri-based estimator provides a theoretically and empirically grounded alternative to revisit long-standing claims about the role of selection in explaining agricultural productivity gaps.

# 7 Conclusion

This paper revisits a critical question: how much of the agricultural productivity gap (APG) can be explained by individual sorting on unobserved comparative advantage? Using the Indonesia Family Life Survey (IFLS), I focus on the pre-decentralization period of the 1990s, a decade of stability before the structural break of the 2000 "Big Bang" reforms. Unlike much of the APG literature, which pools all five IFLS waves and concludes that selection plays a large role, I show that sorting contributed little to the productivity gap during this earlier period. Instead, sector-wide productivity differences and observable heterogeneity explain most of the earnings disparity between agriculture and non-agriculture.

This study contributes on three fronts. First, it demonstrates that the impact of individual sorting on APG has two distinct components: (i) the selection effect, $\beta$, which measures the average extra returns to unobserved comparative advantage; and (ii) the mean differences in latent abilities across sectors. Previous work typically emphasized only $\beta$, overlooking the role of the second term. Unless agricultural and non-agricultural workers are similar on average in their latent abilities, $\beta$ represents only part of the sorting impact at the sectoral level. This paper addresses this blind spot in the literature.

Second, the paper adapts and applies the CRC framework to evaluate selection on APG, drawing on Suri (2011), Lemieux (1998), and Chamberlain (1982). The framework (i) redefines absolute advantage consistently with APG measurement; (ii) models comparative advantage as individual deviations from sector means; and (iii) recovers $\beta$ and the distribution of $\theta_i$ without parametric assumptions (e.g., joint normality). In doing so, it contrasts directly with prevailing approaches: two-way fixed effects, which absorb all time-invariant heterogeneity (including sector-irrelevant ability), and parametric selection corrections, which hinge on functional form assumptions. Echoing the concerns of Heckman and Honore (1990), the analysis shows that distributional assumptions have major consequences when empirically measuring the selection effect due to unobservable abilities.

Third, the paper formally maps the selection effect $\beta$ to the classic Roy model framework. This mapping anchors Suri's CRC framework in the established labour literature while highlighting its key distinction: unlike the coefficient on an inverse Mills ratio, $\beta$ is recovered

without distributional assumptions. This connection situates the paper's contribution within a broader intellectual lineage and clarifies how the CRC approach relates to the classic Roy framework.

The policy relevance of this study lies in showing that sorting affects APG through two channels, not one. Moreover, the empirical distribution of latent abilities in IFLS is far from normal and shows mass in both tails, which means the two terms can even move in opposite directions. Under joint normality, the selection effect $\beta$ and the mean difference in latent abilities always move in the same direction, differing only in magnitude. In that world, policies based on $\beta$ alone would miss the target but not misfire. In reality, when the distribution is skewed or fat-tailed, as the IFLS evidence suggests, the two terms may diverge, and policies that target only one risk failure or unintended consequences.

Two cases illustrate the stakes. In Sub-Saharan Africa (2000s–2010s), large-scale skills and entrepreneurship programs underperformed because they assumed that general training would yield higher non-farm earnings. In settings where $\beta$ was weakly negative but mean ability differences across sectors were large, reallocating workers with poor sectoral fit produced disappointing outcomes. By contrast, Indonesia's rice-intensification programs in the 1970s–1990s succeeded by raising agricultural productivity itself, benefiting workers regardless of comparative advantage and narrowing earnings gaps without requiring reallocation. Together, these examples underscore that policies should not assume alignment between sorting components; instead, they should account for both the returns to comparative advantage and the average ability gap across sectors.

In summary, this paper offers both a conceptual correction and an empirical reevaluation of selection in APG analysis. By incorporating the missing component of latent mean differences, adopting a distribution-free CRC framework, and situating the analysis within Indonesia's pre-decentralization context, the study demonstrates that individual sorting explained a small portion of the country's APG in 1990s Indonesia. The gap was driven largely by sector-wide technology differences and observable heterogeneity.

More importantly, this paper demonstrates that revealed comparative advantages can help policymakers gather valuable information and design effective policies that align with their intended objectives. In doing so, it provides a roadmap for future research and policy

design that treats self-selection not as a foregone conclusion but as an empirical question, with potentially very different answers across time, space, and institutional context.

Table 3: Descriptive Statistics for Individuals – Part A

| | Number of Survey Round | | | |
|---|---|---|---|---|
| **Variable** | **1** | **2** | **3** | **Total** |
| **N** | 4,615 | 4,615 | 4,615 | 13,845 |
| | (33.3%) | (33.3%) | (33.3%) | (100.0%) |
| **hh_count** | 3,963.000 | 3,991.000 | 4,012.000 | 4,012.667 |
| | (98.8%) | (99.5%) | (100.0%) | (100.0%) |
| **gender** | 0.727 | 0.727 | 0.727 | 0.727 |
| | (0.446) | (0.446) | (0.446) | (0.446) |
| **age** | 40.407 | 44.335 | 47.308 | 44.016 |
| | (11.250) | (11.173) | (11.277) | (11.504) |
| **non-agriculture (primary job)** | 0.655 | 0.657 | 0.630 | 0.647 |
| | (0.475) | (0.475) | (0.483) | (0.478) |
| **wagedwork (primary job)** | 0.454 | 0.446 | 0.408 | 0.436 |
| | (0.498) | (0.497) | (0.491) | (0.496) |
| **income** | 126,194.800 | 246,299.956 | 432,582.566 | 268,359.107 |
| | (167,450.486) | (2,692,023.335) | (1,212,964.123) | (1,711,988.882) |
| **ln_income** | 11.127 | 11.629 | 12.096 | 11.617 |
| | (1.209) | (1.213) | (2.026) | (1.582) |
| **hours worked** | 177.964 | 171.676 | 172.585 | 174.088 |
| | (74.269) | (74.998) | (81.850) | (77.168) |
| **ln_hours worked** | 5.076 | 5.027 | 5.014 | 5.039 |
| | (0.505) | (0.546) | (0.596) | (0.551) |
| **cpi** | 145.195 | 194.337 | 209.439 | 182.990 |
| | (4.604) | (7.064) | (9.432) | (28.083) |
| **ln_cpi** | 4.978 | 5.269 | 5.343 | 5.197 |
| | (0.031) | (0.036) | (0.045) | (0.168) |
| **ruralborn** | 0.746 | 0.749 | 0.792 | 0.763 |
| | (0.435) | (0.434) | (0.406) | (0.425) |
| **moved** | 0.533 | 0.465 | 0.462 | 0.487 |
| | (0.499) | (0.499) | (0.499) | (0.500) |
| **urban** | 0.442 | 0.439 | 0.436 | 0.439 |
| | (0.497) | (0.496) | (0.496) | (0.496) |

*Notes:* Top line reports means (or counts for *N*, *hh_count*); second line reports standard deviations for continuous/binary variables and percentages for *N* and *hh_count*. Monetary values in Indonesian Rupiah. Statistics use the balanced panel (waves 1–3).

Table 4: Descriptive Statistics for Individuals – Part B (Categorical Variables)

| | Number of Survey Round | | | |
|---|---|---|---|---|
| **Category** | **1** | **2** | **3** | **Total** |
| **N** | 4,615 (33.3%) | 4,615 (33.3%) | 4,615 (33.3%) | 13,845 (100.0%) |
| **MARITAL STATUS** | | | | |
| Not yet married | 99 (2.1%) | 54 (1.2%) | 38 (0.8%) | 191 (1.4%) |
| Married | 4,231 (91.7%) | 4,194 (90.9%) | 4,146 (89.8%) | 12,571 (90.8%) |
| Separated | 19 (0.4%) | 22 (0.5%) | 24 (0.5%) | 61 (0.4%) |
| Divorced | 56 (1.2%) | 56 (1.2%) | 59 (1.5%) | 181 (1.3%) |
| Widowed | 210 (4.6%) | 289 (6.3%) | 342 (7.4%) | 841 (6.1%) |
| **EDUCATION** | | | | |
| Unschooled | 698 (15.1%) | 651 (14.1%) | 619 (13.4%) | 1,968 (14.2%) |
| Primary | 2,372 (51.4%) | 2,434 (52.8%) | 2,362 (51.2%) | 7,168 (51.8%) |
| Junior high | 548 (11.9%) | 509 (11.0%) | 485 (10.5%) | 1,542 (11.1%) |
| Senior high | 760 (16.5%) | 745 (16.1%) | 667 (14.5%) | 2,172 (15.7%) |
| College/University | 233 (5.1%) | 273 (5.9%) | 341 (7.4%) | 847 (6.1%) |
| Others | 0 (0.0%) | 2 (0.0%) | 136 (3.0%) | 138 (1.0%) |
| **RELIGION** | | | | |
| Islam | 4,007 (86.8%) | 4,030 (87.3%) | 4,020 (87.1%) | 12,057 (87.1%) |
| Protestant | 187 (4.1%) | 185 (4.0%) | 192 (4.2%) | 564 (4.1%) |
| Catholic | 90 (2.0%) | 92 (2.0%) | 94 (2.0%) | 276 (2.0%) |
| Hindu | 286 (6.2%) | 281 (6.1%) | 284 (6.2%) | 851 (6.1%) |
| Buddhist | 23 (0.5%) | 22 (0.5%) | 19 (0.4%) | 64 (0.5%) |
| Others | 22 (0.5%) | 7 (0.2%) | 4 (0.1%) | 33 (0.2%) |

*Notes:* Each cell shows count with share in parentheses. Shares sum to 100% within each block.

Table 5: Descriptive Statistics for Households

| | Number of Survey Round | | | |
|---|---|---|---|---|
| **Variable** | **1** | **2** | **3** | **Total** |
| **N** | 3,963 | 3,991 | 4,012 | 11,966 |
| | (33.1%) | (33.4%) | (33.5%) | (100.0%) |
| **panel** | . | 0.980 | 0.978 | 0.979 |
| | (.) | (0.141) | (0.146) | (0.144) |
| **household size** | 4.777 | 5.361 | 5.801 | 5.315 |
| | (1.981) | (2.180) | (2.408) | (2.237) |
| **urban** | 0.430 | 0.426 | 0.427 | 0.428 |
| | (0.495) | (0.495) | (0.495) | (0.495) |
| **own farm business** | 0.434 | 0.397 | 0.476 | 0.436 |
| | (0.496) | (0.489) | (0.500) | (0.496) |
| **farm business assets** | 7,144,077.782 | 10,330,989.860 | 26,572,186.430 | 15,729,198.326 |
| | (33,921,100.065) | (21,343,301.269) | (83,485,328.313) | (57,595,891.874) |
| **ln_farm business assets** | 14.064 | 14.798 | 15.550 | 14.864 |
| | (2.216) | (1.998) | (2.127) | (2.208) |
| **owns non-farm business** | 0.374 | 0.398 | 0.523 | 0.432 |
| | (0.484) | (0.490) | (0.500) | (0.495) |
| **non-farm business assets** | 6,374,035.384 | 5,439,189.255 | 9,510,752.091 | 7,351,954.719 |
| | (59,736,751.708) | (24,837,647.647) | (46,064,887.772) | (45,183,582.166) |
| **ln_non-farm business assets** | 12.257 | 12.976 | 13.350 | 12.931 |
| | (2.356) | (2.339) | (2.434) | (2.423) |
| **own family business** | 0.675 | 0.672 | 0.803 | 0.717 |
| | (0.469) | (0.469) | (0.398) | (0.451) |
| **own assets** | 0.976 | 0.999 | 0.999 | 0.991 |
| | (0.153) | (0.035) | (0.035) | (0.093) |
| **total assets not for business** | 14,005,326.895 | 20,270,537.043 | 37,537,032.473 | 24,060,081.890 |
| | (78,347,639.350) | (52,369,917.086) | (85,931,361.293) | (74,282,490.131) |
| **ln_total assets not for business** | 14.709 | 15.654 | 16.392 | 15.595 |
| | (1.785) | (1.621) | (1.586) | (1.801) |
| **real estate not for business** | 11,356,720.903 | 18,508,233.980 | 32,133,196.522 | 20,998,405.810 |
| | (56,657,512.411) | (49,100,027.450) | (67,800,563.543) | (59,108,696.917) |
| **ln_real estate not for business** | 14.826 | 15.623 | 16.309 | 15.611 |
| | (1.573) | (1.519) | (1.522) | (1.651) |
| **shock** | 0.312 | 0.402 | 0.345 | 0.353 |
| | (0.463) | (0.490) | (0.475) | (0.478) |
| **numbers of shock** | 0.392 | 0.567 | 0.443 | 0.467 |
| | (0.653) | (0.821) | (0.700) | (0.732) |

*Notes:* Top line reports means (or counts for $N$); second line reports standard deviations (or column shares for $N$). Monetary values are in Indonesian Rupiah.

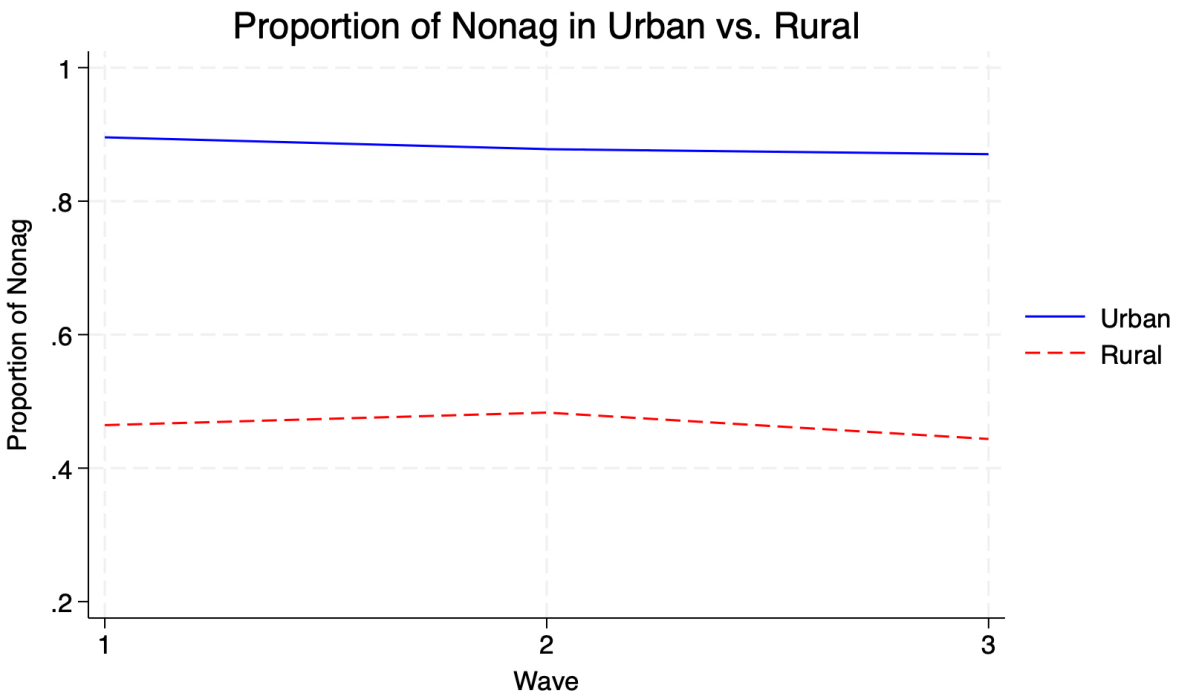


Figure 6: Non-agriculture Share Urban vs. Rural over Three Waves

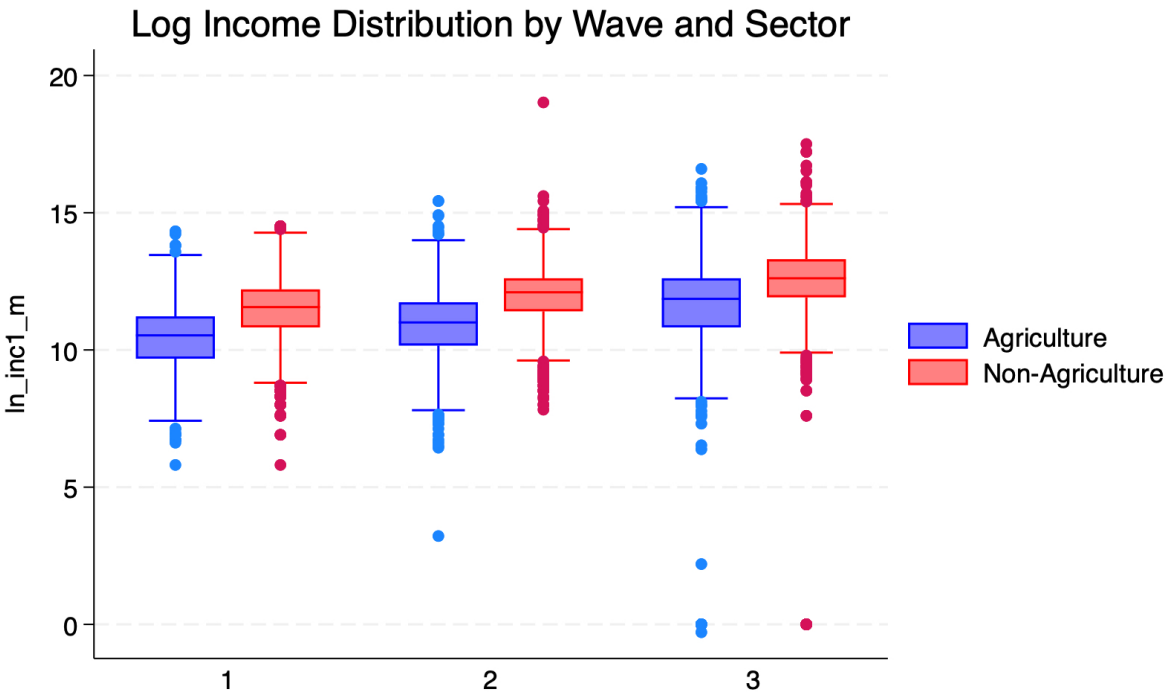


Figure 7: Log Income Non-agriculture vs. Agriculture over Three Waves

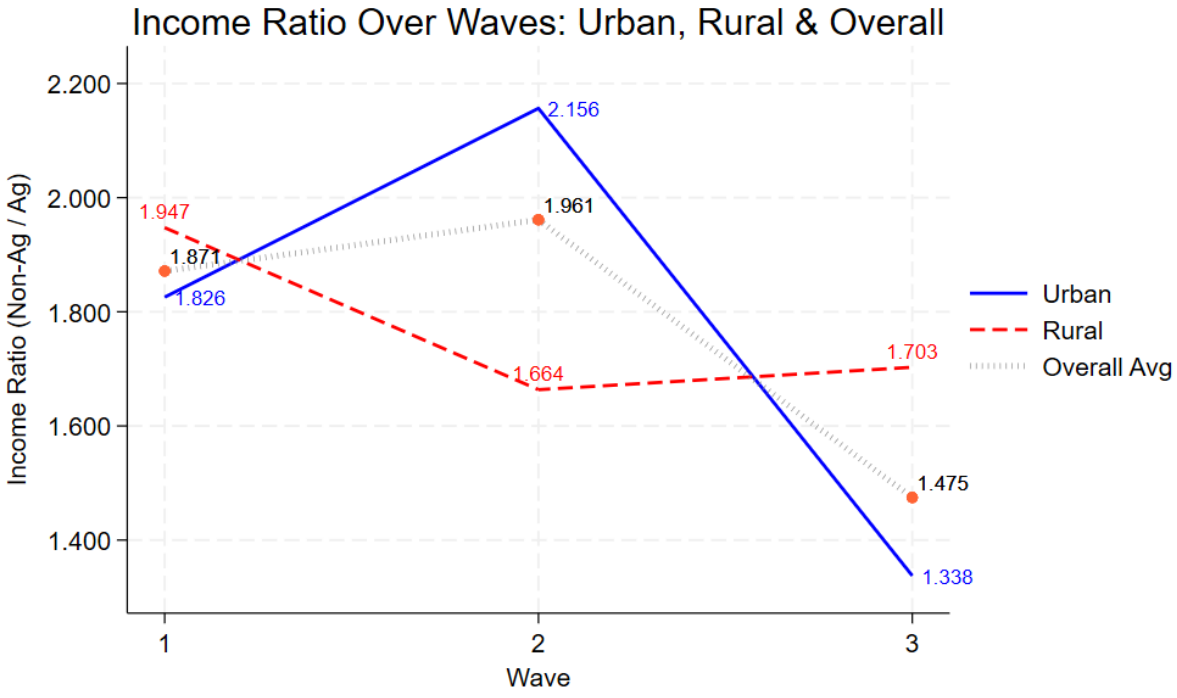


Figure 8: Sectoral Earnings Gaps Urban vs. Rural over Three Waves

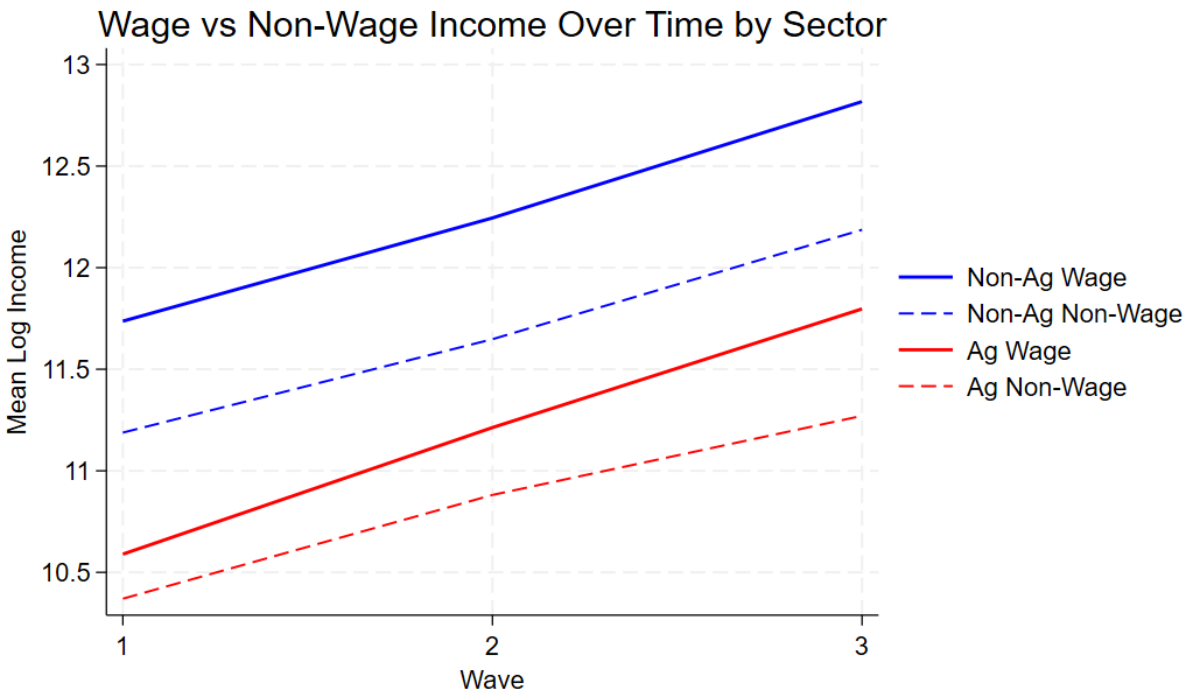


Figure 9: Sectoral Log Earning Difference Formal vs. Informal Workers

Table 6: SUR: Outcome Variable, log Earnings in 1993

| **ln_earnings_1** | **(1)** | **(2)** | **(3)** | **(4)** |
|---|---|---|---|---|
| nonag_1 | 0.693 ** | 0.592 ** | 0.53 ** | 0.539 ** |
| | (0.09) | (0.089) | (0.08) | (0.079) |
| nonag_2 | -0.036 | -0.057 | -0.097 | -0.092 |
| | (0.099) | (0.097) | (0.088) | (0.087) |
| nonag_3 | 0.012 | -0.009 | 0.037 | 0.059 |
| | (0.094) | (0.093) | (0.082) | (0.082) |
| nonag_1 * nonag_2 | 0.028 | 0.043 | 0.075 | 0.100 |
| | (0.148) | (0.145) | (0.13) | (0.129) |
| nonag_1 * nonag_3 | 0.246 | 0.220 | 0.076 | 0.052 |
| | (0.169) | (0.166) | (0.148) | (0.147) |
| nonag_2 * nonag_3 | 0.207 | 0.188 | 0.129 | 0.114 |
| | (0.163) | (0.160) | (0.144) | (0.143) |
| nonag_1 * nonag_2 * nonag_3 | 0.002 | -0.097 | -0.056 | -0.058 |
| | (0.228) | (0.224) | (0.201) | (0.199) |
| urban_1 | | 0.452 ** | 0.185 ** | 0.185 ** |
| | | (0.097) | (0.087) | (0.086) |
| urban_2 | | -0.192 | 0.151 | 0.109 |
| | | (0.114) | (0.103) | (0.101) |
| urban_3 | | 0.067 | -0.067 | -0.047 |
| | | (0.084) | (0.075) | (0.074) |
| ln_hrsworked_1 | | | 0.362 ** | 0.354 ** |
| | | | (0.031) | (0.030) |
| ln_hrsworked_2 | | | 0.053 * | 0.051 ** |
| | | | (0.028) | (0.028) |
| ln_hrsworked_3 | | | 0.026 | 0.026 |
| | | | (0.026) | (0.025) |
| wagedwork_1 | | | 0.106 ** | 0.105 ** |
| | | | (0.042) | (0.042) |
| wagedwork_2 | | | 0.086 ** | 0.088 ** |
| | | | (0.045) | (0.044) |
| wagedwork_3 | | | -0.065 | -0.057 |
| | | | (0.042) | (0.042) |
| ln_cpi_1 | | | | -3.051 ** |
| | | | | (0.941) |
| ln_cpi_2 | | | | 3.609 ** |
| | | | | (0.726) |
| ln_cpi_3 | | | | 1.082 ** |
| | | | | (0.408) |
| shock_1 | | | | -0.021 |
| | | | | (0.031) |
| shock_2 | | | | -0.008 |
| | | | | (0.030) |
| shock_3 | | | | 0.038 |
| | | | | (0.030) |
| province | N | Y | Y | N |
| age | N | N | Y ** | Y ** |
| gender | N | N | Y | Y |
| marital_status | N | N | Y | Y |
| religion | N | N | Y | Y |
| education | N | N | Y ** | Y ** |
| constant | 10.412 ** | 10.391 ** | 6.919 ** | -2.680 |
| | (0.032) | (0.050) | (0.225) | (3.741) |
| $N$ | 4,615 | 4,614 | 4,513 | 4,510 |
| $R^2$ | 0.189 | 0.221 | 0.392 | 0.402 |

*Notes:* Numbers in parentheses are **standard errors**. Asterisks indicate significance: * $p < 0.10$, ** $p < 0.05$.

Table 7: SUR: Outcome Variable, log Earnings in 1997

| **ln_earnings_2** | **(1)** | **(2)** | **(3)** | **(4)** |
|---|---|---|---|---|
| nonag_1 | 0.277 ** | 0.186 * | 0.165 ** | 0.177 ** |
| | (0.091) | (0.090) | (0.079) | (0.078) |
| nonag_2 | 0.462 ** | 0.437 ** | 0.340 ** | 0.352 ** |
| | (0.099) | (0.098) | (0.087) | (0.087) |
| nonag_3 | -0.045 | -0.064 | -0.070 | -0.055 |
| | (0.095) | (0.093) | (0.082) | (0.081) |
| nonag_1 * nonag_2 | -0.071 | -0.054 | -0.036 | -0.021 |
| | (0.149) | (0.147) | (0.130) | (0.128) |
| nonag_1 * nonag_3 | 0.295 | 0.276 * | 0.141 | 0.114 |
| | (0.170) | (0.168) | (0.147) | (0.146) |
| nonag_2 * nonag_3 | 0.114 | 0.105 | 0.049 | 0.030 |
| | (0.164) | (0.162) | (0.143) | (0.142) |
| nonag_1 * nonag_2 * nonag_3 | 0.137 | 0.358 | 0.087 | 0.101 |
| | (0.230) | (0.226) | (0.199) | (0.198) |
| urban_1 | | 0.438 ** | 0.180 ** | 0.199 ** |
| | | (0.097) | (0.087) | (0.086) |
| urban_2 | | -0.127 | 0.051 | 0.041 |
| | | (0.115) | (0.102) | (0.101) |
| urban_3 | | 0.131 | 0.0003 | -0.041 |
| | | (0.084) | (0.075) | (0.074) |
| ln_hrsworked_1 | | | 0.099 ** | 0.087 ** |
| | | | (0.030) | (0.030) |
| ln_hrsworked_2 | | | 0.27 ** | 0.269 ** |
| | | | (0.028) | (0.028) |
| ln_hrsworked_3 | | | 0.057 ** | 0.058 ** |
| | | | (0.025) | (0.025) |
| wagedwork_1 | | | -0.018 | -0.008 ** |
| | | | (0.041) | (0.041) |
| wagedwork_2 | | | 0.185 ** | 0.181 ** |
| | | | (0.044) | (0.044) |
| wagedwork_3 | | | -0.165 | -0.011 |
| | | | (0.042) | (0.041) |
| ln_cpi_1 | | | | -5.449 ** |
| | | | | (0.933) |
| ln_cpi_2 | | | | 4.849 ** |
| | | | | (0.720) |
| ln_cpi_3 | | | | 0.444 |
| | | | | (0.405) |
| shock_1 | | | | 0.063 ** |
| | | | | (0.031) |
| shock_2 | | | | -0.044 |
| | | | | (0.030) |
| shock_3 | | | | 0.003 |
| | | | | (0.030) |
| province | N | Y | Y | N |
| age | N | N | Y ** | Y ** |
| gender | N | N | Y | Y |
| marital_status | N | N | Y | Y |
| religion | N | N | Y | Y |
| education | N | N | Y ** | Y ** |
| constant | 10.906 ** | 10.992 ** | 7.902 ** | 7.007 ** |
| | (0.032) | (0.050) | (0.224) | (3.713) |
| N | 4,615 | 4,614 | 4,513 | 4,510 |
| $R^2$ | 0.185 | 0.212 | 0.406 | 0.417 |

*Notes:* Numbers in parentheses are **standard errors**. Asterisks indicate significance: * $p < 0.10$, ** $p < 0.05$.

Table 8: SUR: Outcome Variable, log Earnings in 2000

| **ln_earnings_3** | **(1)** | **(2)** | **(3)** | **(4)** |
|---|---|---|---|---|
| nonag_1 | 0.111 | -0.048 | 0.014 | 0.038 |
| | (0.161) | (0.161) | (0.156) | (0.156) |
| nonag_2 | -0.221 | -0.237 ** | -0.408 ** | -0.403 ** |
| | (0.176) | (0.175) | (0.173) | (0.172) |
| nonag_3 | 0.173 | 0.163 | 0.071 | 0.044 |
| | (0.168) | (0.168) | (0.162) | (0.162) |
| nonag_1 * nonag_2 | -0.059 | -0.036 | 0.159 | 0.164 |
| | (0.264) | (0.264) | (0.256) | (0.255) |
| nonag_1 * nonag_3 | 0.729 ** | 0.723 ** | 0.592 ** | 0.586 ** |
| | (0.302) | (0.264) | (0.291) | (0.290) |
| nonag_2 * nonag_3 | 0.516 * | 0.511 * | 0.454 | 0.467 * |
| | (0.291) | (0.291) | (0.283) | (0.282) |
| nonag_1 * nonag_2 * nonag_3 | -0.065 | -0.097 | -0.216 | -0.202 |
| | (0.407) | (0.402) | (0.394) | (0.392) |
| urban_1 | | 0.294 | 0.001 | 0.020 |
| | | (0.177) | (0.171) | (0.170) |
| urban_2 | | -0.192 | -0.002 | -0.055 |
| | | (0.208) | (0.202) | (0.200) |
| urban_3 | | 0.161 | 0.077 | 0.071 |
| | | (0.152) | (0.148) | (0.146) |
| ln_hrsworked_1 | | | 0.035 | 0.029 |
| | | | (0.060) | (0.060) |
| ln_hrsworked_2 | | | 0.098 * | 0.106 ** |
| | | | (0.055) | (0.055) |
| ln_hrsworked_3 | | | 0.323 ** | 0.313 ** |
| | | | (0.050) | (0.050) |
| wagedwork_1 | | | -0.149 * | -0.139 * |
| | | | (0.082) | (0.082) |
| wagedwork_2 | | | 0.069 | 0.075 |
| | | | (0.088) | (0.087) |
| wagedwork_3 | | | 0.308 ** | 0.303 ** |
| | | | (0.082) | (0.082) |
| ln_cpi_1 | | | | -3.748 ** |
| | | | | (1.854) |
| ln_cpi_2 | | | | 3.671 ** |
| | | | | (1.431) |
| ln_cpi_3 | | | | 0.937 |
| | | | | (0.805) |
| shock_1 | | | | 0.201 ** |
| | | | | (0.061) |
| shock_2 | | | | 0.006 |
| | | | | (0.059) |
| shock_3 | | | | -0.242 ** |
| | | | | (0.060) |
| province | N | Y | Y | N |
| age | N | N | Y ** | Y |
| gender | N | N | Y | Y |
| marital_status | N | N | Y | Y |
| religion | N | N | Y | Y |
| education | N | N | Y ** | Y ** |
| constant | 11.418 ** | 11.441 ** | 8.704 ** | 3.066 |
| | (0.057) | (0.091) | (0.443) | (7.377) |
| $N$ | 4,615 | 4,614 | 4,513 | 4,510 |
| $R^2$ | 0.083 | 0.087 | 0.168 | 0.175 |

*Notes:* Numbers in parentheses are **standard errors**. Asterisks indicate significance: * $p < 0.10$, ** $p < 0.05$.

| ln_inc1_m | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| nonag_main | 1.053 ** | 1.071 ** | 0.666 ** | 0.654 ** | 0.653 ** | 0.656 ** | 0.646 ** | 0.651 ** |
| | 0.027 | 0.033 | 0.03 | 0.034 | 0.029 | 0.034 | 0.029 | 0.033 |
| ln_hrsworked1_m | | | 0.337 ** | 0.349 ** | 0.353 ** | 0.35 ** | 0.351 ** | 0.349 ** |
| | | | 0.022 | 0.028 | 0.021 | 0.028 | 0.021 | 0.028 |
| wagedwork_main | | | 0.188 ** | 0.202 ** | 0.195 ** | 0.203 ** | 0.186 ** | 0.196 ** |
| | | | 0.027 | 0.026 | 0.026 | 0.026 | 0.026 | 0.026 |
| age | | | 0.004 ** | -0.006 ** | -0.005 ** | -0.006 ** | -0.005 ** | -0.006 ** |
| | | | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| gender | | | 0.392 ** | 0.406 ** | 0.401 ** | 0.405 ** | 0.403 ** | 0.407 ** |
| | | | 0.03 | 0.03 | 0.029 | 0.03 | 0.029 | 0.03 |
| education | | | 0.368 ** | 0.323 ** | 0.331 ** | 0.323 ** | 0.333 ** | 0.325 ** |
| | | | 0.012 | 0.013 | 0.011 | 0.013 | 0.011 | 0.013 |
| marital_status | | | -0.038 ** | -0.038 * | -0.039 ** | -0.039 ** | -0.036 ** | -0.037 ** |
| | | | 0.018 | 0.019 | 0.017 | 0.019 | 0.017 | 0.02 |
| urban | | | 0.238 ** | 0.257 ** | 0.252 ** | 0.253 ** | 0.249 ** | 0.251 ** |
| | | | 0.027 | 0.029 | 0.026 | 0.029 | 0.026 | 0.029 |
| ln_cpi | | | | | 2.457 ** | 1.926 ** | 2.469 ** | 1.933 ** |
| | | | | | 0.073 | 0.328 | 0.073 | 0.328 |
| shock | | | | | | | -0.086 ** | -0.07 ** |
| | | | | | | | 0.024 | 0.026 |
| Year | | | | | | | | |
| 1997 | | 0.5 ** | | 0.537 ** | | -0.024 | | -0.021 |
| | | 0.016 | | 0.017 | | 0.097 | | 0.097 |
| 2000 | | 0.996 ** | | 1.013 ** | | 0.307 ** | | 0.306 ** |
| | | 0.029 | | 0.029 | | 0.124 | | 0.124 |
| cons | 10.936 ** | 10.425 ** | 8.01 ** | 7.956 ** | -4.381 ** | -1.649 | -4.395 ** | -1.65 |
| | 0.0215 | 0.027 | 0.126 | 0.017 | 0.388 | 1.658 | 0.388 | 1.658 |
| clustered | N | Y | N | Y | N | Y | N | Y |
| N | 13,845 | 13,845 | 13,742 | 13,742 | 13,742 | 13,742 | 13,739 | 13,739 |
| $R^2$ | 0.101 | 0.167 | 0.225 | 0.2886 | 0.285 | 0.291 | 0.286 | 0.291 |

Table 9: Pooled OLS Regressions, With and Without Year Fixed Effects

| ln_inc1_m | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| nonag_main | 0.966 ** | 0.515 ** | 0.634 ** | 0.446 ** | 0.636 ** | 0.446 ** | 0.629 ** | 0.438 ** |
| | 0.035 | 0.072 | 0.035 | 0.069 | 0.035 | 0.069 | 0.035 | 0.069 |
| ln_hrsworked1_m | | | 0.326 ** | 0.268 ** | 0.328 ** | 0.269 ** | 0.326 ** | 0.264 ** |
| | | | 0.028 | 0.037 | 0.028 | 0.037 | 0.028 | 0.037 |
| wagedwork_main | | | 0.217 ** | 0.227 ** | 0.217 ** | 0.226 ** | 0.208 ** | 0.221 ** |
| | | | 0.026 | 0.055 | 0.026 | 0.055 | 0.026 | 0.054 |
| age | | | -0.007 ** | 0.0001 | -0.006 ** | 0.0001 | -0.006 ** | -0.001 |
| | | | 0.001 | 0.007 | 0.001 | 0.007 | 0.001 | 0.007 |
| gender | | | 0.406 ** | 0.874 | 0.405 ** | 0.863 | 0.408 ** | 0.946 * |
| | | | 0.031 | 0.531 | 0.03 | 0.539 | 0.305 | 0.495 |
| education | | | 0.3 ** | 0.3 | 0.302 ** | 0.031 | 0.303 ** | 0.032 |
| | | | 0.013 | 0.013 | 0.013 | 0.031 | 0.013 | 0.031 |
| marital_status | | | -0.041 ** | -0.027 | -0.041 ** | -0.027 | -0.038 * | -0.014 |
| | | | 0.019 | 0.036 | 0.019 | 0.036 | 0.019 | 0.037 |
| urban | | | 0.271 ** | 0.084 | 0.267 ** | 0.084 | 0.264 ** | 0.082 |
| | | | 0.029 | 0.081 | 0.029 | 0.08 | 0.029 | 0.081 |
| ln_cpi | | | | | 1.532 ** | 0.274 | 1.529 ** | 0.219 |
| | | | | | 0.335 | 0.449 | 0.335 | 0.449 |
| shock | | | | | | | -0.09 ** | -0.15 ** |
| | | | | | | | 0.026 | 0.029 |
| Year | | | | | | | | |
| 1997 | 0.5 ** | 0.501 ** | 0.54 ** | 0.518 ** | 0.093 | 0.439 ** | 0.102 | 0.469 ** |
| | 0.016 | 0.016 | 0.017 | 0.328 | 0.099 | 0.135 | 0.099 | 0.135 |
| 2000 | 0.994 ** | 0.982 ** | 1.02 ** | 1.005 ** | 0.458 ** | 0.905 ** | 0.462 ** | 0.932 ** |
| | 0.287 | 0.029 | 0.029 | 0.057 | 0.126 | 0.173 | 0.126 | 0.173 |
| cons | 10.494 ** | 10.936 ** | 8.158 ** | 8.676 ** | 0.512 | 7.316 ** | 0.57 | 7.306 ** |
| | 0.029 | 0.0215 | 0.163 | 0.517 | 1.688 | 2.303 | 1.689 | 2.299 |
| individual fixed-effect | N | Y | N | Y | N | Y | N | Y |
| clustered | Y | Y | Y | Y | Y | Y | Y | Y |
| sigma_u | 0.774 | 1.084 | 0.557 | 1.032 | 0.549 | 1.031 | 0.55 | 1.039 |
| sigma_e | 1.212 | 1.215 | 1.206 | 1.206 | 1.206 | 1.206 | 1.205 | 1.205 |
| rho | 0.289 | 0.445 | 0.176 | 0.423 | 0.172 | 0.422 | 0.172 | 0.426 |
| N | 13,845 | 13,845 | 13,742 | 13,742 | 13,742 | 13,742 | 13,739 | 13,739 |
| group | 4,615 | 4,615 | 4,615 | 4,615 | 4,615 | 4,615 | 4,615 | 4,615 |
| F-Test | 2,912.75 | 579.99 | 5,589.42 | 196.35 | 5,652.31 | 178.58 | 5,693.10 | 176.01 |

Table 10: Panel Regressions with Year Fixed Effects vs. TWFE

| ln_inc1_m | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **(a) nonag_main** | 1.053 ** | 0.666 ** | 0.653 ** | 0.646 ** |
| *pooled & no year fixed effect* | 0.027 | 0.03 | 0.029 | 0.029 |
| **(b) nonag_main** | 1.071 ** | 0.654 ** | 0.656 ** | 0.651 ** |
| *pooled, year fixed effect & clustered* | 0.033 | 0.034 | 0.034 | 0.033 |
| **(c) nonag_main** | 0.966 ** | 0.634 ** | 0.636 ** | 0.629 ** |
| *panel, year fixed effect & clustered* | 0.035 | 0.035 | 0.035 | 0.035 |
| **(d) nonag_main** | 0.515 ** | 0.446 ** | 0.446 ** | 0.438 ** |
| *panel, TWFE & clustered* | 0.072 | 0.069 | 0.069 | 0.069 |
| **differences before & after controlling for individual fixed effect:** | | | | |
| **(b) - (d)** | **0.556 ** ** | **0.208 ** ** | **0.21 ** ** | **0.213 ** ** |
| | **0.079** | **0.077** | **0.077** | **0.076** |
| **(c) - (d)** | **0.451 ** ** | **0.188 ** ** | **0.19 ** ** | **0.191 ** ** |
| | 0.080 | 0.077 | 0.077 | 0.077 |

Table 11: Estimated Selection Effects on the APG using TWFE

| ln_inc1_m *non-agriculture* | (1) | (2) | (3) | ln_inc1_m *agriculture* | (1) | (2) | (3) |
|---|---|---|---|---|---|---|---|
| age | -0.001 | -0.001 | -0.001 | age | -0.012 ** | -0.011 ** | -0.012 ** |
| | 0.002 | 0.002 | 0.002 | | 0.002 | 0.002 | 0.002 |
| gender | 0.401 ** | 0.402 ** | 0.403 ** | gender | 0.513 ** | 0.504 ** | 0.502 ** |
| | 0.032 | 0.032 | 0.032 | | 0.082 | 0.082 | 0.083 |
| educlevel2 | 0.351 ** | 0.352 ** | 0.352 ** | educlevel2 | 0.237 ** | 0.236 ** | 0.238 ** |
| | 0.013 | 0.013 | 0.013 | | 0.035 | 0.035 | 0.034 |
| marital_status | -0.08 ** | -0.078 ** | -0.077 ** | marital_status | 0.021 | 0.015 | 0.017 |
| | 0.214 | 0.021 | 0.021 | | 0.038 | 0.038 | 0.038 |
| urban | 0.177 ** | 0.176 ** | 0.176 ** | urban | 0.262 ** | 0.27 ** | 0.269 ** |
| | 0.03 | 0.03 | 0.03 | | 0.087 | 0.086 | 0.086 |
| wagedwork_main | 0.144 ** | 0.149 ** | 0.147 ** | wagedwork_main | 0.239 ** | 0.253 ** | 0.245 ** |
| | 0.034 | 0.034 | 0.034 | | 0.069 | 0.069 | 0.069 |
| ln_hrsworked_m | 0.307 ** | 0.309 ** | 0.309 ** | ln_hrsworked_m | 0.435 ** | 0.435 ** | 0.432 ** |
| | 0.028 | 0.028 | 0.028 | | 0.064 | 0.064 | 0.064 |
| nfarmbiz | 0.017 | 0.024 | 0.025 | nfarmbiz | 0.129 ** | 0.135 ** | 0.135 ** |
| | 0.026 | 0.026 | 0.026 | | 0.064 | 0.064 | 0.065 |
| farmbiz | -0.091 ** | -0.083 ** | -0.078 ** | farmbiz | -0.134 * | -0.112 | -0.094 |
| | 0.034 | 0.034 | 0.034 | | 0.072 | 0.073 | 0.074 |
| ruralborn | -0.105 ** | -0.102 ** | -0.103 ** | ruralborn | -0.138 | -0.122 | -0.117 |
| | 0.029 | 0.029 | 0.029 | | 0.108 | 0.107 | 0.107 |
| ln_cpi | ** | 1.447 ** | 1.451 ** | ln_cpi | | 2.599 ** | 2.629 ** |
| | | 0.331 | 0.331 | | | 0.713 | 0.715 |
| shock | | | -0.03 | shock | | | -0.104 * |
| | | | 0.026 | | | | 0.056 |
| wave | | | | wave | | | |
| 2 | 0.513 ** | 0.089 | 0.091 | 2 | 0.601 ** | -0.154 | -0.151 |
| | 0.018 | 0.098 | 0.099 | | 0.038 | 0.209 | 0.209 |
| 3 | 0.99 ** | 0.463 ** | 0.462 ** | 3 | 1.044 ** | 0.075 | 0.069 |
| | 0.028 | 0.125 | 0.125 | | 0.064 | 0.276 | 0.276 |
| cons | 8.796 ** | 1.567 | 1.553 | cons | 7.926 ** | -5.048 | -5.153 |
| | 0.166 | 1.666 | 1.665 | | 0.38 | 3.629 | 0.209 |
| clustered | Y | Y | Y | clustered | Y | Y | Y |
| N | 8,867 | 8,867 | 8.867 | N | 4,875 | 4,875 | 4,875 |
| $R^2$ | 0.333 | 0.335 | 0.335 | $R^2$ | 0.113 | 0.116 | 0.117 |

Table 12: Exclusion Restriction Evaluation in Outcome Equations

| nonag_main | (1) | | ag_main | (2) | |
|---|---|---|---|---|---|
| age | -0.003 | ** | age | 0.003 | ** |
| | 0.0004 | | | 0.0004 | |
| nfarmbiz | 0.256 | ** | nfarmbiz | -0.256 | ** |
| | 0.008 | | | 0.008 | |
| gender | -0.175 | ** | gender | 0.175 | ** |
| | 0.01 | | | 0.01 | |
| educlevel2 | 0.06 | ** | educlevel2 | -0.06 | ** |
| | 0.004 | | | 0.004 | |
| marital_status | -0.022 | ** | marital_status | 0.022 | ** |
| | 0.006 | | | 0.006 | |
| urban | 0.159 | ** | urban | -0.159 | ** |
| | 0.01 | | | 0.01 | |
| wagedwork_main | 0.153 | ** | wagedwork_main | -0.153 | ** |
| | 0.01 | | | 0.01 | |
| ln_hrsworked_m | 0.05 | ** | ln_hrsworked_m | -0.05 | ** |
| | 0.006 | | | 0.006 | |
| farmbiz | -0.289 | ** | farmbiz | 0.289 | ** |
| | 0.01 | | | 0.01 | |
| ruralborn | -0.016 | * | ruralborn | 0.016 | * |
| | 0.008 | | | 0.008 | |
| | | | | | |
| wave | | | wave | | |
| 2 | -0.0009 | | 2 | 0.0009 | |
| | 0.0057 | | | 0.0057 | |
| 3 | -0.028 | ** | 3 | 0.028 | ** |
| | 0.007 | | | 0.007 | |
| | | | | | |
| cons | 0.421 | ** | cons | 0.579 | ** |
| | 0.044 | | | 0.044 | |
| clustered | Y | | clustered | Y | |
| N | 13,742 | | N | 13,742 | |
| $R^2$ | 0.463 | | $R^2$ | 0.463 | |

Table 13: Exclusion Restriction Evaluation in Selection Equations

| ln_inc1_m non-agriculture | (1) | (2) | (3) | ln_inc1_m agriculture | (1) | (2) | (3) |
|---|---|---|---|---|---|---|---|
| *Step 2 Outcome* | | | | *Step 2 Outcome* | | | |
| gender | 0.444 ** | 0.445 ** | 0.446 ** | gender | 0.601 ** | 0.592 ** | 0.578 ** |
| | 0.029 | 0.029 | 0.029 | | 0.093 | 0.093 | 0.093 |
| educlevel2 | 0.339 ** | 0.339 ** | 0.339 ** | educlevel2 | 0.195 ** | 0.197 ** | 0.203 ** |
| | 0.011 | 0.011 | 0.011 | | 0.035 | 0.035 | 0.036 |
| marital_status | -0.074 ** | -0.072 ** | -0.071 ** | age | -0.011 ** | -0.01 ** | -0.01 ** |
| | 0.016 | 0.016 | 0.016 | | 0.002 | 0.002 | 0.002 |
| urban | 0.151 ** | 0.146 ** | 0.144 ** | urban | 0.173 ** | 0.188 ** | 0.197 ** |
| | 0.03 | 0.03 | 0.03 | | 0.094 | 0.094 | 0.094 |
| wagedwork_main | 0.127 ** | 0.127 ** | 0.125 ** | wagedwork_main | 0.182 ** | 0.193 ** | 0.187 ** |
| | 0.025 | 0.025 | 0.025 | | 0.081 | 0.084 | 0.081 |
| ln_hrsworked1_m | 0.296 ** | 0.297 ** | 0.296 ** | ln_hrsworked1_m | 0.402 ** | 0.404 ** | 0.403 ** |
| | 0.021 | 0.027 | 0.021 | | 0.051 | 0.05 | 0.05 |
| ln_cpi | | 1.513 ** | 1.514 ** | ln_cpi | | 2.601 ** | 2.627 ** |
| | | 0.289 | 0.289 | | | 0.657 | 0.657 |
| shock | | | -0.03 | shock | | | -0.096 * |
| | | | 0.024 | | | | 0.052 |
| wave | | | | wave | | | |
| 2 | 0.517 ** | 0.078 | 0.079 | 2 | 0.595 ** | -0.159 | -0.155 |
| | 0.027 | 0.089 | 0.089 | | 0.063 | 0.201 | 0.201 |
| 3 | 0.984 ** | 0.438 ** | 0.437 ** | 3 | 1.062 ** | 0.093 | 0.089 |
| | 0.027 | 0.109 | 0.109 | | 0.063 | 0.253 | 0.253 |
| cons | 8.888 ** | 1.345 | 1.349 | cons | 7.743 ** | -5.216 | -5.276 |
| | 0.134 | 1.449 | 1.449 | | 0.283 | 3.287 | 3.286 |
| *Step 1: Selection* | | | | *Step 1: Selection* | | | |
| age | -0.014 ** | -0.013 ** | -0.014 ** | age | 0.01 ** | 0.01 ** | 0.01 ** |
| | 0.001 | 0.001 | 0.001 | | 0.001 | 0.001 | 0.001 |
| nfarmbiz | 1.12 ** | 1.129 ** | 1.128 ** | fambiz | 1.109 ** | 1.11 ** | 1.103 ** |
| | 0.03 | 0.031 | 0.031 | | 0.03 | 0.03 | 0.03 |
| gender | -0.853 ** | -0.853 ** | -0.846 ** | gender | 0.904 ** | 0.904 ** | 0.903 ** |
| | 0.037 | 0.037 | 0.037 | | 0.036 | 0.036 | 0.036 |
| educlevel2 | 0.257 ** | 0.261 ** | 0.265 ** | educlevel2 | -0.309 ** | -0.31 ** | -0.311 ** |
| | 0.129 | 0.014 | 0.014 | | 0.014 | 0.014 | 0.014 |
| marital_status | -0.074 ** | -0.074 ** | -0.067 ** | marital_status | 0.134 ** | 0.134 ** | 0.132 ** |
| | 0.019 | 0.019 | 0.019 | | 0.019 | 0.019 | 0.019 |
| urban | 0.933 ** | 0.934 ** | 0.926 ** | urban | -0.689 ** | -0.689 ** | -0.689 ** |
| | 0.031 | 0.031 | 0.031 | | 0.032 | 0.032 | 0.032 |
| wagedwork_main | 0.868 ** | 0.872 ** | 0.853 ** | wagedwork_main | -0.087 ** | -0.087 ** | -0.084 ** |
| | 0.032 | 0.032 | 0.032 | | 0.03 | 0.03 | 0.03 |
| ln_hrsworked1_m | 0.244 ** | 0.242 ** | 0.237 ** | ln_hrsworked1_m | -0.195 ** | -0.195 ** | -0.194 ** |
| | 0.024 | 0.024 | 0.024 | | 0.024 | 0.024 | 0.024 |
| ruralborn | -0.261 ** | -0.255 ** | -0.252 ** | ruralborn | 0.201 ** | 0.201 ** | 0.2 ** |
| | 0.042 | 0.042 | 0.042 | | 0.043 | 0.043 | 0.043 |
| ln_cpi | | -0.226 ** | -0.201 ** | ln_cpi | | 0.011 ** | 0.005 |
| | | 0.086 | 0.087 | | | 0.085 | 0.085 |
| shock | | | -0.182 ** | shock | | | 0.047 * |
| | | | 0.028 | | | | 0.029 |
| cons | -0.979 ** | 0.149 | 0.099 | cons | -0.523 ** | -0.579 | -0.561 |
| | 0.156 | 0.459 | 0.461 | | 0.158 | 0.458 | 0.459 |
| *Inverse Mills Ratio* | | | | *Inverse Mills Ratio* | | | |
| λ | -0.138 ** | -0.142 ** | -0.143 ** | λ | 0.245 ** | 0.226 ** | 0.202 * |
| | 0.052 | 0.052 | 0.052 | | 0.105 | 0.105 | 0.106 |
| rho | -0.133 | -0.137 | -0.138 | rho | 0.139 | 0.129 | 0.116 |
| sigma | 1.038 | 1.036 | 1.036 | sigma | 1.756 | 1.752 | 1.749 |
| N | 13,742 | 13,742 | 13,739 | N | 13,742 | 13,742 | 13,739 |
| Selected | 8,867 | 8,867 | 8,864 | Selected | 4,875 | 4,875 | 4,875 |
| Wald chi2 | 3,777.43 ** | 3,868.35 ** | 3,867.37 ** | Wald chi2 | 523.65 ** | 540.57 ** | 544.21 ** |

| *Exclusion Restriction (Nonag)* |
|---|
| *age* |
| *nfarmbiz* |

| *Exclusion Restrictions (Ag)* |
|---|
| *marital_status* |
| *farmbiz* |
| *ruralborn* |

Table 14: Heckman Two-Step Estimation (Pooled IFLS 1-3 Waves)

| ln_inc1_m<br>*pooled, year fixed effect, Heckman twostep* | (1) | (2) | (3) |
|---|---|---|---|
| **nonag** | | | |
| *selection effect ( $\lambda^n$ )* | -0.138 ** | -0.142 ** | -0.143 ** |
| | 0.052 | 0.052 | 0.052 |
| **ag** | | | |
| *selection effect ($\lambda^a$ )* | 0.245 ** | 0.226 ** | 0.202 * |
| | 0.105 | 0.105 | 0.106 |
| **observed APG** | 0.654 ** | 0.656 ** | 0.651 ** |
| *pooled, year fixed-effect & clustered* | 0.034 | 0.034 | 0.033 |
| **Selection effect explains observed APG** | | | |
| ***nonag*** | ***21.1%*** | ***21.6%*** | ***22.0%*** |
| ***ag*** | ***37.5%*** | ***34.5%*** | ***31.0%*** |

Table 15: Heckman Selection Effect on APG (Pooled)

| xthechman | (1) | (2) |
|---|---|---|
| **Outcome Equation** | | |
| **ln_inc1_m** | | |
| **gender** | 0.451 ** | 0.651 ** |
| | 0.041 | 0.033 |
| **educlevel2** | 0.375 ** | 0.267 ** |
| | 0.014 | 0.012 |
| **marital_status** | -0.02 | -0.067 ** |
| | 0.019 | 0.019 |
| **urban** | 0.204 ** | -0.034 |
| | 0.042 | 0.03 |
| **wagedwork_main** | 0.056 * | -0.033 |
| | 0.029 | 0.028 |
| **lm_hrworked1_m** | 0.258 ** | 0.206 ** |
| | 0.023 | 0.021 |
| **ln_cpi** | | 2.486 ** |
| | | 0.059 |
| **cons** | 9.241 ** | -2.723 ** |
| | 0.156 | 0.329 |
| **Selection Equation** | | |
| **nonag_main** | | |
| **age** | -0.029 ** | -0.024 ** |
| | 0.003 | 0.002 |
| **nfarmbiz** | 1.303 ** | 1.21 ** |
| | 0.054 | 0.046 |
| **gender** | -1.319 ** | -1.466 ** |
| | 0.094 | 0.075 |
| **educlevel2** | 0.413 ** | 0.357 ** |
| | 0.029 | 0.025 |
| **marital_status** | -0.112 ** | -0.117 ** |
| | 0.041 | 0.035 |
| **urban** | 1.787 ** | 1.553 ** |
| | 0.075 | 0.059 |
| **wagedwork_main** | 1.128 ** | 0.944 ** |
| | 0.058 | 0.051 |
| **ln_hrsworked1_m** | 0.347 ** | 0.312 ** |
| | 0.041 | 0.035 |
| **cons** | -0.909 ** | -0.197 |
| | 0.283 | 0.564 |
| **var(e.ln_inc1_m)** | 0.972 | 0.845 |
| | 0.019 | 0.015 |
| **var(ln_inc1_m[i])** | 0.266 | 0.412 |
| | 0.019 | 0.021 |
| **var(nonag_main[i])** | 2.45 | 2.314 |
| | 0.17 | 0.148 |
| **corr(e.nonag_main, e.ln_inc1_m)** | -0.235 ** | -0.973 |
| | 0.093 | . |
| **corr(nonag_main[i], ln_inc1_m[i])** | 0.166 ** | -0.241 ** |
| | 0.08 | 0.148 |
| **N** | 13,742 | 13,742 |
| **Selected** | 8,867 | 8,867 |
| **# of groups** | 4,615 | 4,615 |
| **Wald chi2** | 1,602.49 ** | 3,124.52 ** |

***Exclusion Restriction (Nonag)***
*age*
*nfarmbiz*

***Note: Convergence not achieved in (1) and (2)**

Table 16: Panel Heckman Estimation using **xtheckman** (IFLS Waves 1–3)

# A Appendix: Background on Suri's Empirical Approach

This appendix provides background on the empirical approach of Suri (2011), which this paper adapts to study agricultural productivity gaps. While the main text explains how the framework is modified for the APG context, this appendix reviews Suri's original application to technology adoption in Kenya and outlines the underlying logic of the correlated random coefficient (CRC) model. The purpose is to provide readers less familiar with this method with a clear understanding of its origins, intuition, and technical foundations. Readers already acquainted with Suri's work may skip directly to Section 3, where the adapted framework is presented.

To address the two challenges identified in the previous subsection, I adopt the Correlated Random Coefficient (CRC) model, as employed by Tavneet Suri (2011), when explaining the low adoption rates of hybrid seeds in Kenya, despite their high yields. This empirical approach allows me to estimate individual sorting based on the sector-specific unobserved abilities without imposing parametric distributional assumptions.

In Suri's empirical strategy, expected potential returns determine each farmer's adoption decision on hybrid seeds, which follows Heckman and Vytlacil (1998) under the generalized Roy's model framework (Roy, 1951); what's more, this method does not assume any functional form for unobserved heterogeity by exploring the fact that farmer's adoption choice history contains the information on farmer's net benefits from using hybrid seeds, which is in the spirit of Chamberlain's estimation of fixed-effect in panel data (Chamberlain, 1982, 1984). This framework consists of two key appealing features: First, it considers each farmer's net benefit as a deviation from the average net benefit of hybrid seed adoption, which explicitly models heterogeneity. Second, it exploits the revealed comparative advantages that can be projected by hybrid seeds adoption trajectories, which allows for estimating the selection effect without distributional assumptions. As a result, Suri's (2011) empirical approach is preferable for tackling the two challenges that the research question of this paper must overcome.

# B Selection Effect and Map to the Classic Roy Model

This appendix derives the link between the CRC model selection parameter $\beta$ and the selection terms in the classic Roy framework represented in Borjas (1987). To further illustrate what exactly the selection effect, $\beta$, is measured by this model, I first map it to different types of selection in the classic Roy model framework. Then, I discuss how the selection is determined by the variance of the latent skills in each sector and their correlations.

In this model, individual sorting based on the unobserved comparative advantage is summarized by the structural parameter $\beta$. Let $\sigma_n = Var(\theta_i^n)$, $\sigma_a = Var(\theta_i^a)$, and $\sigma_{na} = COV(\theta_i^n, \theta_i^a)$. Then, $b_n$ and $b_a$ as coefficients for equations (18) and (19) take form, as shown in equation (37). Therefore, in equation (38), the numerator $\sigma_n^2 - \sigma_{na}$ is the covariance-adjusted dispersion of latent absolute advantage in non-agriculture; the denominator $\sigma_{na} - \sigma_a^2$ is the analogous term for agriculture. Intuitively, the selection effect, $\beta$, measures the relative, covariance-adjusted dispersion of unobserved absolute advantages in non-agriculture relative to agriculture — i.e., how much more the non-agriculture sector loads on the latent skill variation once the skill correlation between the two sectors is netted out.

$$\begin{aligned}
\beta &\equiv \frac{b_n}{b_a} - 1 \\
&= \frac{(\sigma_n^2 - \sigma_{na})/(\sigma_n^2 + \sigma_a^2 - 2\sigma_{na})}{(\sigma_{na} - \sigma_a^2)/(\sigma_n^2 + \sigma_a^2 - 2\sigma_{na})} - 1 \qquad (37) \\
&= \frac{\sigma_n^2 - \sigma_{na}}{\sigma_{na} - \sigma_a^2} - 1 \qquad (38)
\end{aligned}$$

Under the assumption of a joint normal distribution for the latent skills, the inverse Mills ratio (IMR) can be derived as a selection-bias factor that summarizes the selectivity of the sample and adjusts the results by serving as a proxy for latent abilities. Therefore, the coefficient of the IMR in a regression represents the selection effect that arises when a joint-normal distribution is imposed on the latent skills.

Borjas (1987) studies the earnings and immigration choices in the United States and assumes a joint normal distribution for latent skills between two countries. Under the classic Roy's model framework, Borjas captures the selection effect and bias correction by $Q_1$ and $Q_0$, which are defined as the differential earnings between the average immigrants and the average in the country of destination (referred to as Country 1) and in the country of origin

(referred to as Country 0), respectively. As shown in the equations (39) and (40), $\frac{\phi(z)}{1-\Phi(z)}$ is IMR and $1/\sigma_\nu$ is the scaling factor for IMR as it is transformed to the standard normal when applying the closed-form solution. Hence, the selection effect in Borjas' paper is captured by the coefficients, $\sigma_0\sigma_1(\rho_{0,1} - \frac{\sigma_0}{\sigma_1})$ and $\sigma_0\sigma_1(\frac{\sigma_1}{\sigma_0} - \rho_{0,1})$ for the country of origin and destination, respectively.

In Borjas (1987), the Roy framework is applied to immigration: country 0 represents the origin country and country 1 the destination (United States). In my APG setting, these roles naturally map to agriculture ($a$) as the origin sector and non-agriculture ($n$) as the destination. Borjas defines two selection terms, $Q_0$ and $Q_1$, which measure differential earnings between immigrants and sectoral averages in the origin and destination, respectively:

$$Q_1 = \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\frac{\sigma_1}{\sigma_0} - \rho_{0,1}\right)\left(\frac{\phi(z)}{1-\Phi(z)}\right) \tag{39}$$

$$Q_0 = \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\rho_{0,1} - \frac{\sigma_0}{\sigma_1}\right)\left(\frac{\phi(z)}{1-\Phi(z)}\right) \tag{40}$$

These two coefficients can be further expressed as equations (41) and (42) by multiplying $\sigma_0\sigma_1$ to the terms within the first bracket in each equation (39) and (40), respectively. The country 0 in Borjas' paper corresponds to the agricultural sector, and country 1 represents the nonagricultural sector in my setting. Notably, $\beta$ in my model includes the selection effect formulation in the classic Roy model. Instead of measuring selection in each sector separately, $\beta$ refers to the differences in the selection effects between two sectors, with the selection effect in the agricultural sector serving as a benchmark.

$$\begin{aligned}\sigma_0\sigma_1\left(\frac{\sigma_1}{\sigma_0} - \rho_{0,1}\right) &= \sigma_1^2 - \rho_{0,1}\sigma_0\sigma_1 \\ &= \sigma_1^2 - \sigma_{0,1}\end{aligned} \tag{41}$$

$$\begin{aligned}\sigma_0\sigma_1\left(\rho_{0,1} - \frac{\sigma_0}{\sigma_1}\right) &= \rho_{0,1}\sigma_0\sigma_1 - \sigma_0^2 \\ &= \sigma_{0,1} - \sigma_0^2\end{aligned} \tag{42}$$

Analogously, I define two distribution-free counterparts in my model: $\Delta_a$ for agriculture and $\Delta_n$ for non-agriculture. These capture the differential earnings between sectoral switchers and the corresponding sectoral averages, just as $Q_0$ and $Q_1$ do in Borjas. The critical difference is that $\Delta_a$ and $\Delta_n$ do not rely on a joint-normality assumption on latent skills.

Instead, they are constructed directly from the Roy-style choice problem. Formal definitions and derivations are provided in Appendix C, but for interpretation it suffices to note that $(\Delta_a, \Delta_n)$ serve as the natural analogues to $(Q_0, Q_1)$ in a distribution-free setting.

Under a weak assumption on monotone sorting, the selection effects in this paper can be mapped into the positive selection, negative selection, and refugee cases illustrated in Borjas (1987). The derivation of the conditions for the different types of selection effect $\beta$ in Appendix C. The summary of the conditions compared to those in Borjas (1987) is in Table 17

Table 17: Selection Types in Roy vs. CRC Model

| **Selection Type** | **Roy Model (Borjas, 1987)** | **CRC (Lemieux, 1998)** |
|---|---|---|
| **Positive Selection** | $\rho_{0,1} > \frac{\sigma_0}{\sigma_1},\ \sigma_1 > \sigma_0$<br>$Q_0 > 0,\ Q_1 > 0$<br>In both upper tails | $\beta > 0 \iff \rho_{na} > \frac{\sigma_a}{\sigma_n},\ \sigma_n > \sigma_a$<br>$\Delta_a > 0,\ \Delta_n > 0$<br>In both upper tails |
| **Negative Selection** | $\rho_{0,1} < \frac{\sigma_0}{\sigma_1},\ \sigma_0 > \sigma_1$<br>$Q_0 < 0,\ Q_1 < 0$<br>In both lower tails | $-1 < \beta < 0 \implies \rho_{na} < \frac{\sigma_a}{\sigma_n}$<br>$\Delta_a < 0,\ \Delta_n < 0$<br>In both lower tails |
| **Refugee Selection** | $\rho_{0,1} < \min\left\{\frac{\sigma_0}{\sigma_1}, \frac{\sigma_1}{\sigma_0}\right\}$<br>$Q_0 < 0,\ Q_1 > 0$ | $\beta < -1 \implies \rho_{na} < \frac{\sigma_a}{\sigma_n}$<br>$\Delta_a < 0,\ \Delta_n > 0$ |
| **Null on One-Side** | No such case | $\beta = -1 \implies \rho_{na} < \frac{\sigma_a}{\sigma_n}$<br>$\Delta_a < 0,\ \Delta_n = 0$ |

*Notes:* In Borjas' paper, 1 refers to the country of destination, 0 to the country of origin. In this paper, $n$ is non-agriculture and $a$ is agriculture. $\Delta_a$ and $\Delta_n$ are the differential earnings between the average of those who choose to switch to non-agriculture and the average in agriculture and non-agriculture, respectively.

When $\beta > 0$, indicating positive selection—individuals are drawn from the upper tail of agriculture and fall in the upper tail of non-agriculture. However, when $\beta < 0$, there are three cases: (1) If $-1 < \beta < 0$, workers are drawn from the lower tail in agriculture and contribute to the lower tail in non-agriculture, hence negative selection. (2) If $\beta < -1$, it is drawn from the lower tail in agriculture, but those workers earn higher wages than the average workers in non-agriculture. This case would occur when the negative selection is

sufficiently large. (3) If $\beta = -1$, the switchers from agriculture earn the same as the average workers in non-agriculture, which is not in Borjas (1987).

Essentially, the selection effect and unobserved comparative advantages formulated by Lemieux (1998) capture the equivalent selection effect from unobserved heterogeneity that affects individuals' choices, as modelled in the classic Roy choice framework. The difference is that this selection effect is not the coefficient of IMR, thereby allowing for the flexibility to estimate underlying unobserved comparative advantages, as opposed to assuming a specific functional form.

# C Appendix: Sector Choice and Differential Returns

This appendix lays out the sector choice individual faces in the Roy's model framework and then defines the differential returns sector switchers relative to the average earnings in each sector, $\Delta_n$ and $\Delta_a$.

I start from the potential earnings representation in Section 3.3 and make the individual choice rule explicit under the Roy model choice framework. We then define the differential earnings objects $\Delta_n$ and $\Delta_a$ and show how they correspond to Borjas's $Q_1$ and $Q_0$.

**Step 1. Potential earnings in each sector.** Recall the log potential earnings for individual $i$ in sector $j \in \{n, a\}$ at time $t$ (see Eqs. (20)–(21)):

$$w_{it}^n = \delta_t^n \; + \; (1+\beta)\,\theta_i \; + \; \tau_i \; + \; X_{it}\gamma^n \; + \; \xi_{it}^n, \tag{26}$$

$$w_{it}^a = \delta_t^a \; + \; \theta_i \; + \; \tau_i \; + \; X_{it}\gamma^a \; + \; \xi_{it}^a. \tag{27}$$

Here, $\theta_i$ is the unobserved comparative advantage (sector-relevant, time-invariant), $\tau_i$ is the sector-invariant component (irrelevant for choice), $\beta$ captures the relative loading of latent skills across sectors, $X_{it}$ are observables, and $\xi_{it}^j$ are transitory shocks with zero conditional mean.

**Step 2. Sectoral choice rule.** Individual $i$ chooses non-agriculture ($D_{it} = 1$) iff $w_{it}^n \geq w_{it}^a$. Using (26)–(27):

$$D_{it} = 1 \quad \Longleftrightarrow \quad (\delta_t^n - \delta_t^a) \; + \; \beta\, \theta_i \; + \; X_{it}(\gamma^n - \gamma^a) \; + \; (\xi_{it}^n - \xi_{it}^a) \; \geq \; 0. \tag{43}$$

Conditional on observables, sorting is governed by the comparative advantage term $\beta\, \theta_i$; the common component $\tau_i$ cancels in the difference.

**Step 3. Differential earnings objects.** Define selection premia as differences between conditional and unconditional sector means:

$$\Delta_n \equiv E[w^n \mid D = 1, X] - E[w^n \mid X]\,, \tag{44}$$

$$\Delta_a \equiv E[w^a \mid D = 0, X] - E[w^a \mid X]\,. \tag{45}$$

Intuitively, $\Delta_n$ measures how the earnings of those who select into non-agriculture differ from the non-agriculture sector mean; $\Delta_a$ is the analogous object for agriculture.[2]

Using (26)–(27) and the choice rule (43), we can write

$$\Delta_n = (1 + \beta)\,\Big(E[\theta_i \mid D = 1, X] - E[\theta_i \mid X]\Big), \tag{46}$$

$$\Delta_a = \qquad\quad \Big(E[\theta_i \mid D = 0, X] - E[\theta_i \mid X]\Big). \tag{47}$$

Thus, both selection premia are linear in the *comparative advantage* component; $\beta$ scales the non-agriculture premium relative to agriculture.

**Step 4. Connection to Borjas (1987).** In the Borjas–Roy migration setting, country 1 (destination) and 0 (origin) abilities are jointly normal with variances $\sigma_1^2, \sigma_0^2$ and covariance $\sigma_{01}$. The standard selection corrections are

$$Q_1 \;=\; \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\frac{\sigma_1}{\sigma_0} - \rho_{01}\right)\lambda(z), \qquad Q_0 \;=\; \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\rho_{01} - \frac{\sigma_0}{\sigma_1}\right)\lambda(z), \tag{48}$$

where $\lambda(z) = \phi(z)/[1 - \Phi(z)]$ is the inverse Mills ratio and $\sigma_\nu > 0$ is a scale from the latent index. The *coefficients on the IMR* are

$$\underbrace{\sigma_1^2 - \sigma_{01}}_{\text{destination (1)}}\,, \qquad \underbrace{\sigma_{01} - \sigma_0^2}_{\text{origin (0)}}. \tag{49}$$

[2] We suppress $t$ and condition on $X$ to lighten notation. The transitory shocks $\xi_{it}^j$ integrate out by zero conditional mean.

Identify agriculture with origin $(0 \leftrightarrow a)$ and non-agriculture with destination $(1 \leftrightarrow n)$, so that

$$\sigma_1^2 - \sigma_{01} \;\longleftrightarrow\; \sigma_n^2 - \sigma_{na}, \qquad \sigma_{01} - \sigma_0^2 \;\longleftrightarrow\; \sigma_{na} - \sigma_a^2. \tag{50}$$

From Section 3.2, the CRC selection parameter satisfies

$$\beta + 1 \;=\; \frac{\sigma_n^2 - \sigma_{na}}{\sigma_{na} - \sigma_a^2}. \tag{51}$$

Comparing (49)–(50) with (51) yields the algebraic identity

$$\boxed{\beta + 1 \;=\; \frac{\text{IMR coefficient at destination (non-agriculture)}}{\text{IMR coefficient at origin (agriculture)}}} \tag{52}$$

up to a common positive scale that cancels in the ratio.[3]

Finally, observe that (46)–(47) are distribution-free analogues of Borjas's $Q_1, Q_0$:

$$\Delta_n \;\leftrightarrow\; Q_1, \qquad \Delta_a \;\leftrightarrow\; Q_0, \tag{53}$$

where the CRC framework replaces the IMR with conditional means of $\theta_i$ governed by the monotone selection rule (43).

# D Appendix: Further Analysis of $\beta$

This appendix examines how the variance and correlation of latent abilities between sectors determine selection. I can rewrite $\beta$ in equation (38) into the expression in equation (54). Define $r$ as the ratio of the standard deviation of unobserved absolute advantages between two sectors, as in equation (55), which represents how widely spread the absolute advantages in nonagriculture are relative to those in agriculture. Then, $\beta$ can be further expressed as in equation (56), where $\rho$ is the correlation coefficient of the covariance of absolute advantages between agriculture and nonagriculture, taking values $-1 \leq \rho \leq 1$.

[3] Any multiplicative factors such as $1/\sigma_\nu$ and $\lambda(z)$ in (48) are common across $Q_1, Q_0$ conditional on the selection index and therefore cancel in the ratio.

$$\begin{aligned}\beta &= \frac{\sigma_n^2 - \sigma_{na}}{\sigma_{na} - \sigma_a^2} - 1 \\ &= \frac{(\sigma_n^2 - \sigma_{na}) - (\sigma_{na} - \sigma_a^2)}{\sigma_{na} - \sigma_a^2} \\ &= \frac{\sigma_n^2 + \sigma_a^2 - 2\sigma_{na}}{\sigma_{na} - \sigma_a^2}\end{aligned} \tag{54}$$

$$r \equiv \frac{\sigma_n}{\sigma_a} \tag{55}$$

$$\beta = \frac{\sigma_n^2 + \sigma_a^2 - 2\sigma_{na}}{\sigma_{na} - \sigma_a^2} = \frac{r^2 + 1 - 2\rho r}{\rho r - 1} \tag{56}$$

$$\frac{\partial \beta}{\partial \rho} = \frac{r(1 - r^2)}{(\rho r - 1)^2} \tag{57}$$

Given a fixed $r$ in the market, meaning how spread unobserved absolute advantages in nonagriculture relative to agriculture, the partial derivatives of $\beta$ with respect to $\rho$, shown in equation (57), indicate four cases:

(i) If $r > 1$, then $\beta$ decreases in $\rho$;

(ii) If $0 < r < 1$, then $\beta$ increases in $\rho$;

(iii) If $r = 0$, then $\beta = -1$ for all feasible $\rho$;

(iv) If $r = 1$, then $\beta = -2$ for all feasible $\rho < 1$.

Since $r$ is the ratio of standard deviations, its value will always be non-negative. In the case (i), this ratio of spread between latent absolute advantages is sufficiently large ($r > 1$). Holding $r$ constant, as those latent skills become more similar across sectors (an increase in $\rho$), it dulls the comparative advantage in the respective sector that workers enjoy and weakens the selection effect; hence, $\beta$ falls. In the case (ii), the dispersion of latent absolute advantages is alike in two sectors ($0 < r < 1$). For a fixed $r$, as the latent skills become more transferable across two sectors (a rising $\rho$), it raises the comparative advantage in the sector where workers don't possess and strengthens the selection effect; thus, $\beta$ rises. The remaining

two cases are not very interesting. Both cases will make the selection effect degenerate into a constant number, where case (iii) is when absolute advantages in nonagriculture are without any dispersion, and case (iv) is when the spread of latent skills is precisely the same in two sectors.

Furthermore, the values of the selection effect $\beta$ can provide some policy insights. When $\beta > 0$, the selection effect is bounded from below, $\beta \in [r-1, +\infty)$, which can only occur in the positive selection case. The minimum value of the selection effect is the differential spread of latent skills between sectors after removing the correlation coefficient at the upper boundary ($\rho = 1$). When $\beta < 0$, there are two cases: (i) If the spreads of latent skills across sectors are sufficiently large ($r > 1$), the selection flips to a negative value when passing the threshold, $\frac{1}{r}$. In this case, $\beta$ is bounded from above, $\beta \in (-\infty, -(r+1)]$, where the least negative value is at two skills perfectly negatively correlated ($\rho = -1$). (ii) If the spreads of latent skills are similar ($0 < r < 1$), the selection is bounded, $\beta \in [-(r+1), r-1]$.

The threshold $\rho = \frac{1}{r}$ is where the positive and negative selection switches. In the environment where the spread in two sectors is sufficiently large, when the correlation of latent skills approaches the threshold, the selection effect becomes a large positive number on the right side and a large negative number on the left side. In the environment, the dispersions in latent skills are similar, and the selection effect is tightly bound by the correlation coefficient $\rho \in [-1, 1]$. Policies promoting transferable skills can influence the correlations, and interventions enhancing educational levels may bridge the disparities in latent skills between the two sectors. As the selection effect intensifies at the threshold, this insight can help design policies that align the selection effect with the main policy goals, thereby avoiding unintended consequences arising from the selection effect.

# E Appendix: Recovering Structural Parameters

This appendix is to demonstrate how structure parameter unobserved comparative advantage, $\theta_i$, is recovered in a simplified two-period no covariant model. $\theta_i$ is unobserved in the data. I will, now, demonstrate how the structural parameters of interest can be recovered without imposing distributional assumptions. Following the procedure proposed by

Suri (2011), I present the recovery of structural parameters in the simplest setting, without covariates, over two periods, as expressed in equation (58).

$$w_{it} = \eta + \alpha D_{it} + \theta_i + \beta \theta_i D_{it} + u_{it} \tag{58}$$

where $\delta_t^a = \eta \quad \forall t$, $\alpha \equiv \delta_t^n - \delta_t^a \quad \forall t$, and $u_{it} \equiv \tau_i + \epsilon_{it}$.

I can do this because structural parameters $\beta$ and $\theta_i$ do not enter covariates in the main estimation equation (23). First, I disentangle the dependency between $\theta_i$ and $D_{it}$ by linearly projecting $\theta_i$ onto the entire history of sector choices and their interaction terms. This method was first developed by Chamberlain (Chamberlain, 1982, 1984) to estimate individual unobserved fixed effects in panel data. Later, Suri's (2011) generalized Chamberlain's fixed-effect estimation by including interactions of choice histories to purge the dependency between unobserved abilities and individual choices fully. Chamberlain and Suri treat this as a purely technical step to separate $\theta_i$ into two parts: one is related to sectoral choice ($D_{it}$), and the remainder is orthogonal to the sectoral choices. Since individuals choose, $D_{it}$, is a dummy variable, the projection $\theta_i$ onto a complete history and interaction terms will be a saturated model to purge the correlation between $\theta_i$ and $D_{it}$, which is formally expressed in the equation (59). However, I interpret the equation (59) as an individual choice trajectories that reveal her unobserved comparative advantages, $\theta_i$.

$$\theta_i = \lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_3 D_{i1} D_{i2} + \nu_i \tag{59}$$

Next, I substitute equation (59) into the wage equation (58) to obtain the log earnings for each period as a function of choice histories and their interactions, see equations (60) and (61).

$$\begin{aligned} w_{i1} = {} & (\eta + \lambda_0) + (\lambda_1(1+\beta) + \alpha + \beta\lambda_0) D_{i1} \\ & + \lambda_2 D_{i2} + (\lambda_3(1+\beta) + \beta\lambda_2) D_{i1} D_{i2} + (\nu_i + \beta\nu_i D_{i1} + u_{i1}) \end{aligned} \tag{60}$$

$$
\begin{aligned}
w_{i2} = {} & (\eta + \lambda_0) + \lambda_1 D_{i1} \\
& + (\lambda_2(1+\beta) + \alpha + \beta\lambda_0) D_{i2} + (\lambda_3(1+\beta) + \beta\lambda_1) D_{i1} D_{i2} \\
& + (\nu_i + \beta\nu_i D_{i2} + u_{i2})
\end{aligned} \tag{61}
$$

Since sector choices are observed in each period, I can run a reduced-form regression of earnings on the choice history and their interactions in this stacked system of equations (60) and (61). To simplify the coefficients in the reduced form regression in equations (60) and (61), I can rewrite them as equations (62) and (63).

$$w_{i1} = \eta + \phi_1 D_{i1} + \phi_2 D_{i2} + \phi_3 D_{i1} D_{i2} + e_{it} \tag{62}$$

$$w_{i2} = \eta + \phi_4 D_{i1} + \phi_5 D_{i2} + \phi_6 D_{i1} D_{i2} + e_{it} \tag{63}$$

The reduced form regression of each period earnings on the entire history of the sector choice, including interaction terms across periods, will obtain reduced form coefficients $\phi_1$, $\phi_2$, $\phi_3$, $\phi_4$, $\phi_5$, and $\phi_6$. Combining equations (60) to (63), the information that I have learned from the reduced form coefficients can yield the following equations:

$$\phi_1 = \lambda_1(1+\beta) + \alpha + \beta\lambda_0 \tag{64}$$

$$\phi_2 = \lambda_2 \tag{65}$$

$$\phi_3 = \lambda_3(1+\beta) + \beta\lambda_2 \tag{66}$$

$$\phi_4 = \lambda_1 \tag{67}$$

$$\phi_5 = \lambda_2(1+\beta) + \alpha + \beta\lambda_0 \tag{68}$$

$$\phi_6 = \lambda_3(1+\beta) + \beta\lambda_1 \tag{69}$$

Equations (64) to (69) explicitly express the reduced-form coefficients as functions of underlying structural parameters. Solving this system of equations, I can estimate five underlying parameters: $\lambda_1$, $\lambda_2$, $\lambda_3$, $\alpha$, and $\beta$. Under the condition that $\lambda_1 \neq \lambda_2$, the parameter $\beta$ is identified. The first objective is to obtain the structural parameter $\beta$, which captures the extra returns of comparative advantages in nonagriculture, i.e., the selection effect. Then, the

second task is to estimate the distribution of comparative advantage $\theta_i$ by using equation (59) and normalizing $\sum \theta_i = 0$. Specifically, I can obtain $\lambda_0$ by using $\lambda_0 = -\lambda_1 \overline{D_{i1}} - \lambda_2 \overline{D_{i2}} - \lambda_3 \overline{D_{i1} D_{i2}}$. It is noted that $\lambda_1$, $\lambda_2$, and $\lambda_3$ can be obtained by solving the system of equations, and sectoral choices are observed in the data. Once $\lambda_0$ is available, I can use (59), $\lambda$'s, and $D_{it}$'s to estimate the distribution of the unobserved comparative advantage, $\theta_i$.

# F Appendix: Revealed Comparative Advantages

This appendix shows that the projection of $\theta_i$ can be interpreted as revealed comparative advantages, which contain rich information empirically in the context of Indonesia.

Although Chamberlain (1982, 1984) and Suri (2011) explicitly emphasize that equation (59) is primarily a technical device for eliminating correlation between $\theta_i$ and choice variable $D_{it}$, it can equivalently be interpreted as a regression of the latent comparative advantage $\theta_i$ on indicators of the choice trajectory (choices at each $t$ and their interaction). In this formulation, the fitted values $\hat{\theta}_i$ represent the component of individual underlying comparative advantage explained by the trajectory, thereby providing an empirical measure of unobserved heterogeneity across groups.

Drawing loosely on the intuition of revealed preference theory (Samuelson, 1938, 1948), sectoral choices can be interpreted as revealing information about underlying comparative advantages. In this context, individuals implicitly conduct a cost–benefit analysis, where potential earnings in each sector represent are the benefits, and constraints such as schooling, time, and ability represent the costs. Over three waves, each agent's sequence of sectoral choices yields one of eight possible trajectories ($2^3 = 8$), reflecting how unobserved comparative advantages shape decisions. Although this analogy to revealed preference is only suggestive rather than a formal extension, the observed choice histories and their interactions provide an empirical basis for capturing latent abilities. I therefore refer to the fitted values from this procedure, $\hat{\theta}_i$, as revealed comparative advantages.

Figures 10 and 11 illustrate the composition of education levels and waged work across the eight trajectory groups, denoted by $t000, t001, t010, t100, t110, t101, t011, t111$, where 1 indicates non-agriculture and 0 indicates agriculture in a given period. For example, $t000$

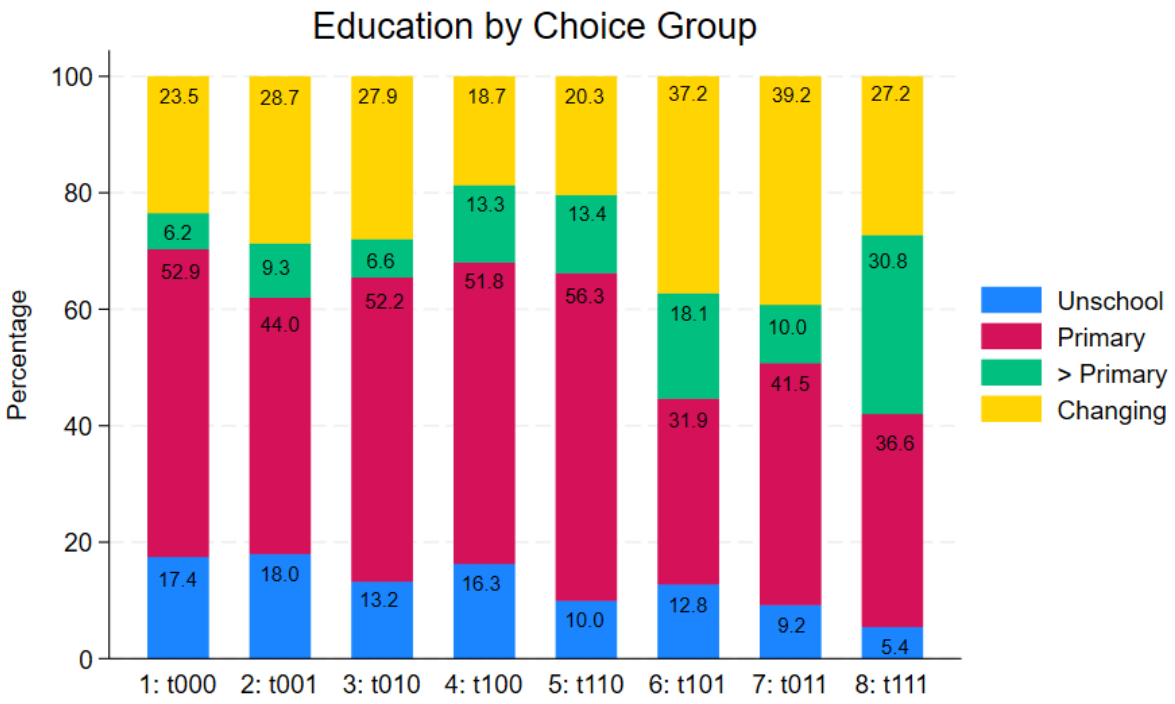


Figure 10

corresponds to staying in agriculture in all three waves, while $t111$ corresponds to remaining in non-agriculture. Figure 10 shows clear contrasts: non-agricultural stayers ($t111$) are disproportionately drawn from individuals with education beyond primary school, whereas agricultural stayers ($t000$) contain a higher share of unschooled individuals. Sector switchers, by contrast, display larger shifts in educational attainment across waves. A parallel pattern emerges in Figure 11: switchers exhibit greater transitions between self-employment and waged work, while stayers tend to remain more stable in their employment types.

Figures 12 and 13 provide additional evidence that choice trajectories reveal information about underlying comparative advantages. A well-documented puzzle in the APG literature using IFLS data is that individuals who move from non-agriculture to agriculture appear to experience substantial earnings losses (Pulido and Świecki, 2019; Hamory et al., 2021). When earnings are instead examined by trajectory groups, as shown in Figure 12, the distributions of log earnings for all groups shift to the right over time, indicating earnings growth for each trajectory group. This perspective suggests that trajectory groups differ in their initial mean earnings, reflecting underlying abilities and costs across sectors. When outcomes

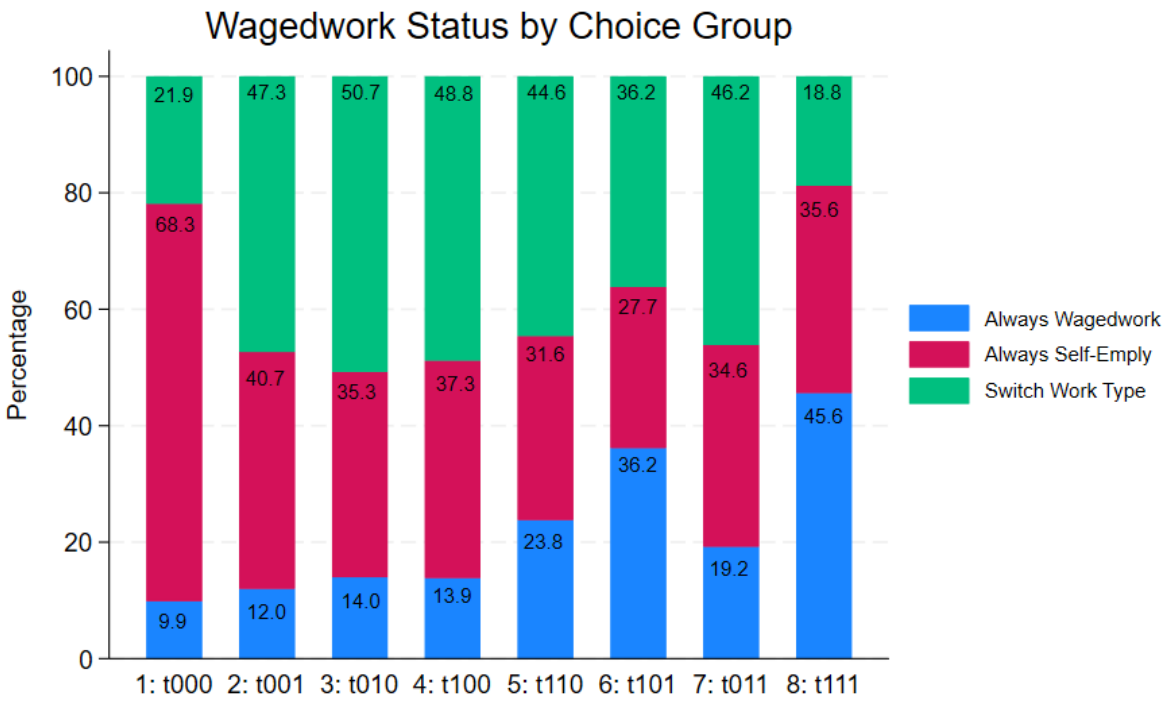


Figure 11

are aggregated to sector-level averages, these initial differences are masked, creating the appearance of earnings losses that are not evident once trajectories are taken into account.

Figure 13 further shows that hours worked remain relatively stable for stayers but fluctuate for switchers. Taken together, these results suggest that trajectory groups capture systematic heterogeneity consistent with comparative advantages on which sectoral choices are made.

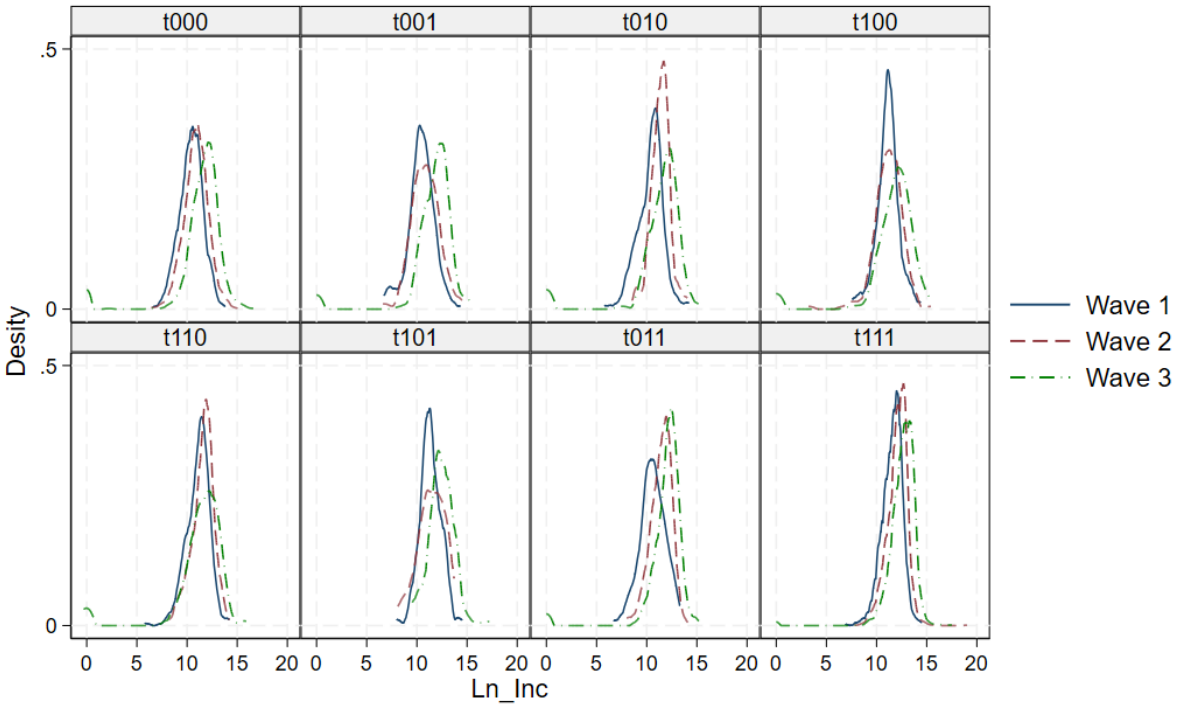


Figure 12: Log Earnings Distribution in 3 Waves by Choice Trajectory

ht

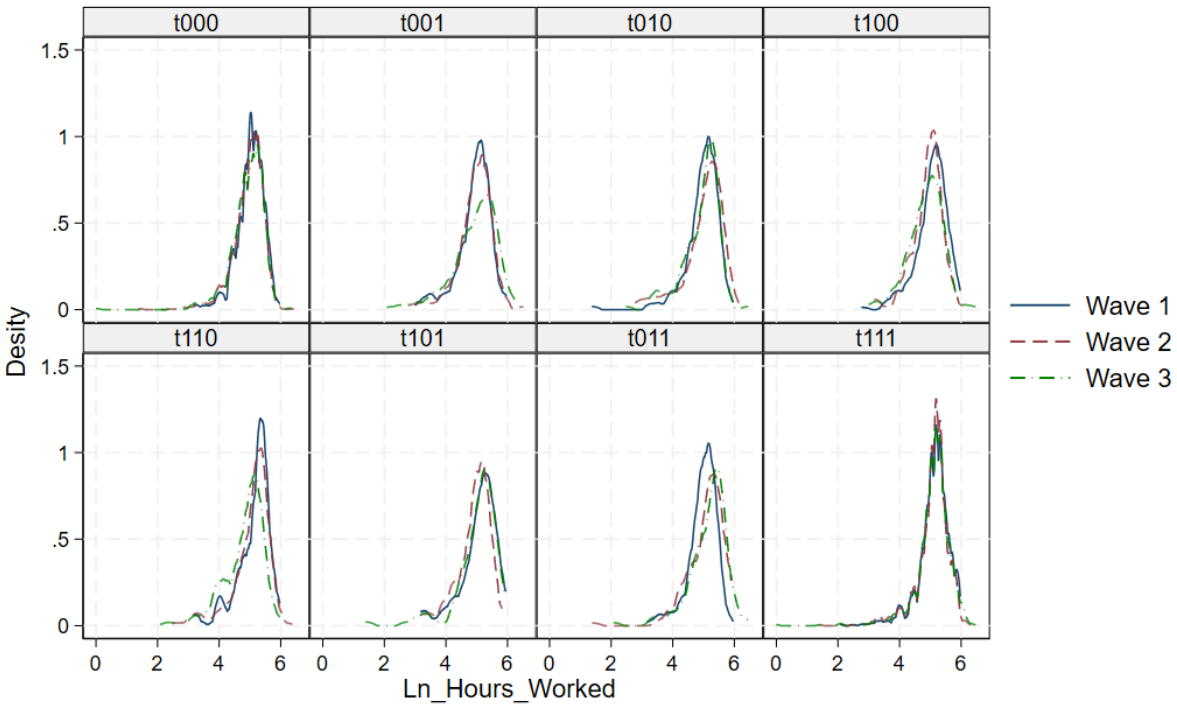


Figure 13: Log Hours Worked Distribution in 3 Waves by Choice Trajectory

# G Appendix: Model Structures and Estimation Details

## G.1 Heckman Two-step Estimation (Pooled Panel)

The canonical Heckman two-step estimator is designed to correct for selection bias in cross-sectional data. To evaluate the implications of parametric assumptions on the selection effect in the APG context, I estimate the selection-corrected wage equations under the assumption of joint normality in unobserved sectoral abilities using the Heckman two-step method (Heckman, 1979).

This approach, when applied to pooled panel data from the first three waves of IFLS (used in this paper), reveals a statistically significant selection effect. Recall the main result in the present paper: the selection effect due to comparative advantages at the individual level does not impact the sectoral productivity gap significantly when avoiding such a functional form assumption. By imposing a joint normal distribution on the latent skills, the pool sample from the same dataset finds that individual comparative advantages explain a significant portion of the sectoral productivity gap, which is consistent with the finding in Pulido and Świecki (2019) with the same distributional assumptions.

Abstracting from the time dimension, the Roy model framework is based on the assumption that an individual $i$ has potential earnings $y_i^n$ and $y_i^a$ in the non-agricultural sector $(n)$ and the agricultural sector $(a)$, respectively.

$$y_i^n = X_i\gamma^n + \epsilon_i^n \tag{70}$$

$$y_i^a = X_i\gamma^a + \epsilon_i^a \tag{71}$$

This setup mirrors the Roy model and is structurally comparable to the potential outcomes framework in Section 3. However, the estimation strategy differs: Heckman's method treats sectoral choice as endogenous and estimates the selection correction term explicitly, while the main framework in this paper treats sectoral choice as informative of comparative advantage and avoids strong distributional assumptions

In Heckman's estimation, for an individual who chooses to work in the non-agricultural sector, the selection equation governing sectoral choice is given by:

$$D_i = \begin{cases} 1 & \text{if} \quad Pr(y_i^n > y_i^a | X_i = x) \\ 0 & \text{otherwise} \end{cases} \tag{72}$$

Therefore, the probability of choosing the nonagricultural sector ($n$) follows:

$$\begin{aligned} Pr(D_i = 1 | X_i = x) &= Pr(y_i^n > y_i^a | X_i = x) \\ &= Pr(X_i\gamma^n + \epsilon_i^n > X_i\gamma^a + \epsilon_i^a) \\ &= Pr(\epsilon_i^a - \epsilon_i^n < X_i(\gamma^n - \gamma^a)) \end{aligned} \tag{73}$$

Assuming sector-specific unobserved abilities $\epsilon_i^n$ and $\epsilon_i^a$ are jointly normally distributed, the difference in unobserved abilities $\epsilon_i^n - \epsilon_i^a$ also follows a normal distribution. The resulting selection equation can be estimated as a probit model. Let $\hat{\gamma}$ denote the estimated coefficients from this probit regression of the sectoral choice equation. Then, IMR for individual $i$, denoted $\lambda_i$), is given by:

$$\lambda_i = \frac{\phi(X_i'\hat{\gamma})}{\Phi(X_i'\hat{\gamma})} \tag{74}$$

where $\phi(.)$ and $\Phi(.)$ refer to the standard normal density and Cumulative Density Function (CDF), respectively. Note that the derivation of IMR is well known and can refer to Heckman's seminal paper (1979). The IMR enters the outcome equation to correct for selection bias as follows:

$$y_i = X_i\gamma^n + \lambda_i\beta + u_i \quad \text{for} \quad D_i = 1 \tag{75}$$

In this outcome equation (75), $\beta$ captures the direction and magnitude of selection bias, and $y_i$ is an individual's observed wage working in the non-agricultural sectors, which is only observed for those with $D_i = 1$. $X_i$ is a vector of observed characteristics, and $\lambda_i$ is the estimated IMRs coming out of the selection equation. $\beta$ captures direction and magnitude of selection bias.

To bolster identification, I incorporate exclusion restrictions $Z_i$ in the selection equations but omit them from the outcome equation. This yields the augmented selection equation (76):

$$\begin{aligned} Pr(D_i = 1|X_i = x) &= Pr(y_i^n > y_i^a|X_i = x, Z_i = z) \\ &= Pr(X_i\gamma^n + Z_i^n\gamma_z^n + \epsilon_i^n > X_i\gamma^a + Z_i^a\gamma_z^a + \epsilon_i^a) \\ &= Pr(\epsilon_i^a - \epsilon_i^n < Z_i^n\gamma_z^n - Z_i^a\gamma_z^a + X_i(\gamma^n - \gamma^a)) \end{aligned} \tag{76}$$

In practice, the first-stage probit regression uses $Z_i$ as exclusion restrictions, the variables that affect sectoral choice but are assumed not to directly influence wages. For non-agricultural wage estimation, valid exclusion restrictions include age and non-farm business ownership. For the agricultural sector, the exclusion restrictions are rural-born, marital status, and farm business ownership. Valid exclusion restrictions are supported empirically (see Tables 12 and 13). The three regressions on the left of Table 12 show that the two variables, age and non-farm business, are not significant in the outcome equation for the non-agricultural workers; the left regression on Table 13 indicate they are both highly relevant to the sectoral decision in the selection equation. The evidence for exclusion restrictions for agriculture is on the right side of Tables 12 and 13.

Columns (1) to (3) in Table 14 present the Heckman two-step estimates for the non-agricultural sector. The table is divided into three blocks: (i) outcome equation estimates, (ii) selection equation estimates, and (iii) IMR coefficients. Column (1) includes basic covariates; column (2) adds log CPI; and column (3) includes both log CPI and a shock variable. All regressions include time fixed effects. Across all specifications, the IMR coefficient is negative and statistically significant, indicating adverse selection into the non-agricultural sector. The estimated selection effects are -0.138, -0.142, and -0.143, respectively. Selection effects for agricultural workers are estimated in a similar manner. The right block of Table 14, labelled as agriculture, shows corresponding regressions for agricultural workers under specifications (1) to (3), mirroring those used in the non-agricultural sector. In all three specifications, the IMR coefficients are positive and statistically significant, ranging from 0.202 to 0.245.

The wage for the agricultural workers can also be observed in the data. Similarly, the selection effect for the agricultural workers can be estimated. The right side of Table 12 and

Table 13 demonstrate that three variables, marital status, farm business, and rural born, are suitable exclusion restrictions for the agricultural sector. Table 14 presents the estimation of the Heckman selection correction for the agricultural workers. The three columns on the right side of the table correspond to three specifications used in the non-agriculture sector.

On average, the IMR coefficient for agricultural workers is positive and statistically significant, with estimated values of 0.245, 0.226, and 0.202 across specifications (Table 14). Taken together, under the assumption of bivariate normality in unobserved abilities, the selection effect in the non-agricultural sector is negative and statistically significant (ranging from -0.138 to -0.143), while that in the agricultural sector is positive and significant (ranging from 0.202 to 0.245). These magnitudes remain stable across model specifications, both with and without additional controls (log CPI and shock).

To benchmark these magnitudes, Column (3) of Table 9 estimates an average productivity gap of 0.654 using a pooled OLS regression with the same covariates as column (1) of Table 14. This implies that selection effects, as captured by the Heckman two-step estimator, explain approximately 21.1% to 37.5% of the observed APG in the baseline specification, and between 22% and 31% in the most complete specification (see Table 15).

This finding echoes the conclusion in Pulido and Świecki (2019), who report substantial selection effects under joint normality. In contrast, the empirical strategy employed in this paper imposes no functional form assumptions on unobserved heterogeneity and finds no significant selection effect. This divergence underscores the sensitivity of selection estimates to distributional assumptions.

## G.2 Heckman Selection Estimation with Panel Structure (xtheckman)

To evaluate whether panel structure affects the magnitude or direction of selection estimates under joint normality, I estimate the selection-corrected wage equation using Stata's **xtheckman** command. This approach extends the Heckman framework to panel data and fits a full maximum likelihood model accounting for both unobserved individual heterogeneity and time-varying error components.

Under the panel Roy model, individual $i$'s potential log earnings in sector $s \in \{n, a\}$ at time $t$ are given by:

$$y_{it}^n = X_{it}\gamma^n + \theta_i^n + \epsilon_{it}^n \tag{77}$$

$$y_{it}^a = X_{it}\gamma^a + \theta_i^a + \epsilon_{it}^a \tag{78}$$

where $(X_{it})$ are observed characteristics, $(\theta_i^s)$ are unobserved time-invariant individual sector-specific abilities, and $(\epsilon_{it}^s)$ are time-varying unobserved shocks. The model assumes that both $(\theta_i^n, \theta_i^a)$ and $(\epsilon_{it}^n, \epsilon_{it}^a)$ follow a joint normal distribution across sectors.

Sectoral choice at each period is modelled by a latent selection equation:

$$D_{it} = \begin{cases} 1 & \text{if } y_{it}^n > y_{it}^a \\ 0 & \text{otherwise} \end{cases} \tag{79}$$

Observed wages for those working in the non-agricultural sector ($D_{it} = 1$) are:

$$y_{it} = X_{it}\gamma^n + \theta_i^n + \epsilon_{it}^n \tag{80}$$

Unlike the pooled two-step Heckman estimator, **xtheckman** does not report estimated Inverse Mills Ratios (IMRs) or directly quantify the magnitude of selection effects. Instead, it provides estimated correlations across unobserved components of the outcome and selection equations, which imply the presence and direction of selection bias.

Table 16 reports results from the **xtheckman** estimation for non-agricultural workers. Two specifications are presented, mirroring columns (1) and (2) of Table 14. The model includes valid exclusion restrictions in the selection equation: age and non-farm business ownership. While estimation is computationally demanding, and convergence was not achieved in either specification, the results still reveal significant correlation in unobserved components, consistent with selection.

Despite limitations—long run times, convergence issues, and lack of explicit IMR estimates—the results from **xtheckman** reinforce the key message: under the joint-normality assumption, significant selection effects are recovered even in a panel framework. These findings align with those from the pooled Heckman model and underscore the influence of

distributional assumptions on empirical conclusions regarding comparative advantage and selection in the agricultural productivity gap.

## G.3 Control Function Approach for Panel Data

To further understand the impact of the distributional assumptions on the estimation results, I deploy the panel selection bias correction developed by Wooldridge (1995) to the first three waves of the IFLS dataset. This method takes into account the panel structure and uses a control function with a weaker assumption regarding the unobserved heterogeneity. Under the same Roy model framework in the panel structure, the key departure of Wooldridge (1995) from **xtheckman** is that the distribution of unobserved individual abilities remains unspecified; at the same time, IMRs are calculated for individual $i$ and each period $t$. Under the assumption that the error term in the outcome equation satisfies the conditional mean assumption (see Equation (85), Wooldridge (1995) exploits the panel structure by regressing the de-meaned log earnings on de-meaned regressors (the mathematical form expressed in (88)), including the transformed IMRs for selection bias correction.

In this method, each period latent choice variable $D_{it}^*$ follows (81). Instead, assuming a bivariate normal distribution of unobserved abilities between sectors for individuals, the selection equation (81) assumes that the error term $\nu_{it}$ is independent of $\mathbf{x_i}$ and normally distributed, which allows for calculating IMRs for each period. In Equation (82), $\mathbf{x_i'}$ refers to a vector of observed characteristics, including 1 for the intercept, $\delta_{t0}$ in Equation (81). Therefore, Equation (82) is a shorthand expression of Equation (81) using a vector form.

$$D_{it}^* = \delta_{t0} + x_{i1}\delta_{t1} + ... + x_{iT}\delta_{tT} + \nu_{it} \tag{81}$$

$$D_{it}^* = \mathbf{x_i'}\delta_\mathbf{t} + \nu_{it}, \quad t = 1, 2, ..., T \tag{82}$$

In the case that $D_{it} = 1$ represents individuals who choose the non-agricultural sector, Wooldridge (1995) uses the observed characteristics in all periods for a person's probability of choosing a non-agricultural sector. This method assumes $\nu_{it}$ is conditional mean zero and normally distribute, $\nu_{it} \sim$ Normal $(0, \sigma_t^2)$. Only when the latent choice variable $D_{it}^* > 0$, it will be observed, expressed in (83).

$$D_{it} = \begin{cases} 1 & \text{if} \quad D_{it}^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{83}$$

Subsequently, in the outcome equation, when $D_{it} = 1$, the log earnings will be observed. Hence, for both agricultural and non-agricultural workers, we can observe their earnings in the data; however, we don't know what they would have earned if they had chosen the other sector instead. For non-agricultural workers, the outcome equation is

$$y_{it} = \theta_i + X_{it}\gamma + u_{it} \quad \text{when} \quad D_{it} = 1 \tag{84}$$

In Equation (84), $y_{it}$ is observed earning for non-agricultural workers, and the assumption is that $u_{it}$ is strictly exogenous conditional on unobserved ability $\theta_i$ and observed characteristics $X_{it}$, see Equation (85). Let $\rho$ be the correlation between $u_{it}$ in the outcome equation and $\nu_{it}$ in the selection equation, representing the selection bias in the dataset (refer to Equation (86). Hence, the outcome equation in the expectation has the expression in Equation (87).

$$E(u_{it}|\theta_i, \mathbf{x_i}) = 0 \tag{85}$$

$$E(u_{it}|\theta_i, \mathbf{x_i}, \nu_\mathbf{i}) = E(u_{it}|\nu_{it}) = \rho\nu_{it} \tag{86}$$

$$E(y_{it}|\theta_i, \mathbf{x_i}, \nu_\mathbf{i}, \mathbf{D_i}) = \theta_i + X_{it}\gamma + \rho\nu_{it} \tag{87}$$

When estimating the outcome equation, Wooldridge (1995) first calculates the IMR for each period in the selection equation and then includes the IMRs as a control function in the outcome equation to correct for selection bias (mathematical form expressed in (88)). In estimating the outcome equation, all variables are transformed into demeaned variables (see Equations (89) and (90)), including IMRs (Equation (91)). This method can deliver consistent estimates while imposing much weaker assumptions on the unobserved components. By applying this method to the same dataset, the results will provide an assessment of the selection effect in the sample with a weaker distributional assumption.

$$\ddot{y}_{it} = \ddot{X}_{it}\gamma + \rho\ddot{\nu}_{it} + e_{it} \tag{88}$$

$$\ddot{y}_{it} \equiv y_{it} - \frac{1}{T_i}\sum_{r=1}^{T} D_{ir}y_{ir} \tag{89}$$

$$\ddot{X}_{it} \equiv X_{it} - \frac{1}{T_i}\sum_{r=1}^{T} D_{ir}X_{ir} \tag{90}$$

$$\ddot{\nu}_{it} \equiv \nu_{it} - \frac{1}{T_i}\sum_{r=1}^{T} D_{ir}\nu_{ir} \tag{91}$$

Table 18 represents three different sets of probit estimation as the first stage of Wooldridge's selection correction for panel data. The primary purpose of this step is to obtain the IMRs for each period by assuming the error terms are distributed as a normal distribution. As sector choice is a binary variable, i.e., either working in the non-agricultural or agricultural sector, the IMRs can be calculated for both agricultural and non-agricultural workers under the normal distribution of the error terms in the probit estimation. As Equation (82) shows, the regressors include all the history of observed characteristics. The exclusion restrictions are not required in this Wooldridge (1995) approach; however, they help improve the estimation if available. Let $\lambda(.)$ represent the Inverse Mills Ratio (IMR) and $\hat{\delta}_{\mathbf{t}}$ be estimated coefficients in vector form. Then, IMR for the non-agriculture sector is expressed in Equation (92), and IMR for agriculture workers is calculated as in Equation (93).

$$\lambda(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}}) = \frac{\phi(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}})}{\Phi(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}})} \tag{92}$$

$$\lambda(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}}) = \frac{-\phi(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}})}{1-\Phi(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}})} \tag{93}$$

The first block in Table 18 does not include either log CPI or shock; the second and third blocks add shock and log CPI, respectively. Each block estimates the selection probability for each period, which enables the calculation of IMRs for each individual for each period. Note that shock does not have any predictive power on the sectoral selection (shown in the second block). This estimation result provides evidential support for the earlier claim that individuals do not switch sectors in response to shocks. However, shocks will affect the

earnings in the outcome equation. Since the outcome equation only includes time-varying variables, the time-invariant variables, such as rural born and gender, can serve as exclusion restrictions in the selection equation.

With the IMRs acquired, a demeaned IMR will be calculated for each individual and included in the outcome equation estimation as a control function to correct selection bias. Table 19 uses IMRs calculated in the first block of Table 18, and the dependent variables are demeaned log income in the primary job. The left three regressions of Table 19 are for non-agriculture workers, and the right three ones are for agriculture. For each sector, three specifications are estimated, from the basic numbers of regressors in Column (1) to gradually adding log CPI in Column (2) and shock in Column (3); all the variables in the outcome equations are demeaned, as expressed in Equations (88) to (91). Table 19 shows that the selection effect is not statistically significant in both sectors, as indicated by the estimated coefficient of $\lambda$ (in the first line of the table). Moreover, the signs of the estimated coefficients are opposed to the results in the pooled dataset using the Heckman two-step method.

Wooldridge (1995) avoids making explicit the joint-normal distribution of individual unobserved heterogeneity across sectors; instead, this method uses the control function approach and exploits the panel structure. When applying this method to the three waves of the IFLS balanced dataset, selection effects are not statistically significant in either sector. This finding is aligned with the main analysis results in this paper by using Suri (2011)'s approach. Importantly, this method provides evidence on the consequences of the joint-normal distribution assumption on unobserved abilities. Pulido and Świecki (2019) use the same IFLS dataset and find a significant selection effect when imposing a joint-distributional assumption on unobserved components. Examining the first three waves of the IFLS dataset, the pooled Heckman and **xtheckman** with the same assumption as Pulido and Świecki (2019) produce a significant selection effect. Using the same dataset, Wooldridge's (1995) method relaxes the assumption of the bivariate normal and applies a weaker control function on the outcome equation; the estimation results show an insignificant selection effect. The selection effect varies significantly with the distributional assumptions imposed on the unobserved heterogeneity. Hence, it is more desirable to impose weaker assumptions on the individual unobserved abilities when estimating the magnitude of the selection effect. Heck-

man and Honore (1990) have warned of the consequences of such normal distribution in the empirical studies, even though this assumption provides meaningful insight in the theoretical studies.

This paper adopts the empirical approach by Suri (2011) to directly model individual latent skills in each sector while avoiding explicitly imposing functional forms; instead, this method exploits the information of each individual's sectoral choices over time. Unlike the Wooldridge (1995) approach, the methodology used in this paper estimates the individual latent comparative advantages through the choice trajectories and panel structure, which is a more desirable approach when the primary goal is to study the magnitude of the selection effect empirically. However, this method has its limitations. First, data on earnings and choices are required for each individual in each period. Second, the variations of earnings among different groups of choice trajectories need to be sufficiently large for the solutions to be stable when solving a system of equations. Tjernström et al. (2023) thoroughly discuss this limitation and provide evaluations. Despite those limitations, this method offers an attractive solution to the two primary challenges that the research question in this paper faces.

In this section, I first estimate the selection effect by using one of the prevailing approaches in the APG literature, TWFE on panel data, on the same dataset for the primary analysis of the present paper. The estimation results show a significant individual selection effect on sectoral productivity gaps, which reconciles with the findings in the APG literature applying this method. While controlling for individual fixed effects on panel data can remove unobserved heterogeneity, this method is inadequate to model comparative advantages due to the inability to model sector-specific unobserved individual abilities. As a result, this method confounds the effect relevant to sector choice with those that are irrelevant. Then, I present the consequences of the joint-normal distributions on unobserved components in the estimation results. With the same dataset, under the conventional joint-normal assumption, canonical Heckman two-step on pooled data and **xtheckman** both produce significant selection effect, which is aligned with the finding in Pulido and Świecki (2019). Once relaxing the joint normal distribution and using a weaker control function approach, Wooldridge's (1995) finds no significant selection effect in the same dataset. The empirical approach used in the

present paper avoids making such an assumption and exploits the information embedded in the trajectories of choices, which is a more desirable method to study the selection effect on the APG empirically.

| nonag_main | (1) | | | (1) + shock | | | (1) + ln_cpi | | |
|---|---|---|---|---|---|---|---|---|---|
| | *t = 1* | *t = 2* | *t = 3* | *t = 1* | *t = 2* | *t = 3* | *t = 1* | *t = 2* | *t = 3* |
| **nfarmbiz1** | 0.798 ** | 0.514 ** | 0.357 ** | 0.801 ** | 0.513 ** | 0.355 ** | 0.809 ** | 0.522 ** | 0.361 ** |
| | 0.065 | 0.064 | 0.063 | 0.065 | 0.064 | 0.063 | 0.065 | 0.064 | 0.063 |
| **farmbiz1** | -0.702 ** | -0.351 ** | -0.374 ** | -0.708 ** | -0.344 ** | -0.374 ** | -0.681 ** | -0.321 ** | -0.356 ** |
| | 0.068 | 0.068 | 0.067 | 0.069 | 0.068 | 0.068 | 0.069 | 0.069 | 0.068 |
| **ruralborn1** | -0.155 * | 0.009 | -0.142 | -0.154 * | 0.009 | -0.148 * | -0.156 * | 0.013 | -0.136 |
| | 0.091 | 0.086 | 0.088 | 0.091 | 0.087 | 0.088 | 0.092 | 0.087 | 0.088 |
| **shock1** | | | | -0.006 | -0.023 | 0.027 | | | |
| | | | | 0.058 | 0.057 | 0.057 | | | |
| **age1** | 0.003 | -0.001 | -0.024 ** | 0.003 | -0.001 | -0.024 ** | 0.006 | 0.001 | -0.023 ** |
| | 0.011 | 0.011 | 0.01 | 0.011 | 0.011 | 0.01 | 0.011 | 0.011 | 0.01 |
| **gender1** | -4.919 | 0.763 | 0.739 | -4.953 | 0.77 | 0.732 | -4.986 | 0.642 | 0.673 |
| | 85.969 | 0.894 | 0.874 | 85.321 | 0.891 | 0.863 | 99.891 | 0.88 | 0.87 |
| **educlevel21** | 0.215 ** | 0.149 ** | 0.211 ** | 0.217 ** | 0.149 ** | 0.211 ** | 0.228 ** | 0.163 ** | 0.216 ** |
| | 0.063 | 0.062 | 0.06 | 0.063 | 0.062 | 0.06 | 0.063 | 0.063 | 0.06 |
| **marital_status1** | -0.213 ** | -0.213 ** | -0.168 ** | -0.206 ** | -0.214 ** | -0.175 * | -0.203 ** | -0.203 ** | -0.164 ** |
| | 0.068 | 0.066 | 0.065 | 0.069 | 0.066 | 0.065 | 0.069 | 0.066 | 0.065 |
| **urban1** | 0.14 | -0.1 | 0.111 | 0.139 | -0.1 | 0.117 | 0.157 | -0.077 | 0.122 |
| | 0.177 | 0.177 | 0.173 | 0.176 | 0.177 | 0.174 | 0.176 | 0.177 | 0.174 |
| **wagedwork_main1** | 0.43 ** | 0.054 | -0.063 | 0.434 ** | 0.051 | -0.069 | 0.411 ** | 0.04 | -0.069 |
| | 0.074 | 0.074 | 0.074 | 0.074 | 0.074 | 0.075 | 0.074 | 0.074 | 0.074 |
| **ln_hrsworked1_m1** | 0.186 ** | 0.033 | -0.048 | 0.184 ** | 0.034 | -0.045 | 0.193 ** | 0.043 | -0.044 |
| | 0.054 | 0.053 | 0.054 | 0.055 | 0.053 | 0.054 | 0.055 | 0.053 | 0.054 |
| **ln_cpi1** | | | | | | | -3.22 ** | -1.755 | -0.951 |
| | | | | | | | 1.323 | 1.318 | 1.291 |
| **nfarmbiz2** | 0.387 ** | 0.657 ** | 0.275 ** | 0.385 ** | 0.658 ** | 0.278 ** | -0.228 ** | 0.648 ** | 0.273 ** |
| | 0.065 | 0.064 | 0.064 | 0.065 | 0.064 | 0.064 | 0.068 | 0.065 | 0.065 |
| **farmbiz2** | -0.215 ** | -0.585 ** | -0.26 ** | -0.227 ** | -0.581 ** | -0.256 ** | -0.228 ** | -0.6 ** | -0.265 ** |
| | 0.067 | 0.065 | 0.066 | 0.068 | 0.066 | 0.067 | 0.068 | 0.066 | 0.067 |
| **ruralborn2** | -0.038 | -0.184 | -0.091 | -0.033 | -0.186 | -0.095 | -0.052 | -0.193 | -0.095 |
| | 0.162 | 0.151 | 0.149 | 0.162 | 0.151 | 0.149 | 0.163 | 0.152 | 0.149 |
| **shock2** | | | | 0.074 | 0.017 | 0.003 | | | |
| | | | | 0.056 | 0.055 | 0.055 | | | |
| **age2** | -0.001 | 0.004 | 0.028 * | -0.002 | 0.004 | 0.027 * | -0.003 | 0.003 | 0.027 * |
| | 0.016 | 0.016 | 0.015 | 0.016 | 0.016 | 0.015 | 0.016 | 0.016 | 0.015 |
| **gender2** | 4.156 | -1.52 * | -1.528 * | 4.187 | -1.526 * | -1.521 * | 4.21 | -1.409 | -1.465 * |
| | 85.929 | 0.898 | 0.877 | 85.322 | 0.894 | 0.867 | 99.891 | 0.883 | 0.873 |
| **educlevel22** | 0.2 ** | 0.186 ** | 0.063 | 0.197 ** | 0.168 ** | 0.064 | 0.215 ** | 0.197 ** | 0.068 |
| | 0.062 | 0.063 | 0.06 | 0.063 | 0.06 | 0.06 | 0.063 | 0.063 | 0.06 |
| **marital_status2** | 0.097 | 0.118 | 0.021 | 0.092 | 0.117 | 0.013 | 0.096 | 0.118 | 0.023 |
| | 0.081 | 0.08 | 0.078 | 0.082 | 0.079 | 0.078 | 0.082 | 0.08 | 0.078 |
| **urban2** | -0.555 ** | -0.564 ** | -0.714 ** | -0.561 ** | -0.561 ** | -0.717 ** | -0.521 ** | -0.53 ** | -0.706 ** |
| | 0.205 | 0.203 | 0.202 | 0.204 | 0.203 | 0.202 | 0.206 | 0.204 | 0.203 |
| **wagedwork_main2** | 0.193 ** | 0.561 ** | 0.116 | 0.193 ** | 0.561 ** | 0.119 | 0.174 ** | 0.536 ** | 0.109 |
| | 0.078 | 0.075 | 0.078 | 0.078 | 0.075 | 0.076 | 0.078 | 0.076 | 0.076 |
| **ln_hrsworked1_m2** | -0.012 | 0.129 ** | 0.015 | -0.011 ** | 0.128 ** | 0.018 | -0.016 | 0.125 ** | 0.013 |
| | 0.051 | 0.049 | 0.049 | 0.051 | 0.049 | 0.049 | 0.051 | 0.049 | 0.049 |
| **ln_cpi2** | | | | | | | 2.57 ** | 2.04 * | 1.291 |
| | | | | | | | 1.281 | 1.248 | 1.223 |
| **nfarmbiz3** | 0.418 ** | 0.375 ** | 0.73 ** | 0.418 ** | 0.376 ** | -0.729 ** | 0.399 ** | 0.358 ** | 0.719 ** |
| | 0.062 | 0.061 | 0.062 | 0.062 | 0.061 | 0.062 | 0.062 | 0.061 | 0.063 |
| **farmbiz3** | -0.314 ** | -0.336 ** | -0.72 ** | -0.314 ** | -0.336 ** | -0.709 ** | -0.304 ** | -0.326 ** | -0.716 ** |
| | 0.069 | 0.068 | 0.066 | 0.069 | 0.069 | 0.066 | 0.069 | 0.068 | 0.066 |
| **ruralborn3** | -0.085 | -0.047 | 0.062 | -0.089 | -0.046 | 0.068 | -0.075 | -0.038 | 0.068 |
| | 0.097 | 0.095 | 0.093 | 0.097 | 0.095 | 0.093 | 0.098 | 0.096 | 0.094 |
| **shock3** | | | | 0.013 | -0.0146 | -0.132 ** | | | |
| | | | | 0.057 | 0.056 | 0.055 | | | |
| **age3** | -0.004 | -0.011 | -0.012 | -0.004 | -0.011 | -0.012 | -0.005 | -0.012 | -0.013 |
| | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 | 0.013 |
| **gender3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | (omitted) | (omotted) | (omitted) | (omitted) | (omitted) | (omitted) | (omitted) | (omitted) | (omitted) |
| **educlevel23** | -0.026 | -0.017 | 0.092 ** | -0.027 | -0.017 | 0.093 ** | -0.034 | -0.023 | 0.09 ** |
| | 0.034 | 0.034 | 0.033 | 0.034 | 0.034 | 0.033 | 0.034 | 0.034 | 0.033 |
| **marital_status3** | -0.055 | -0.079 | -0.017 | -0.059 | -0.076 | -0.003 | -0.062 | -0.086 | -0.02 |
| | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 | 0.063 |
| **urban3** | 1.031 ** | 1.142 ** | 1.079 ** | 1.041 ** | 1.139 ** | 1.073 ** | 0.999 ** | 1.106 ** | 1.069 ** |
| | 0.123 | 0.123 | 0.121 | 0.124 | 0.123 | 0.122 | 0.125 | 0.125 | 0.123 |
| **wagedwork_main3** | 0.17 ** | 0.162 ** | 0.556 ** | 0.172 ** | 0.16 ** | 0.551 ** | 0.164 ** | 0.162 ** | 0.553 ** |
| | 0.076 | 0.075 | 0.075 | 0.076 | 0.075 | 0.075 | 0.076 | 0.075 | 0.075 |
| **ln_hrsworked1_m3** | -0.015 | 0.071 | 0.208 ** | -0.011 | 0.07 | 0.204 ** | -0.01 | 0.072 | 0.207 ** |
| | 0.047 | 0.046 | 0.044 | 0.047 | 0.046 | 0.045 | 0.047 | 0.046 | 0.044 |
| **ln_cpi3** | | | | | | | -3.217 ** | -2.934 ** | -1.334 ** |
| | | | | | | | 0.673 | 0.66 | 0.649 |
| **cons** | -0.582 | -0.499 | -0.378 | -0.622 ** | -0.485 ** | 0.353 ** | 19.087 ** | 13.148 | 4.674 |
| | 0.439 | 0.426 | 0.42 | 0.44 | 0.428 | 0.422 | 5.596 | 5.561 | 5.447 |
| **N** | 4,513 | 4,513 | 4,513 | 4,510 | 4,510 | 4,510 | 4,513 | 4,513 | 4,513 |
| **LR chi2** | 2,885.33 | 2,773.61 | 2,859.02 | 2,884.73 | 2,771.46 | 2,862.19 | 2,909.10 | 2,793.80 | 2,863.38 |
| **Pseudo R2** | 0.494 | 0.475 | 0.479 | 0.494 | 0.475 | 0.479 | 0.498 | 0.478 | 0.479 |

Table 18: Wooldridge Probit Estimation (IFLS 1-3 Waves)

| dm_ln_inc1_m | *nonag_main* | | | ag_main | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) |
| dm_lambda | 0.502 ** | 0.334 | 0.33 | -0.208 | -0.3 | -0.26 |
| | 0.236 | 0.227 | 0.227 | 0.432 | 0.425 | 0.426 |
| dm_age | 0.126 ** | 0.058 ** | 0.057 ** | 0.112 ** | 0.045 ** | 0.042 ** |
| | 0.005 | 0.01 | 0.01 | 0.008 | 0.011 | 0.011 |
| dm_educ | 0.098 ** | 0.092 ** | 0.093 ** | 0.076 | 0.071 | 0.07 |
| | 0.023 | 0.022 | 0.022 | 0.065 | 0.064 | 0.064 |
| dm_marital | -0.055 | -0.077 * | -0.072 | 0.176 ** | 0.145 * | 0.162 ** |
| | 0.048 | 0.047 | 0.047 | 0.078 | 0.078 | 0.08 |
| dm_urban | 0.013 | 0.013 | 0.011 | -0.044 | 0.005 | 0.026 |
| | 0.057 | 0.055 | 0.055 | 0.168 | 0.165 | 0.168 |
| dm_wagedwork | 0.033 | 0.022 | 0.02 | 0.394 ** | 0.371 ** | 0.363 ** |
| | 0.067 | 0.066 | 0.066 | 0.128 | 0.125 | 0.125 |
| dm_ln_hrsworked | 0.234 ** | 0.231 ** | 0.23 ** | 0.243 ** | 0.254 ** | 0.247 ** |
| | 0.048 | 0.047 | 0.047 | 0.077 | 0.076 | 0.077 |
| dm_farmbiz | -0.049 | -0.011 | -0.005 | | | |
| | 0.045 | 0.044 | 0.044 | | | |
| dm_nfarmbiz | | | | -0.031 | -0.08 | -0.074 |
| | | | | 0.117 | 0.115 | 0.115 |
| dm_ln_cpi | | 1.488 ** | 1.499 ** | | 1.652 ** | 1.746 ** |
| | | 0.168 | 0.167 | | 0.21 | 0.207 |
| dm_shock | | | -0.048 * | | | -0.301 ** |
| | | | 0.027 | | | 0.067 |
| control for individual fixed effect | Y | Y | Y | Y | Y | Y |
| clustered | Y | Y | Y | Y | Y | Y |
| cons | 4.05E-08 ** | -4.11E-09 | 0.000011 | 3.75E-08 ** | -3.82E-08 ** | -4.49E-08 ** |
| | 1.72E-09 | 5.65E-09 | 7.87E-06 | 1.99E-09 | 9.60E-09 | 9.73E-09 |
| sigma_u | 2.78E-07 | 3.32E-07 | 0.002 | 2.67E-07 | 3.44E-07 | 3.53E-07 |
| sigma_epsilon | 0.828 | 0.816 | 0.816 | 1.598 | 1.588 | 1.58E+00 |
| ICC | 1.13E-13 | 1.66E-13 | 6.67E-06 | 2.79E-14 | 4.68E-14 | 4.97E-14 |
| corr(u_i, X_it) | 0 | 0 | 0.007 | 0 | 0 | 0 |
| N | 8,691 | 8,691 | 8,688 | 4,848 | 4,848 | 4,848 |
| Selected | 3,343 | 3,343 | 3,343 | 2,063 | 2,063 | 2,063 |
| F | 117.91 ** | 273.16 ** | 259.14 | 36.65 ** | 42.09 ** | 42.62 ** |
| | (8, 3,342) | (9, 3,342) | (10, 3,342) | (8, 2,062) | (9, 2062) | (10, 2,062) |

Table 19: Wooldridge Outcome Equation Estimation 1 (IFLS 1-3 Waves)

| dm_ln_inc1_m | *nonag_main* | | | ag_main | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (1) | (2) | (3) |
| dm_lambda | 0.505 ** | 0.386 * | 0.386 * | -0.143 | -0.161 | -0.116 |
| | 0.245 | 0.237 | 0.237 | 0.423 | 0.419 | 0.42 |
| dm_age | 0.126 ** | 0.058 ** | 0.057 ** | 0.111 ** | 0.044 ** | 0.042 ** |
| | 0.005 | 0.01 | 0.01 | 0.008 | 0.011 | 0.011 |
| dm_educ | 0.098 ** | 0.094 ** | 0.095 ** | 0.079 | 0.076 | 0.075 |
| | 0.023 | 0.022 | 0.022 | 0.065 | 0.065 | 0.064 |
| dm_marital | -0.055 | -0.077 * | -0.072 | 0.176 ** | 0.147 * | 0.164 ** |
| | 0.048 | 0.047 | 0.047 | 0.078 | 0.078 | 0.08 |
| dm_urban | 0.015 | 0.015 | 0.014 | -0.04 | 0.01 | 0.031 |
| | 0.057 | 0.055 | 0.055 | 0.168 | 0.165 | 0.168 |
| dm_wagedwork | 0.033 | 0.026 | 0.024 | 0.411 ** | 0.407 ** | 0.401 ** |
| | 0.067 | 0.066 | 0.066 | 0.126 | 0.124 | 0.123 |
| dm_ln_hrsworked | 0.234 ** | 0.234 ** | 0.233 ** | 0.248 ** | 0.264 ** | 0.258 ** |
| | 0.048 | 0.047 | 0.047 | 0.077 | 0.076 | 0.076 |
| dm_farmbiz | -0.049 | -0.019 | -0.013 | | | |
| | 0.045 | 0.044 | 0.044 | | | |
| dm_nfarmbiz | | | | -0.021 | -0.058 | -0.051 |
| | | | | 0.116 | 0.114 | 0.114 |
| dm_ln_cpi | | 1.49 ** | 1.501 ** | | 1.642 ** | 1.737 ** |
| | | 0.167 | 0.166 | | 0.209 | 0.207 |
| dm_shock | | | -0.049 * | | | -0.301 ** |
| | | | 0.028 | | | 0.067 |
| control for individual fixed effect | Y | Y | Y | Y | Y | Y |
| clustered | Y | Y | Y | Y | Y | Y |
| cons | 4.03E-08 ** | -4.38E-09 | 0.000011 | 3.78E-08 ** | -3.73E-08 ** | -4.41E-08 ** |
| | 1.74E-09 | 5.62E-09 | 7.90E-06 | 1.86E-09 | 9.66E-09 | 9.61E-09 |
| sigma_u | 2.78E-07 | 3.33E-07 | 0.002 | 2.67E-07 | 3.43E-07 | 3.52E-07 |
| sigma_epsilon | 0.828 | 0.816 | 0.816 | 1.598 | 1.588 | 1.583 |
| ICC | 1.13E-13 | 1.66E-13 | 6.54E-06 | 2.79E-14 | 4.66E-14 | 4.95E-14 |
| corr(u_i, X_it) | 0 | 0 | 0.007 | 0 | 0 | 0 |
| N | 8,691 | 8,691 | 8,688 | 4,848 | 4,848 | 4,848 |
| Selected | 3,343 | 3,343 | 3,343 | 2,063 | 2,063 | 2,063 |
| F | 116.48 ** | 272.41 ** | 258.68 | 36.67 ** | 41.89 ** | 42.62 ** |
| | (8, 3,342) | (9, 3,342) | (10, 3,342) | (8, 2,062) | (9, 2,062) | (10, 2,062) |

Table 20: Wooldridge Outcome Equation Estimation 2 (IFLS 1-3 Waves)

# References


Adamopoulos, T., Brandt, L., Leight, J., and Restuccia, D. (2022). Misallocation, selection and productivity: A quantitative analysis with panel data from china. *Econometrica*, 90:1261–1282.

Alvarez, J. A. (2020). The agricultural wage gap: Evidence from brazilian microdata. *American Economic Journal: Macroeconomics*, 12:153–173.

Alvarez-Cuadrado, F., Amodio, F., and Poschke, M. (2020). Selection, absolute advantage, and the agricultural productivity gap. *Centre for Economic Policy Research*, 4:1–82.

Alvarez-Cuadrado, F., Long, N. V., and Poschke, M. (2017). Capital-labour substitution, structural change and growth. *Theoretical Economics*, 12(3):1229–1266.

Blattman, C. and Ralston, L. (2015). Generating employment in poor and fragil states: evidence from labor market and entrepreneurship programs. *Poverty Action Lab*, pages 1–52.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *The American Economic Review*, 77(4):531–553.

Bryan, G., Chowdhury, S., and Mobarak, A. M. (2014). Underinvestment in a profitable technology: the case of seasonal migration in bangladesh. *Econometrica*, 82(5):1671–1748.

Cabanillas, O. B., Michler, J. D., Michuda, A., and Tjernstrom, E. (2018). Fitting and interpreting correlated random-coefficient models using stata. *The Stata Journal*, 18(1):159–173.

Caselli, F. (2005). Accounting for cross-country income differences. In Aghion, P. and N., D. S., editors, *Handbook of Economics Growth, Vol. 1A*, pages 679–741. Elsevier, Amsterdam.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, 18(1):5–46.

Chamberlain, G. (1984). Panel data. *Handbook of Econometrics*, 2:1247–1318.

Chanda, A. and Dalgaard, C.-J. (2008). Dual economies and international total factor productivity differences: channelling the impact from institutions, trade, and geography. *Economica*, 75(300):629–661.

David, C. C. and Otsuka, K. (1994). Differntial impact of modern rice varities in asia: an overview. In *Modern Rice Technology and Income Distribution in Asia*, pages 11–21. Boulder: Lynne Rienner Publisher.

Evenson, R. and Gollin, D. (May 2, 2003). Assessing the impact of the green revolution. *Science*, 300(5620):758–762.

Frankerberg, E., Karoly, L. A., Gertler, P., Achmad, S., Agung, I. G. N., Hatmadji, S. H., and Sudharto, P. (1995). The 1993 indonesia family life survey: overview and field report. *RAND*, 1(DRU-1195/1-NICHD/AID).

Gai, Q., Guo, N., Li, B., Shi, Q., and Zhu, X. (2021). Migration costs, sorting, and the agriculture productivity gap. *University of Toronto*, Working Paper(693).

Gollin, D., Lagakos, D., and Waugh, M. E. (2014). The agricultural productivity gap. *The Quarterly Journal of Economics*, 129(2):939–993.

Gollin, D., Parente, S., and Rogerson, R. (2002). The role of agriculture in development. *American Economic Review*, 92(2).

Gouriéroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8:S85–S118.

Group, W. B. (2025). Databank: World development indicators, ind.

Hamory, J., Keelmans, M., Li, N. Y., and Miguel, E. (2021). Reevaluating agricultural gaps with longitudinal microdata. *Journal of the European Economic Association*, 19(3):1522–1555.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Heckman, J. and Vytlacil, E. (1998). Instrumental variables methods for the correlated random coefficient model: estimating the average rate of return to schooling when the return is correlated with schooling. *The Journal of Human Resources*, 33(4):974–987.

Heckman, J. J. and Honore, B. E. (1990). The empirical content of roy model. *Econometrica*, 58(5):1121–1149.

Herrendorf, B. and Schoellman, T. (2018). Wages, human capital, and barriers to structural transformation. *American Economic Journal: Macroeconomics*, 10:1–23.

Hofman, B. and Kaiser, K. (2002). The making of the big bang and its aftermath: A political economy perspective. In *Can Decentralization Help Rebuild Indonesia?*, volume Working Paper 02-25 of *International Studies Program*, Atlanta, Georgia. Andrew Young School of Policy Studies, Georgia State University.

House, F. (1998). Freedom in the world 1998 - indonesia.

Indonesa, B.-S. (2019). Consumer price index (general).

Kuznets, S. (1971). *Economic Growth of Nations: Total Output and Production Structure*. Harvard University Press, Cambridge, MA.

Lagakos, D. (2020). Urban-rural gaps in the developing world: does internal migration offer opportunities. *Journal of Economic Perspectives*, 34(3):174–192.

Lagakos, D., Marshall, S., Mobarak, A. M., Vernot, C., and Waugh, M. E. (2020). Migration costs and observational returns to migration in the developing world. *Journal of Monetary Economics*, 113:138–154.

Lagakos, D. and Waugh, M. E. (2013). Agriculture, and cross-country productivity differences. *The American Economic Review*, 103(2):948–980.

Lemieux, T. (1998). Estimating the effects of unions on wage inequality in two-sector model with comparative advantage and non-random selection. *Journal of Labour Economics*, 16:261–291.

Lewis, W. A. (1954). Economic development with unlimited supplies of labour. *The Manchester School*, 22(2):115–227.

McKenzie, D. (2017). How effective are active labor market policies in developinf countries? a critical review of recent evidence. *World Bank, Policy Research Working Paper*, WPS(8011).

McMillan, M. and Rodrik, D. (2014). Globalization, structural change, and productivity growth, with an update on africa. *World Development*, 63:11–32.

Munshi, K. and Rosenzweig, M. (2016). Networks and misallocation: insurance, migration, and the rural-urban wage gap. *American Economic Review*, 106(1):46–98.

Pulido, J. and Świecki, T. (2019). Barriers to mobility or sorting? sources and aggregate implications of income gaps across sectors and locations in indonesia. *Working Paper*.

Restuccia, D., Yang, D. T., and Zhu, X. (2008). Agriculture and aggregate productivity: a quantitative cross-country analysis. *Journal of Monetary Economics*, 55(2):234–250.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers, New Series*, 3(2):135–146.

Samuelson, P. A. (1938). A note on pure theory of consumer's behaviour. *Economica*, 5(17):61–71.

Samuelson, P. A. (1948). Consumer theory in terms of revealed preference. *Economica*, 15(60):243–253.

StataCorp (2025). Stata statistical software: Release 19. kdensity.

Strauss, J., Witoelar, F., and Sikoki, B. (2016). The fifth wave of indonesia family life survey (ifls4): overview and field report. *RAND*, 5(WR-1143/1-NIA/NICHD).

Suri, T. (2011). Selection and comparative advantage in technology adoption. *Econometrica*, 79:159–209.

Tjernström, E., Ghanem, D., Cabanillas, O. B., Lybbert, T., Michuda, A., and Michler, J. (2023). Comment on suri (2011) "selection and comparative advantage in technology adoption". Working paper.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, 68:115–132.