Expectations versus Reality in Business Formation*

Emin Dinlersoz† Yueyuan Ma‡

March 2026

Abstract

Using administrative data on 17 million U.S. business applications linked to outcomes, we compare potential entrants' expectations about employer entry and first-year employment with realizations. On average, applicants overestimate employment, mainly because many expect to enter but do not. Among those who expect and achieve entry, employment is typically underestimated. Expected employment predicts entry and realized employment, but conditional on entry realized employment rises less than one-for-one with expectations. Expectation errors are highly heterogeneous and systematically related to application characteristics and local economic conditions, and they predict near-term employment outcomes. A parsimonious model with heterogeneous priors, learning, and pre-entry selection rationalizes these patterns.

Keywords: business applications, expectation errors, entry, employment, misallocation

* Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. We thank Timothy Dunne, Max Farrell, Lucia Foster, Cheryl Grim, Ron Jarmin, Vitaliy Novik, Timo Reinelt, Ryan Oprea, Alisa Tazhitdinova, Jeffrey Yang, and the participants at the UCSB Applied Micro Seminar for comments and suggestions. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure-avoidance protection (Project 7500391: CBDRB-FY25-0346).

†Center for Economic Studies, U.S. Census Bureau; emin.m.dinlersoz@census.gov

‡Department of Economics, University of California Santa Barbara; yueyuanma@ucsb.edu

1 Introduction

Expectations about the future are central to models of firm behavior and entrepreneurial decision-making. Entrepreneurs form beliefs about whether they will be able to launch a business, how many employees they will have, and how their venture will evolve. These expectations influence initial commitments, resource acquisition, financing needs, and organizational planning. Because these decisions shape the allocation of labor, capital, and other inputs, errors in expectations can distort the efficient allocation of resources. For example, overly optimistic expectations may lead to excessive entry and misallocation, while pessimistic expectations may deter otherwise productive ventures. Yet, despite their foundational role in both theoretical models and empirical implications, little is known about expectation errors around the time of business formation.

Most existing studies on firms' expectations focus on businesses that have already begun operating (Guiso and Parigi (1999); Bloom (2007); Coibion, Gorodnichenko, and Kumar (2018)). Early-stage ventures, however, typically face greater informational constraints, frictions, and uncertainty. As a result, expectation errors may be both systematic and large at the point of entry (Camerer and Lovallo (1999); Koellinger, Minniti, and Schade (2007)). Importantly, such errors may take the form of both overestimation and underestimation, with potentially distinct patterns at the extensive margin of entry and the intensive margin of initial employment. In addition, expectations at the formation stage can be informative on unobserved pre-entry heterogeneity across businesses, which is relevant for variation in firm growth (Sterk, Sedláček, and Pugsley (2021)).

Classic models of firm dynamics assume that potential entrants possess rational expectations or at least forecasts that are efficiently formed given available information (see, e.g., Jovanovic (1982); Hopenhayn (1992)).[1] In Jovanovic's model, firms enter with imperfect information about their own unknown fixed productivity and learn about it over time, while in Hopenhayn's framework, firms do not know their productivity ex-ante and observe it upon entry, after which productivity evolves stochastically. Behavioral models, by contrast, allow for expectation errors due to optimism bias, bounded rationality, miscalibration, or overconfidence.[2] Others highlight the role of information frictions and institutional constraints.[3] While a rich literature examines entrepreneurial beliefs

[1] Other firm dynamics frameworks with similar assumptions include Nocke (2006) and Ericson and Pakes (1995).
[2] See, e.g., Camerer and Lovallo (1999); Koellinger et al. (2007).
[3] See, e.g., Hurst and Pugsley (2011); Hall (2010); Gennaioli, La Porta, Lopez-de Silanes, and Shleifer (2013).

using experimental settings or surveys, few studies provide observational evidence linking ex-ante expectations to ex-post outcomes at the business formation stage.[4] The rarity of such studies owes in part to data limitations: expectations are seldom recorded systematically near the point of entry, and are rarely linked to subsequent outcomes.

This paper provides novel observational evidence on entrepreneurial expectations and how they compare with realized outcomes using administrative microdata from about 17 million applications to start new businesses, submitted between 2017 and 2021 in the United States. These data, which underlie U.S. Census Bureau's Business Formation Statistics (BFS), are based on applications made to the IRS for Employer Identification Numbers (EINs) using IRS Form SS-4 and capture nearly all economically meaningful applications for new businesses, including all employer businesses, corporations, and partnerships. The data include expected initial employment, and the links to Census Bureau's Longitudinal Business Database (LBD) allow us to observe whether an application becomes an employer business (a business with at least one paid employee), which we define as *entry*, and track its realized employment over time.[5] We use this structure to study the accuracy of expected initial employment reported at the time of application.

Our analysis proceeds in several parts. First, we document broad patterns of forecast error. On average, applicants overestimate initial employment, with substantial heterogeneity: many expect entry but do not enter, some enter without anticipating it, and those who both expect and achieve entry tend to underestimate. Both extensive (entry) and intensive (employment) margins drive errors. Tests based on Mincer–Zarnowitz type regressions indicate biased and inefficient forecasts: realized employment falls short of expectations overall, and a one-unit increase in expected employment translates into less-than-one-unit increase in realized employment. A variety of econometric specifications designed to address non-linearity and selection in the entry process reveals that this overestimation at the margin is robust. Year-by-year estimates further show a weakening correlation between expected and realized employment, particularly after 2020.

Second, we relate errors to application attributes and local conditions, without a causal inter-

[4] See, e.g., Manski (2004); Astebro, Herz, Nanda, and Weber (2014); Puri and Robinson (2007) for the literature based on surveys and experiments.

[5] Entry as non-employer businesses is also important, as it provides income for a large number of individuals. However, in terms of employment generation, the outcome of a non-employer business is equivalent to a business not materializing at all. Because our focus is on employment generation, we do not distinguish non-employer business formation from no business formation. Work in progress (Dinlersoz, Kroff, Luque, and Novik, 2025) analyzes patterns of both employer and non-employer entry originating from business applications.

pretation. Corporations and LLCs report expectations that are more conservative and closer to realizations than those of sole proprietors or partnerships, consistent with greater planning capacity, formality, or access to advisors (Levine and Rubinstein, 2017; Hurst and Pugsley, 2011). Higher tract-level college share is associated with lower overestimation, whereas higher nonwhite share is linked to larger overestimation—patterns consistent with differences in access to planning resources, information networks, or structural constraints (Gennaioli et al., 2013). Greater local firm density correlates with lower overestimation, in line with agglomeration benefits (Moretti, 2004b; Carlino and Kerr, 2009). By contrast, greater dispersion in incumbent employment and productivity is associated with larger overprediction, pointing to the potential role of unpredictability and heterogeneous shocks (Bachmann, Elstner, and Sims, 2013; Wu and Knott, 2006).

Sectoral differences are also pronounced: high-tech and manufacturing applicants exhibit higher overestimation, which may reflect greater uncertainty, reliance on intangibles, and longer development cycles (Hall, 2010). At the same time, relatively more standardized service sectors (e.g., professional services, real estate) display smaller errors. Overall, business environment (e.g. market structure, industry characteristics) and expectations are systematically linked.

Third, we examine how expectation errors evolved around the onset of COVID-19. Starting in 2019, errors rose as realized employment weakened more sharply than expected employment, reflecting that applications filed from March 2019 onward had 12-month horizons extending into the pandemic, when actual startup sizes fell. In the run-up to March 2020, expectations declined only modestly and failed to anticipate the severity of the shock, while realized employment declined more abruptly, leading to growing overestimation. After the shock, expectations leveled off, but realizations remained depressed, leaving a persistent gap. Dispersion in errors, already trending down before 2019, compressed further around the pandemic onset, driven by a rise in applications projecting zero employment, before widening again in the recovery (especially among positive-expectation applications) as heterogeneity re-emerged. These patterns illustrate how a large, unexpected macroeconomic shock can shape the level and dispersion of expectation errors.

Fourth, we relate expectation errors to near-term outcomes: third-year employer status (any positive employment three years after application) and employment levels (counting failures as zero). Controlling for observables, greater initial overestimation is positively associated with both outcomes. In canonical models with common priors (e.g., Jovanovic (1982); Hopenhayn (1992)),

forecast errors at entry should carry no predictive content for future performance once initial outcomes are controlled for. The fact that errors remain predictive suggests heterogeneous priors with private signals not fully revealed at entry.

Finally, we propose an entry model that rationalizes our key findings by minimally extending Jovanovic (1982). Prior to application, potential entrants hold heterogeneous beliefs with an *informative* component (private signals about business quality) and an *uninformative* component (e.g., heterogeneous interpretation of common information, such as over- or under-confidence). The informative component generates a positive relationship between initial forecast errors and near-term employment, while selection at the application stage, combined with the uninformative component, produces aggregate overestimation among applicants. Together, the two components also imply overestimation at the margin: a one-unit increase in expected employment translates into a less-than-one-unit increase in realized employment. The declining correlation between expected and realized employment especially after 2020 suggests that the variance of the uninformative component has increased relative to underlying business quality or that the informative signal has become less precise.

Our contributions are threefold. First, we extend the literature on early-stage business activities by providing new evidence on expectation errors near the point of business formation, distinguishing between extensive margin of entry and intensive margin of employment outcomes.[6] We document how these errors vary with applicant characteristics, local conditions, industry environments, and in response to the COVID-19 shock. Second, we provide large-scale observational evidence relevant for both classical and behavioral theories of entrepreneurship, complementing prior experimental and survey-based findings. In particular, systematic overestimation at the time of business formation may help explain entrepreneurial entry despite low risk-adjusted returns, addressing a puzzle regarding the roots of entrepreneurship (Moskowitz and Vissing-Jørgensen (2002); Astebro et al. (2014)). Third, we show that expectations at the application stage contain valuable information about ex-ante heterogeneity and ex-post performance. Our predictive econometric analyses that incorporate expectations can help assess near-term job-creation potential from business applications even prior to realization. More broadly, the analysis underscores the importance of studying

[6] See, e.g., Guzman and Stern (2020); Bayard, Dinlersoz, Dunne, Haltiwanger, Miranda, and Stevens (2018); Dinlersoz, Dunne, Haltiwanger, and Penciakova (2023) on early-stage business activities.

early-stage expectations, an area that remains underexplored in both theory and empirical research.

2 Conceptual framework

We study potential entrants who file a business application in period t, contemplating the formation of a new business—employer or nonemployer.[7] In the data, potential entrants reveal expectations about whether they will start an employer business, which we refer to as "entry", and the highest employment their business will achieve within a finite horizon of τ periods (e.g., months or quarters). Actual entry and size that are realized over this horizon may diverge from initial expectations. Some applications may become non-employer businesses—an important source of income for many individuals. Since non-employer businesses generate no employment, we do not distinguish them from no formation and classify both as no entry/zero employment based on our definition of entry.

2.1 Entry and employment choice

Let V_{it} denote the random variable representing the maximum net expected value from starting an employer business at any period between t and $t+\tau$. Entry occurs if and only if the realization v_{it} of this variable is positive. Let $E_{it} = I(V_{it} > 0)$ denote the random variable representing entry outcomes. The probability of entry is given by $p_{it} = \Pr(E_{it} = 1) = \Pr(V_{it} > 0)$. The event of no entry corresponds to no employment generation within τ periods (e.g., no or late employer business formation, or a nonemployer business formation). Define $e_{it} = I(v_{it} > 0)$ as the entry realization. When $e_{it} = 1$, entry occurs within τ periods; when $e_{it} = 0$, no entry takes place in that horizon.

Let M_{it} denote the maximum employment of the business over the horizon — a random variable. $\mu_{it} = \mathbb{E}[M_{it}]$ denotes the underlying (population) expected maximum employment.[8] If entry occurs, the observed maximum employment level is $m_{it} > 0$. If no entry occurs, $m_{it} = 0$. The relationship between business value and employment is then

$$m_{it} \begin{cases} > 0 & \text{if } v_{it} > 0, \\ = 0 & \text{if } v_{it} \leq 0. \end{cases} \tag{1}$$

[7] Employer businesses are those that hire workers, as opposed to nonemployer businesses, which are operated without paid employees.

[8] For instance, one can write this as $\mu_{it} = \theta_t + \nu_{it}$, where θ_t is the component common to all potential entrants of a given business type at time t and ν_{it} is an idiosyncratic, applicant-specific component.

In other words, the realized maximum employment depends on both the extensive margin (whether the business starts as an employer) and the intensive margin (employment conditional on entry).

2.2 Expectations at time of application

At the time of application (t), the potential entrant forms expectations based on available information. Let the applicant's information set be $\mathcal{I}_{it}$. Define the *subjective* time-t estimate of the mean of the random variable M_{it} as

$$\tilde{\mu}_{it} = \mathbb{E}[\, M_{it} \mid \mathcal{I}_{it}]. \tag{2}$$

The expectation $\tilde{\mu}_{it}$ reflects beliefs about demand, profitability, ability, competition, and various sources of uncertainty, including macroeconomic factors. Entry is expected if and only if expected maximum employment is strictly positive. The indicator $\tilde{e}_{it} = I(\tilde{\mu}_{it} > 0)$ denotes expected entry. In a business application (made using the IRS form SS-4), applicants directly report $\tilde{e}_{it}$ and $\tilde{\mu}_{it}$.

2.3 Expectation errors

Define the realized entry expectation (or forecast) error as

$$\omega_{it} = \tilde{e}_{it} - e_{it}. \tag{3}$$

Here, $\omega_{it} = 1$ if the potential entrant expected to enter but did not, $\omega_{it} = -1$ if entry occurred despite no expectation of entry, and $\omega_{it} = 0$ if expectations and outcomes match. Similarly, define the realized expectation (or forecast) error of employment as

$$\delta_{it} = \tilde{\mu}_{it} - m_{it}. \tag{4}$$

In this case, $\delta_{it} > 0$ indicates overestimation, $\delta_{it} < 0$ underestimation, and $\delta_{it} = 0$ perfect accuracy.

2.4 Decomposition of expectation errors in the population

Denote the population shares of expectation-outcome pairs by $s(\tilde{e}, e)$, for $\tilde{e}, e \in \{0, 1\}$. Let $\bar{\delta}(\tilde{e}, e)$ be the population mean of employment expectation error conditional on group $(\tilde{e}, e)$. Then, the overall population mean employment expectation error can be decomposed as

$$\bar{\delta} = \bar{\delta}(1,1)s(1,1) + \bar{\delta}(1,0)s(1,0) + \bar{\delta}(0,1)s(0,1) + \bar{\delta}(0,0)s(0,0), \tag{5}$$

where $\bar{\delta}(0,0) = 0$ by construction. Similarly, the population mean of entry expectation error is $\bar{\omega} = s(1,0) - s(0,1)$. The decomposition (5) is useful in understanding the behavior expectation errors in different sub-groups of the population.

2.5 Sample implementation and inference

In a sample of n potential entrants, we observe reported expected maximum employment ($\tilde{\mu}_{it}$), realized maximum employment (m_{it}), and expected and actual entry, $\tilde{e}_{it}$ and e_{it}. Using these, the sample averages of the expectation errors are

$$\hat{\bar{\delta}} = \frac{1}{n}\sum_{(i,t)}^{n} \delta_{it}, \quad \hat{\bar{\omega}} = \frac{1}{n}\sum_{(i,t)}^{n} \omega_{it} \tag{6}$$

As the sample size increases ($n \to \infty$), these statistics converge to their respective population counterparts, $\bar{\delta}$ and $\bar{\omega}$, under the Strong Law of Large Numbers, given standard regularity conditions.[9] These statistics can be used to formally test for systematic bias in expectations.

3 Data

Our analysis leverages microdata from the U.S. Census Bureau's Business Formation Statistics (BFS) program, which provide comprehensive coverage of business applications linked to subsequent outcomes. Key features of the data are summarized below; see Bayard et al. (2018) for details.

3.1 Business applications

The BFS compiles statistics on early-stage entrepreneurial activity based on applications for Employer Identification Numbers (EINs) via IRS Form SS-4.[10] The underlying microdata contain rich application-level information, including six-digit NAICS industry codes and geographic identifiers. These fields enable identification of *business applications* – EIN applications with a business intent.

The data encompass all economically significant business applications. All employer businesses are required to obtain an EIN to file payroll taxes. Corporations and partnerships, regardless of

[9] The Strong Law requires that sample averages converge almost surely to their expectations. Sufficient conditions include finite first and second moments and, in the non-*i.i.d.* case, restrictions on the nature of the correlations. In our context, expectation errors are unlikely to be strictly *i.i.d.* and may exhibit cross-sectional dependence arising from local or industry shocks or correlated beliefs.

[10] See the BFS website for additional documentation.

employment status, must also acquire one. While sole proprietors without employees do not need an EIN, many obtain one for banking, licensing, or compliance (Fairlie, Kroff, Miranda, and Zolas (2023)). Nonemployers with EINs represent a substantial share of total nonemployer revenues.[11]

Each application includes information on business name/location, date of filing (week), intended start date, reason for applying (e.g., starting a new business, purchasing an existing business, banking purposes), legal form, prior EIN filing, principal business activity, planned date of first wage payment, and the maximum number of employees expected in the next 12 months. These variables capture salient features of the applicant's plans and provide predictive signals about subsequent employer status (Bayard et al. (2018)).[12]

Our analysis focuses on business applications that report an intent to "start a new business" (by checking the relevant box in the application form for the reason for applying). These constitute 90% of all business applications in the period of analysis. Excluded are those applications with reasons such as banking purposes, purchasing an existing business or changing organizational form, which may not face the same type of uncertainty as new businesses.

A substantial share of the applications do not indicate a planned wage payment date and report zero expected employment; many of these are likely intended as nonemployer businesses or reflect uncertainty about early hiring. Nevertheless, a small but nontrivial proportion of such applications transition to employer status ex post. The applications that report a planned wage payment (classified as Applications with Planned Wages (WBA) in the BFS) exhibit significantly higher employer transition rates (27% within 4 quarters of application), compared to 7% for all applications in our sample. WBA almost always report positive expected employment within 12 months of application – the correlation in our sample is 0.996. Because applications with positive expected employment may reflect more concrete employer business plans, expectation errors could behave differently for this group. Conversely, applications with zero expected employment (likely indicating nonemployer intent or delayed hiring) may yield more accurate expectation errors due to the lower share of actual entry or positive employment. We therefore conduct the analysis on both the full applications sample (including cases with zero expected employment) and the positive

[11] See Davis, Haltiwanger, Krizan, Jarmin, Miranda, Nucci, and Sandusky (2009), who also document that EIN nonemployer firms report significantly higher revenues on average than their non-EIN counterparts.

[12] The BFS program uses these characteristics to project near-term employer business formations for recent periods not yet observable in the LBD.

expected employment sample (nearly identical to WBA), and document the differences.

The analysis sample contains monthly data over 2017m4–2021m12. While BFS microdata are available since 2004, the expected maximum employment variable is available only from April 2017. The end of the sample period (2021m12) was determined by the availability of actual employment information for applications at the time of our analysis. The analysis sample contains nearly 17 million applications with a plan to start a new business.

3.2 Business formations

The BFS links EIN applications to the Longitudinal Business Database (LBD), a comprehensive panel containing data on firm identifiers, location, industry, age, and quarterly employment and payroll. Using this link, we identify transitions to employer status (entry) by matching application EINs to first observed instances of positive employment in the LBD.

Employment is observed at quarterly frequency. Consistent with the horizon for expected maximum employment, we set the forward-looking horizon for actual employment generation to $\tau = 4$ quarters (12 months) including the quarter of filing. Entry occurs if a firm generates positive employment in any of these quarters. Since application dates vary within the quarter, applicants filing later in a quarter have less time to generate employment within a fixed 4-quarter horizon. To address this timing issue, we use the week of application within a quarter to adjust the window of observation to identify the first positive employment. Specifically, applications filed in week $w = 1, \ldots, 13$ have a probability, $w/13$, of being assigned an additional (fifth) quarter for employment generation. This procedure gives all applicants, on average, a full 12-month window to transition into employer status, mitigating bias from within-quarter timing variation.

3.3 Measures of expected entry, realized entry, and initial size

Form SS–4 asks applicants to report the highest number of employees they expect within the first year. The IRS uses this value primarily to assign an employer to quarterly payroll tax filing (Form 941) or to the simplified annual filing regime (Form 944). We use this information to construct two expectation measures: (i) an indicator for expected entry as an employer, $\tilde{e}_{it}$, which equals one if the applicant expects any employees, and (ii) expected maximum employment, $\tilde{\mu}_{it}$, which captures the intensive margin and satisfies $\tilde{\mu}_{it} > 0$ or $\tilde{e}_{it} = 1$. The relevant question, "What is

the highest number of employees expected in the next 12 months? (Enter 0 if none)," has a response recorded for all EIN applications beginning in 2017m4.[13] The indicator $\tilde{e}_{it}$ is nearly identical to the indicator for planning to pay wages (correlation 0.996), which suggests that applicants report consistent expectations about both the decision to hire and expected employment levels.

Information on expected hiring and maximum employment is collected on IRS Form SS-4 primarily to support tax administration (e.g., helping determine which tax filings and withholding obligations may apply). Because these responses are not an enforceable commitment, carry limited immediate financial consequences at the application stage, and can be updated as the business evolves during subsequent tax reporting, they provide a useful window into entrepreneurs' beliefs at the outset. Most applicants complete SS-4 very early, often before they face binding regulatory thresholds or heightened external scrutiny from lenders, landlords, or investors. As a result, reported employment plans are less likely to be shaped by strategic considerations than later-stage disclosures. Finally, EIN issuance is an administrative process that depends on submitting a complete and valid application rather than on performance-based screening, reducing concerns that observed expectations reflect selection by the approval process.

We construct realized (actual) entry and employment measures from the LBD. Entry occurs, i.e., $e_{it} = 1$, if any of the four post-application quarters record positive employment. That is, we use a horizon of $\tau = 4$ periods (quarters) to match the potential entrant's one-year horizon for expectations. Maximum realized employment, m_{it}, is defined as the highest quarterly employment observed during this window. Applications with no observed employment receive a value of zero – indicating no employment generation.[14]

A potential concern is that the LBD measures employment on a quarterly basis, while applicants may interpret the 12-month expectation horizon at a finer temporal resolution—for example, as referring to monthly or even bi-weekly employment levels. As a result, our measure of realized maximum employment could understate what applicants had in mind, even without systematic expectation bias.[15] To account for this possible mismatch in frequency, we perform a robustness

[13] To reduce the influence of some major outliers, we winsorize $\tilde{\mu}_{it}$ at 100 employees; fewer than one percent of observations exceed this threshold.

[14] Similar to the case with the expected maximum employment, we winsorize the realized maximum employment values at 100 employees – non-winsorized cases again representing more than 99% of the observations.

[15] The expected maximum of n i.i.d. random variables is non-decreasing in n. Although this does not always hold for non-i.i.d. variables, the expected maximum can still rise with n.

check by constructing theoretical upper bounds for the expected value of the maximum of realized employment under different measurement frequencies—specifically, $\tau = 12$ (monthly) and $\tau = 24$ (approximately bi-weekly)—as detailed in Section Appendix B. Because these upper bounds exceed the true expected maximum and are not necessarily tight, this is a conservative approach. Even with these upper bounds, the results show that our findings are robust to measurement frequency, mainly because employment within firms exhibits strong persistence across sub-annual periods (conditional on entry). For instance, conditional on entry, the average pairwise correlation between quarterly employment levels is nearly 90%, and likely much higher at monthly frequency. This persistence supports the use of quarterly data to approximate maximum realized employment over a 12-month horizon.

To examine expected and realized entry over horizons other than 12 months, we use the planned first wage payment date reported in each application to measure expected entry. This variable enables us to identify the expected quarter of entry and to compute cumulative expected entry rates by quarter following application. We then compare these with cumulative realized entry rates over time, constructed using the actual quarter of the first wage payment.

4 Analysis

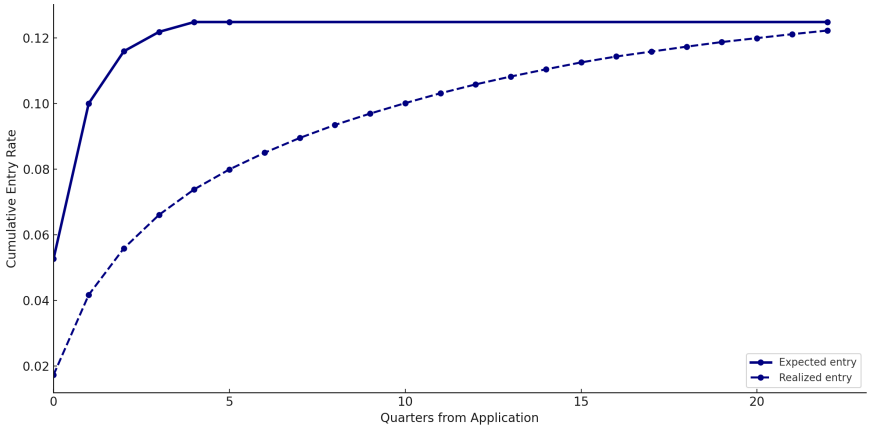
We begin by exploring whether potential entrants have systematic errors in their entry and employment expectations at the time of application.

4.1 Expectations and realizations of entry and initial size

Figure 1 compares cumulative entry rates over time based on expected wage payment quarters and realized entry for all applications. The expected rate rises quickly after application, plateauing at about 12.5% by quarter 6, after which applicants no longer anticipate a first wage payment. The realized rate increases more gradually, continuing past quarter 6 and partly closing the gap with expectations. This suggests applicants overpredict near-term employer business formation, though some delayed entries occur. The persistent gap indicates systematic overestimation of entry likelihood in early-stage expectations (Puri and Robinson (2007); Koellinger et al. (2007)).

Figure 2 plots the average realized employment against the average expected employment 12

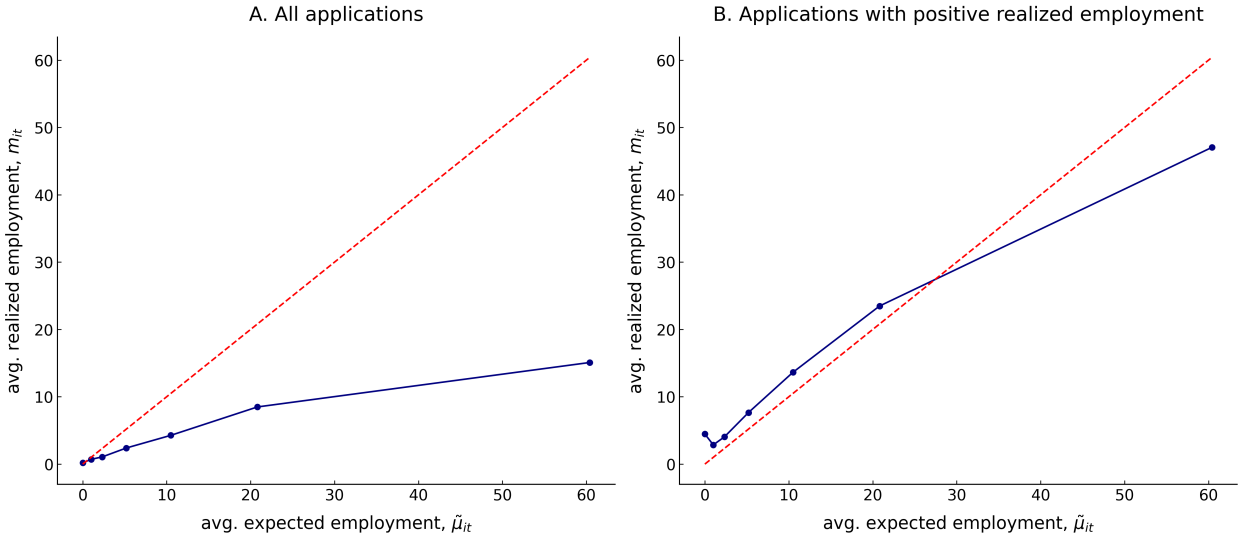
Figure 1: Expected versus realized cumulative entry rate – All applications



Notes: The graph shows expected and realized cumulative entry rate over time starting from the application quarter (corresponding to 0 in the x-axis). All applications between 2017m4-2021m12 are included.

months after business applications across bins of the expected employment distribution. Panel A uses the full sample, while Panel B restricts to applications with positive realized employment. The 45-degree line serves as a benchmark.

Figure 2: Average expected versus realized employment within percentiles of expected employment



Notes: Each point shows the average realized employment against average expected employment within bins of expected employment. The bins correspond to point 0, and the intervals defined by pseudo-percentiles (pct): (0,50th pct), [50th pct,75th pct), [75th pct,90th pct), [90th pct,95th pct), [95th pct,99th pct), and [99th pct,100th pct]. The dashed line is the 45-degree line.

In Panel A, realized employment falls short of expectations throughout the distribution, indicating widespread overestimation. Although realized size rises with expectations, values remain

consistently below the 45-degree line. Panel B shows that among actual entrants realized employment generally exceeds expectations, except in the top bin. The underestimation at the leftmost point (the bin corresponding to expected employment = 0) reflects that only positive errors are possible when no employment is forecast, though their magnitude is not predetermined. Other points indicate that errors stem not only from entry failures but also from underprediction among applicants with lower expected employment. Overall, the figure reveals systematic errors in employment projections, with both entry (extensive margin) and size (intensive margin) contributing to gaps between expectations and outcomes.

4.1.1 Components of the expectation error

We next examine the components of the expectation error based on (5). Table 1 reports average expected employment ($\tilde{\mu}_{it}$), realized employment (m_{it}), and their difference (δ_{it}) across groups defined by expected and actual entry outcomes. It also includes group shares.

Most applications (84.1%) neither expect nor achieve entry. These applications likely represent plans to start nonemployer businesses. Only 3.4% of applications both expect entry and enter, while 8.9% expect entry but do not realize it. Another 3.6% enter despite expecting not to. The average entry expectation error is $\hat{\bar{\omega}}_t = 0.089 - 0.036 = 0.053$, and the total misprediction rate is $|0.089| + |-0.036| = 0.125$.

In the full sample, applicants expect to hire 0.53 employees on average but realize only 0.39, yielding a mean overprediction of 0.14. Disaggregating by entry outcomes reveals sharper differences. Those who expect to enter but do not ($\{\tilde{e}, e\} = \{1, 0\}$) overpredict by 4.087 employees. Conversely, those who unexpectedly enter ($\{0, 1\}$) underpredict by 4.462 on average.

Among those who both expect and achieve entry ($\{1, 1\}$), the average error is –1.903, indicating underprediction overall. Yet, this group is quite heterogeneous: 47% underpredict by 6.237 employees, 24% overpredict by 3.872, and 29% forecast employment perfectly–the latter has a small average expected (and realized) employment (about 2 employees) compared to the other two groups, reflecting the fact that the scale of small business operations may be easier to predict. For the broader group (12.3% of the sample) who expect entry ($\{1, \cdot\}$), the average error is 2.431 – resulting from both overestimation by those who fail to enter and those that enter and underestimate.[16]

[16] This group corresponds to the pairs $\{1, 1\}$ and $\{1, 0\}$ in Table 1.

Table 1: Expectation error and its decomposition

	Sample average of:			
Sample	$\tilde{\mu}_{it}$	m_{it}	δ_{it}	% of N
$\{\tilde{e}, e\} = \{\cdot, \cdot\}$	0.530	0.388	0.142	100%
	(0.001)	(0.001)	(0.001)	
$\{\tilde{e}, e\} = \{1, 1\}$	4.779	6.683	-1.903	3.4%
	(0.011)	(0.014)	(0.011)	
$\hat{\bar{\delta}}_t(1,1) < 0$	4.788	11.030	-6.237	1.6%
	(0.014)	(0.026)	(0.018)	
$\hat{\bar{\delta}}_t(1,1) > 0$	8.233	4.361	3.872	0.8%
	(0.032)	(0.021)	(0.019)	
$\hat{\bar{\delta}}_t(1,1) = 0$	1.988	1.988	0.000	1.0%
	(0.007)	(0.007)	(0.000)	
$\{\tilde{e}, e\} = \{1, 0\}$	4.087	0.000	4.087	8.9%
	(0.006)	(0.000)	(0.006)	
$\{\tilde{e}, e\} = \{0, 1\}$	0.000	4.462	-4.462	3.6%
	(0.000)	(0.011)	(0.011)	
$\{\tilde{e}, e\} = \{0, 0\}$	0.000	0.000	0.000	84.1%
	(0.000)	(0.000)	(0.000)	

Notes: Standard errors are in parentheses. N = 16,800,000–rounded for disclosure avoidance. $\delta_{it} = \tilde{\mu}_{it} - m_{it}$.

Overall, expectation errors reflect both extensive-margin gaps (entry failures and surprises) and intensive-margin deviations among entrants, with significant heterogeneity even within the group that correctly forecast entry. The coexistence of over- and underestimation is unlikely to stem solely from bureaucratic conservatism or strategic understatement (e.g., to reduce perceived tax exposure). If such motives were dominant, we would expect aggregate and wide-spread understatement. The bidirectional variation is instead consistent with models linking entrepreneurial outcomes to diverse

expectations and strategic orientations.[17]

4.1.2 Decomposition of the variance of expectation errors

Table 2 decomposes the sample variance of the employment expectation error using the identity

$$Var(\delta_{it}) = Var(\tilde{\mu}_{it}) + Var(m_{it}) - 2Cov(\tilde{\mu}_{it}, m_{it}). \tag{7}$$

In the full sample, the variance of the expectation error is modest (12.69), driven by the prevalence of cases with zero expected *and* actual employment. Expected employment explains 75% of the error variance, compared to 68% for realized employment. The covariance term contributes –43%. The correlation between expectations and realizations is positive but not too high (0.30). Among applications with positive expected employment, the error variance is 69.37. The contribution of expectations increases to 88%, while that of realizations drops to 58%. The covariance share is –46%, and the correlation is slightly higher at 0.32. These patterns indicate that variation in expectation error is more closely linked to variation in expectations than outcomes, and expectations are positively but not highly correlated with outcomes, even among those anticipating employment.

Table 2: Decomposition of the variance of the expectation error

Sample	$Var(\delta_{it})$	$\frac{Var(\tilde{\mu}_{it})}{Var(\delta_{it})}$	$\frac{Var(m_{it})}{Var(\delta_{it})}$	$\frac{-2\,Cov(\tilde{\mu}_{it},m_{it})}{Var(\delta_{it})}$	$Corr(\tilde{\mu}_{it}, m_{it})$
All applications	12.69	0.75	0.68	-0.43	0.30
Positive Exp. ($\tilde{\mu}_{it} > 0$)	69.37	0.88	0.58	-0.46	0.32

Notes: The functions $Var(\cdot)$, $Cov(\cdot,\cdot)$, and $Corr(\cdot,\cdot)$ give the *sample* variance, covariance, and correlation of their arguments, respectively. "Positive Exp." refers to applications that report positive expected maximum employment.

4.2 Expectation errors and application characteristics

Next, we turn to an analysis of how expectation errors vary by selected application characteristics to illustrate the heterogeneity in errors. We note that the comparisons in this section are unconditional; later, we consider a regression analysis that controls for various application characteristics and other observables to assess the partial correlations without a causal interpretation.

[17] See, e.g., Kihlstrom and Laffont (1979); Lazear (2005); Baron (2004); Astebro, Jeffrey, and Adomdza (2007); Parker (2009); Hurst and Pugsley (2011); Manso (2016).

4.2.1 Legal form of organization

Table 3 summarizes expected and realized employment and associated expectation error metrics by legal form of organization (LFO), for both all business applications and the subset with positive expected employment. We combine all LLCs into a separate group. All non-LLCs are classified into one of the three distinct groups: sole proprietorships, partnerships, or corporations. The results reveal substantial differences in expectation errors across organizational types.

Among all applications, corporations and LLCs exhibit the highest average expected and realized employment, consistent with their more formal structure and growth orientation. Expectation errors, however, vary by form. Sole proprietorships report the lowest realized employment and a relatively large average error, while corporations exhibit a much smaller average error. Restricting to applications with positive expected employment, average errors are higher across all legal forms: they exceed two employees in all cases. Thus, even among more sophisticated forms (corporations and LLCs) the average error remains high, underscoring misalignment between plans and outcomes.

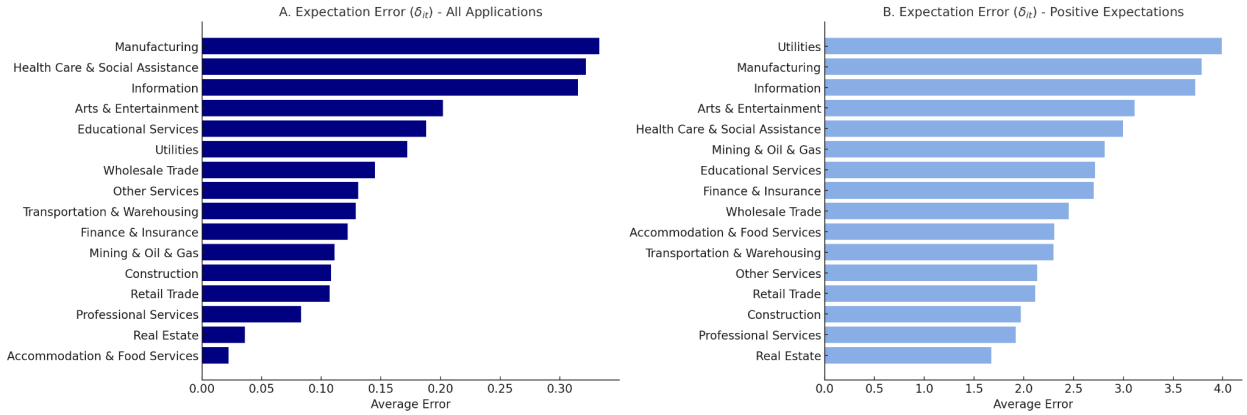
Table 3: Summary statistics for expectation errors by legal form of organization

	All applications			Positive expectations ($\tilde{\mu}_{it} > 0$)		
Legal form	$\tilde{\mu}_{it}$	m_{it}	δ_{it}	$\tilde{\mu}_{it}$	m_{it}	δ_{it}
Sole proprietorship	0.220	0.058	0.161	3.522	1.317	2.205
	(0.001)	(0.000)	(0.001)	(0.003)	(0.001)	(0.003)
Partnership	0.453	0.231	0.222	4.511	1.901	2.610
	(0.001)	(0.001)	(0.001)	(0.007)	(0.003)	(0.008)
Corporation	0.916	0.842	0.075	6.217	3.673	2.544
	(0.002)	(0.003)	(0.002)	(0.007)	(0.005)	(0.008)
LLC	0.521	0.377	0.144	4.763	2.218	2.545
	(0.001)	(0.001)	(0.001)	(0.004)	(0.002)	(0.005)
Other	0.508	0.197	0.311	5.246	2.065	3.181
	(0.002)	(0.001)	(0.002)	(0.013)	(0.005)	(0.014)

Notes: Entries are sample means by legal form of organization. LLCs are classified under one category and non-LLCs are presented in 4 distinct groups (sole proprietorship, corporations, partnerships, others). "Positive expectations" refers to applications reporting positive expected maximum employment. Standard errors are in parentheses. $\delta_{it} = \tilde{\mu}_{it} - m_{it}$.

Differences in forecast errors across organizational forms may reflect variation in information access, venture complexity, business planning capabilities, or access to external advising. The findings for expectation errors align with models in which organizational form proxies for entrepreneurial sophistication or planning capacity (e.g., Hurst and Pugsley (2011)), but also indicate the potential

Figure 3: Average expectation error by sector



Notes: Panels A and B show average expectation error (δ_{it}) by sector.

role of frictions and uncertainty in expectation formation across the full legal-form spectrum.

4.2.2 Sectoral heterogeneity

Figure 3 and Table A.1 in Appendix A reveal substantial sectoral variation in expectation errors, pointing to the interplay between entrepreneurial judgment and the planning environments specific to different sectors.

Manufacturing, Health Care, and Information show the largest raw expectation errors, indicating systematic overprediction. These sectors typically involve longer planning horizons, regulatory complexity, or uncertainty in demand realization, which may lead entrepreneurs to overestimate their initial employment. By contrast, Professional Services, Real Estate, Retail, and Accommodation and Food Services display low errors, possibly due to more standardized business models or smaller initial staffing needs, which may help entrepreneurs form more accurate expectations.

When focusing on applicants with positive expected employment, these discrepancies largely remain. For instance, Utilities and Manufacturing display large expectation errors, suggesting that even conditional on hiring plans, realized hiring often falls short. This pattern may reflect lags in capital deployment, permitting, or sales acquisition in execution-heavy sectors. Sectoral patterns suggest that errors are related to sector-specific environments, supporting theories in which industry features play a central role in shaping entrepreneurial forecasts (e.g., Wu and Knott (2006); Bloom (2009); Hall and Woodward (2010)).

4.2.3 High-tech business applications

As an example of finer industry-level differences in expectation errors, Table 4 reports employment expectations and associated errors by whether applicants intend to operate in high-tech industries. High-tech industries are defined using the concentration of Science, Technology, Engineering, and Math (STEM) occupation employment as in Goldschlag and Miranda (2020) and Hecker (2005).[18]

High-tech applicants report slightly higher expected employment than non–high-tech applicants, yet their realized employment is lower on average, leading to a larger mean expectation error. Among applicants with positive expected employment, high-tech firms overpredict realized employment more, relative to non–high-tech firms. Their average expectation error is 2.9, half an employee higher than that of non–high-tech firms (2.4).

Table 4: Summary statistics for expectation errors by high-tech industry status

	All applications			Positive expectations ($\tilde{\mu}_{it} > 0$)		
High-tech status	$\tilde{\mu}_{it}$	m_{it}	δ_{it}	$\tilde{\mu}_{it}$	m_{it}	δ_{it}
No	0.535	0.395	0.141	4.280	1.863	2.417
	(0.001)	(0.001)	(0.001)	(0.006)	(0.005)	(0.006)
Yes	0.575	0.343	0.232	4.193	1.272	2.921
	(0.004)	(0.003)	(0.005)	(0.028)	(0.016)	(0.028)

Notes: Entries are sample means by high-tech industry status. "Positive expectations" refers to applications that report positive expected maximum employment. Standard errors are in parentheses. $\delta_{it} = \tilde{\mu}_{it} - m_{it}$.

These patterns suggest that employment expectations are generally less well calibrated in high-tech industries. The results are consistent with theories emphasizing the greater volatility and information frictions faced by technology-oriented startups (e.g., Hall and Woodward (2010)), suggesting that high-tech entrepreneurs may form expectations under heightened uncertainty or optimism.

[18]The 4-digit NAICS industries that are classified as high-tech are: 3341 (Computer and Peripheral Equipment Manufacturing), 3342 (Communications Equipment Manufacturing), 3344 (Semiconductor and Other Electronic Component Manufacturing), 3345 (Navigational, Measuring, Electromedical, and Control Instruments Manufacturing), 3364 (Aerospace Product and Parts Manufacturing), 5112 (Software Publishers), 5182 (Data Processing, Hosting, and Related Services), 5191 (Other Information Services), 5413 (Architectural, Engineering, and Related Services), 5415 (Computer Systems Design and Related Services), 5417 (Scientific Research and Development Services). See also Business Dynamics Statistics-High Tech for further details.

4.3 The evolution of expectation errors around the Covid-19 Pandemic

Next, we consider the time series patterns in expectations, actuals, and errors to provide an overview of how they evolved before and after the Covid-19 shock. Figure 4 shows the evolution of monthly average expected and realized employment (Panels A, C) and the corresponding average expectation error (Panels B, D), while Figure 5 plots the time paths of their standard deviation and coefficient of variation (CV). The top panels cover all applications; the bottom panels restrict to those with positive expected employment.

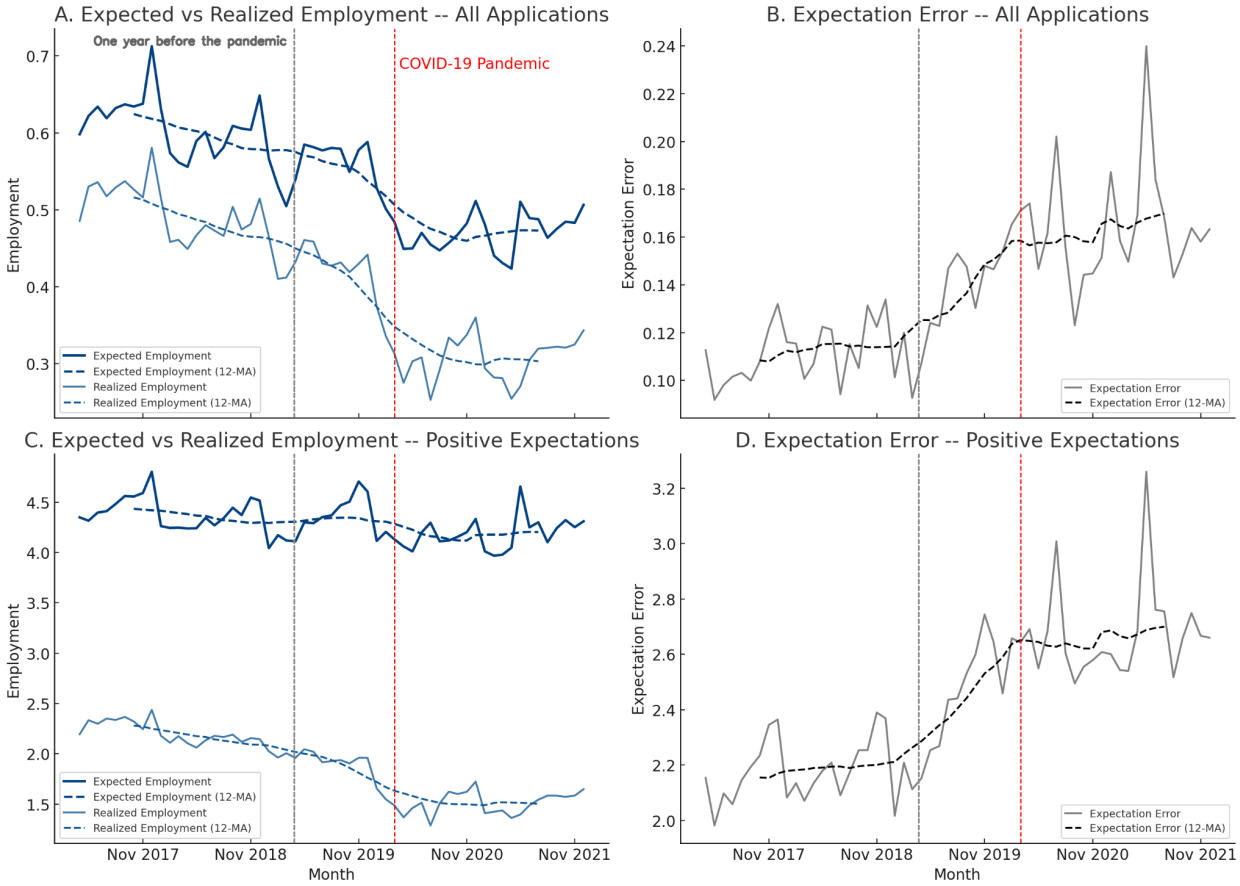
In both samples, expected employment exceeds realized employment, and both series trend downward between 2017 and 2019, reflecting longer-term declines in startup size. A clearer divergence emerges around March 2019, when the 12-month realization window for new applications begins to overlap with the pandemic onset in March 2020. Applications submitted from March 2019 onward can materialize during the pandemic, so their realized outcomes are partly shaped by this shock. Expectations, however, do not fall as steeply as realizations in 2019–2020, leading to rising overestimation. As the pandemic approaches, both series decline more sharply, but after March 2020 expectations level off sooner, while realizations continue to fall before stabilizing, leaving a persistent post-COVID wedge and elevated errors.

Dispersion patterns (Figure 5) show a decline in the standard deviation and CV of expectation errors that begins in 2019 and continues through early 2020. The compression is less pronounced in the positive-expectations sample (Panel C). Part of this convergence reflects a surge in applications with zero expected and realized employment, which increase the mass at zero error. By late 2020/early 2021, dispersion rises again, especially for standard deviation and more strongly in the positive-expectation sample. The initial pandemic shock thus induced a convergence of errors as both expectations and realizations fell, but heterogeneity re-emerged as the recovery began, consistent with heterogeneity in local and sectoral exposures.

4.4 Are expectations unbiased and efficient?

The analysis so far has shown that expectation errors are systematic, exhibit significant variation, and vary meaningfully with application characteristics. To more formally evaluate forecast efficiency and bias, we begin with the standard Mincer-Zarnowitz regression (Mincer and Zarnowitz, 1969),

Figure 4: The evolution of average expected and realized employment, and expectation error



Notes: Dashed curves are 12-month centered moving averages (12-MA). Red vertical line indicates March 2020; grey vertical line indicates March 2019.

which tests whether realized outcomes align with expectations.[19] Under the rational expectations hypothesis, forecasts should fully incorporate all available information and be unbiased. This leads to the following specification in levels

$$m_{it} = \alpha + \beta \tilde{\mu}_{it} + \epsilon_{it}. \tag{8}$$

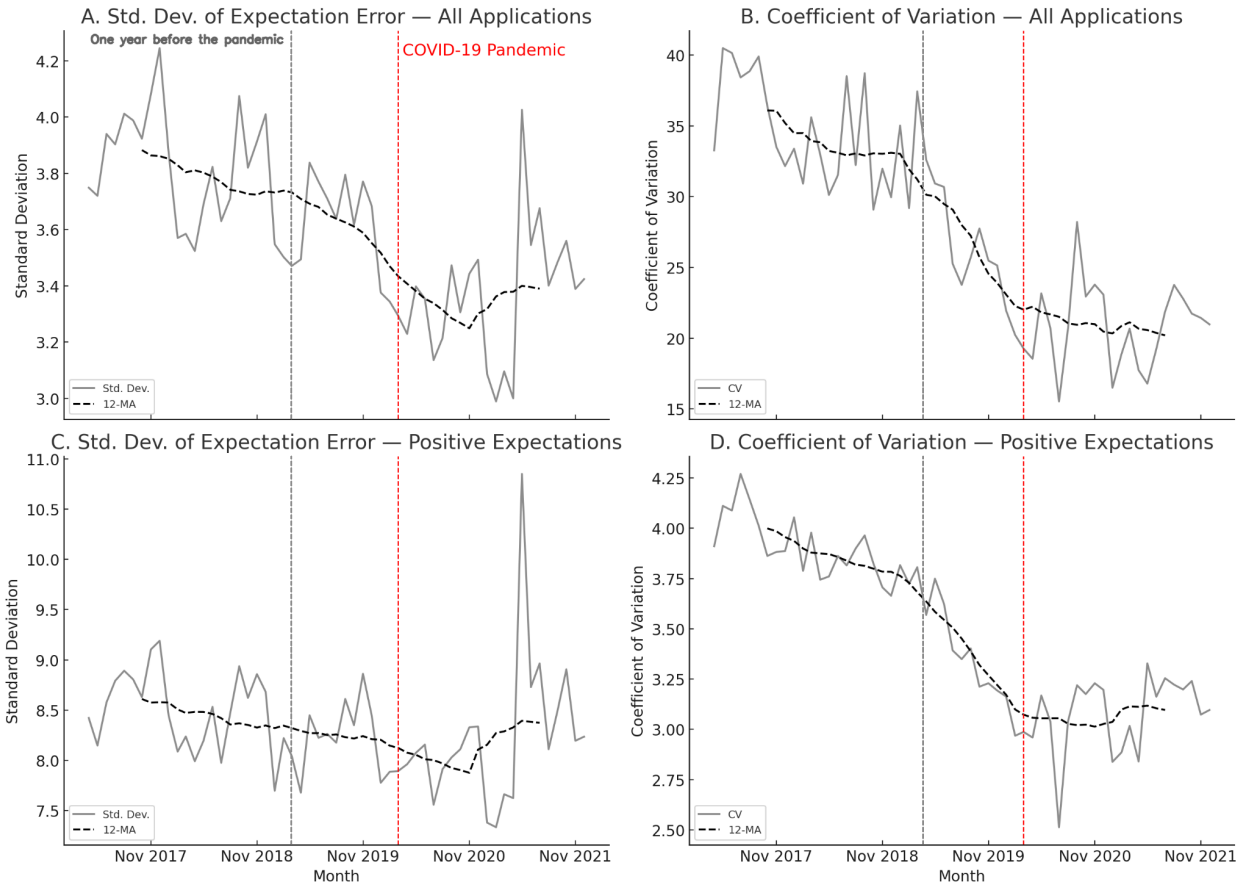
The joint null hypothesis $\alpha = 0$ and $\beta = 1$ implies both unbiasedness (no systematic over- or under-forecasting) and efficiency (forecast errors are orthogonal to the forecast).

As a complementary diagnostic, we also estimate a regression with the expectation error as the dependent variable

$$\delta_{it} = \tilde{\mu}_{it} - m_{it} = \alpha^* + \beta^* \tilde{\mu}_{it} + \epsilon^*_{it}, \tag{9}$$

[19] See also Chapter 3 in Bachmann, Topa, and van der Klaauw, 2022.

Figure 5: The evolution of the standard deviation and coefficient variation of the expectation error



Notes: Dashed curves are 12-month centered moving averages (12-MA). Red vertical line = March 2020; grey vertical line = March 2019.

where $\alpha^* = -\alpha$, $\beta^* = 1 - \beta$, and $\epsilon^*_{it} = -\epsilon_{it}$. While this regression does not substitute for the levels regression (8) in formally testing rationality, it is useful for separately identifying whether forecast errors are systematically biased (via α^*) or depend on the forecast level (via β^*).[20] Under rational expectations, both coefficients should equal zero.

Table 5 reports the results from both regressions. In the levels specification, the slope estimates are far below one (0.29 and 0.26), and the intercepts are significantly positive, rejecting the (joint) null of rational expectations. In the error specification, the slope coefficients (0.71 and 0.74) are each significantly different from zero, further indicating that forecast errors are strongly predictable from the forecasts themselves. The intercepts are again significantly different from zero, consistent

[20] Regression (9) is algebraically equivalent to (8) and thus does not provide an independent test of rational expectations. We report it as a complementary diagnostic because it facilitates interpretation of bias (α^*) and forecast-dependent errors (β^*).

Table 5: Mincer-Zarnowitz test of rational expectations

	All Applications		Positive Exp. ($\tilde{\mu}_{it} > 0$)	
	m_{it}	δ_{it}	m_{it}	δ_{it}
$\tilde{\mu}_{it}$	.288***	.712***	.261***	.739***
	(0.002)	(0.002)	(0.002)	(0.002)
const.	.235***	-.235***	.719***	-.719***
	(0.001)	(0.001)	(0.008)	(0.008)
R^2	0.091	0.381	0.103	0.478
N	16,820,000	16,820,000	2,082,000	2,082,000
F-statistic	68580		171600	
p-value	0.000		0.000	

Notes: Robust standard errors are in parentheses. Stars denote statistical significance: *** p < 0.01, ** p < 0.05, * p < 0.1. F-statistic pertains to the joint hypothesis $H_0 : \alpha = 0, \beta = 1$. N and F-statistic values are rounded for disclosure avoidance.

with unconditional bias.

The R^2 values in the levels regressions (0.09–0.10) indicate that expectations explain a non-trivial share (10%) of the variation in realized employment in a bi-variate setting. However, this alone does not necessarily imply forecast inefficiency, as a large portion of outcome variance may stem from unpredictable shocks even under rational expectations. In contrast, the moderate R^2 values in the error regressions (0.38–0.48) indicate that forecast errors are systematically related to the forecasts themselves—a direct violation of forecast efficiency.

Rejecting the joint null $\alpha = 0$ and $\beta = 1$ in (8) implies that expectations are not *unconditionally* unbiased and efficient. However, this does not necessarily imply that applicants form expectations irrationally. The Mincer–Zarnowitz test assumes that the econometrician observes the same information set as the forecaster. If applicants possess private information (included in $\mathcal{I}_{it}$) that affects both expectations and realized outcomes, the estimated slope may deviate from unity even when $E[m_{it} \mid \mathcal{I}_{it}] = \tilde{\mu}_{it}$. In this case, the departure from $\beta = 1$ reflects differences between the econometrician's and applicants' information sets (an omitted-information problem) rather than a failure of applicants to form conditionally efficient expectations. To address this, we next examine whether forecast errors are predictable using observables that approximate the information available to applicants at the time of filing—a test of *conditional* forecast efficiency.

4.5 Regression analysis of expectation errors

We further assess *conditional* forecast efficiency by examining whether forecast errors are systematically related to information available at the time of application. Under rational expectations, forecasts should fully incorporate the applicant's information set $\mathcal{I}_{it}$, implying $E[\delta_{it} \mid \mathcal{I}_{it}] = 0$. In particular, no variable known at the time of filing should predict the forecast error. To explore this hypothesis, we estimate

$$\delta_{it} = \alpha + \beta\tilde{\mu}_{it} + \gamma' z_{it} + q_t + \iota_{j(i)} + \xi_{s(i)} + \varepsilon_{it}, \tag{10}$$

where $\delta_{it} = \tilde{\mu}_{it} - m_{it}$ is the forecast error and z_{it} denotes observable characteristics of the application and local economic conditions at the time of filing. A positive (negative) coefficient on a component of z_{it} indicates that forecast errors are systematically more positive (negative) when that factor is higher. The specification also includes year–quarter (q_t), four-digit NAICS industry ($\iota_{j(i)}$), and state ($\xi_{s(i)}$) fixed effects to absorb aggregate macroeconomic trends and persistent industry- or state-specific components of forecast errors.

Including $\tilde{\mu}_{it}$ in (10) controls for the mechanical dependence between realized employment and the level of expected employment, ensuring that the coefficients on z_{it}, γ, capture only the incremental predictive content of other information available at filing. A significant estimate of any element of γ therefore indicates a violation of *conditional* forecast efficiency: forecast errors are predictable from information available at the time of forecasting, implying that such information was not fully incorporated into expectations.[21]

Including local demographic, economic, and business conditions in z_{it} is motivated by both theoretical and empirical considerations. Theoretically, models of entry, selection, and learning emphasize that local market conditions shape expected demand, cost structures, and post-entry growth trajectories. Cross-sectional variation in these factors generates heterogeneity in both expectations and realized outcomes across otherwise similar applicants. Empirically, controlling for lagged or contemporaneous local conditions approximates the information set available to applicants at the time of filing and mitigates omitted-information bias arising from spatial sorting or

[21] For the Mincer–Zarnowitz regression to provide a valid test of rational expectations, forecasts must not only incorporate all information available to the forecaster, but also fully summarize the effect of that information on realized outcomes. In other words, $\tilde{\mu}_{it}$ must be a sufficient statistic for $\mathcal{I}_{it}$ with respect to m_{it}. If some element of $\mathcal{I}_{it}$ affects m_{it} through channels not mediated by $\tilde{\mu}_{it}$, then forecast errors may be predictable even when expectations are rational.

local shocks that influence both expectations and outcomes.

4.5.1 Covariates

For the covariates z_{it}, we include four blocks capturing application attributes and local conditions that proxy both the information available to founders and the environment shaping expectations and realizations. The analysis is descriptive, not causal: unobserved traits (e.g., ambition, risk tolerance) may confound associations, and some local variables may lie on proposed channels rather than act as strictly exogenous shifters. To capture information "available at filing", we use pre-application or contemporaneous/lagged local measures. Our goal is not to identify mechanisms but to control for systematic contextual differences so residual variation more plausibly reflects firm or founder heterogeneity. Appendix Table A.2 details variable construction.

Application attributes. Legal form and prior experience proxy for managerial sophistication and resource access. Corporations and LLCs face greater reporting obligations and often employ professional advisers; sole proprietorships and informal partnerships are typically less formal.[22] We include dummies for legal form (non-LLC partnership, non-LLC corporation, non-LLC other, and LLC, with non-LLC sole proprietorship as reference) and an indicator for prior EIN issuance to the applicant: a potential proxy for experience, regulatory familiarity, and serial entrepreneurship. Prior experience may be correlated with better information about a market or industry (Lafontaine and Shaw (2016)) but may also reflect persistent overconfidence. However, in some cases it may also simply indicate a more sophisticated business structure with multiple EINs – though our focus on new business starts reduces this possibility substantially.

Local demographic conditions. Using ACS data as in Dinlersoz et al. (2023), we relate errors to tract-level demographics. Higher college-educated share proxies richer human-capital pools and knowledge spillovers (Moretti, 2004a); non-white share captures entrepreneurial composition, demand shifters, and potential constraints (e.g., financing or hiring frictions) (Fairlie et al., 2012); median age (log) proxies local experience and life-cycle demand. All measures use a two-year lag to ensure availability at filing and mitigate simultaneity.[23]

Local market structure. Market structure shapes competition and stability, and thus may

[22] See U.S. Small Business Administration (2025).

[23] Using a one-year lag for 2017–2020 yields similar results given high persistence; see Dinlersoz et al. (2023).

affect forecast errors. In imperfect competition models, larger markets feature lower markups, lower firm density, higher average firm size, and greater turnover,[24] making outcomes harder to predict and potentially biasing expectations. Low density may also weaken agglomeration benefits and information flows. Higher concentration can grant incumbents market power that entrants underestimate, yet it may also stabilize markets and improve forecasting accuracy; Datta, Iskandar-Datta, and Sharma (2011) show concentrated industries exhibit lower volatility and better information efficiency. To capture market structure prior to application, we construct three proxies from LBD (lagged one year): (i) log firm density (firms per 1,000 population in county–industry (2-digit NAICS)); (ii) the DHS transform of county–industry average employment (local average firm size),[25] and (iii) the top-four-firm employment share.

Local dispersion in firm outcomes. Greater cross-sectional variability can hinder accurate expectations (Bachmann and Bayer, 2014) and foster overconfidence (Wu and Knott, 2006). We compute (lagged one year, LBD) at the county–industry level: (i) the coefficient of variation of firm employment (scale dispersion) and (ii) the coefficient of variation of labor productivity (revenue per employee), capturing heterogeneity in technology, prices, and efficiency.

Local shocks. To gauge contemporaneous shocks, we include an industry–county exposure to COVID-19:

$$\text{COVIDShock}_{i,t} \;=\; (1 - \text{WFHshare}_{j(i)}) \times \frac{\text{NewCases}_{c(i),t}}{\text{Pop}_{c(i),2019}},$$

where $(1 - \text{WFHshare}_{j(i)})$ (from Dingel and Neiman 2020) measures the share of non-teleworkable jobs in industry $j(i)$, and the second term is the per-capita new cases in county $c(i)$ and quarter t. Higher values indicate less teleworkable industries in harder-hit areas. Following Papanikolaou and Schmidt (2022), the negative association between $(1 - \text{WFHshare}_{j(i)})$ and subsequent industry employment growth supports interpreting this as adverse local conditions at filing. $\text{COVIDShock}_{i,t}$ enables tests of whether forecasting errors vary with exposure to an adverse shock.[26] We also control for prior-quarter county GDP growth to separate local exposure from broader macro fluctuations.

Together, these covariates span: (i) application-level information quality, (ii) local market com-

[24] See Campbell and Hopenhayn (2005), Dinlersoz (2004), Nocke (2006), Asplund and Nocke (2006).

[25] Formally, $DHS(\overline{Emp})_{j,c,t-1} = \frac{\overline{Emp}_{j,c,t-1} - \overline{Emp}}{0.5(\overline{Emp}_{j,c,t-1} + \overline{Emp})}$, where j is 2-digit NAICS, c county, t application year, $\overline{Emp}_{j,c,t-1}$ the cell average, and $\overline{Emp}$ the overall average of firm employment.

[26] The COVID shock can affect both the level of activity and uncertainty; moreover, uncertainty itself can operate like a level (aggregate-demand) shock (see Leduc and Liu (2016)).

petition and agglomeration, (iii) cross-sectional heterogeneity in outcomes, and (iv) contemporaneous local shocks, enabling a descriptive assessment of how each correlates with entrepreneurial expectations, in line with work on business formation and managerial forecasting.

4.5.2 Results

Table 6 summarizes the results for the full sample of applications and the subsample with positive expected employment. In both groups, forecast errors are strongly and significantly related to expected employment and observable information at the time of filing, z_{it}, indicating that applicants do not fully incorporate observable information when forming expectations. The coefficient on expected employment, $\tilde{\mu}_{it}$, is stable around 0.75 in both samples, implying that realized employment increases less than one-for-one with expectations. This suggests larger errors at higher forecast levels—i.e., systematic overprediction at the margin.

Corporations and LLCs, are associated with less overprediction, suggesting that more formally structured businesses may engage in more realistic planning. A higher local college share is similarly associated with less overprediction, while a higher nonwhite share is associated with more overprediction. These patterns may reflect variation in access to planning resources, information, or networks.

Local market structure also matters: greater firm density and employment concentration are associated with less overprediction, consistent with information advantages from denser firms and stable market. Greater employment and productivity dispersion are linked to greater overprediction, consistent with heightened uncertainty in less structured markets.

An increase in the COVID-19 shock variable is associated with higher overprediction, while GDP growth has no consistent relationship with forecast errors. This suggests that potential entrants tend to overestimate more during adverse economic shocks – consistent with the descriptive analysis in Section 4.3. Overall model fit is higher in the positive expectations subsample ($R^2 = 0.50$ vs. 0.40), indicating that observables explain a larger portion of the variation in forecast errors among business applicants planning to hire.

Table 6: OLS regressions for expectation error, δ_{it}

Variable	All Applications	Pos. Exp. ($\tilde{\mu}_{it} > 0$)
$\tilde{\mu}_{it}$	0.746***	0.765***
	(0.002)	(0.002)
partnership	-0.065***	-0.227***
	(0.004)	(0.030)
corporation	-0.505***	-1.195***
	(0.003)	(0.013)
llc	-0.213***	-0.794***
	(0.001)	(0.010)
other	-0.088***	-0.140***
	(0.005)	(0.030)
prior EIN	0.040	0.286
	(0.048)	(0.156)
college share	-0.334***	-1.097***
	(0.007)	(0.040)
nonwhite share	0.225***	1.112***
	(0.003)	(0.020)
ln(median age)	0.011*	0.039
	(0.005)	(0.029)
ln(firm density)	-0.045***	-0.145***
	(0.002)	(0.012)
dhs(avg. firm emp.)	-0.036***	-0.151***
	(0.002)	(0.012)
top4 emp. share	-0.087***	-0.224***
	(0.006)	(0.036)
cv(firm emp.)	0.010***	0.032***
	(0.000)	(0.002)
cv(labor prod.)	0.003	0.019
	(0.002)	(0.013)
local gdp growth	0.024	0.260**
	(0.018)	(0.088)
local covid-shock	0.025***	0.066***
	(0.001)	(0.009)
const.	-0.416***	-1.341***
	(0.023)	(0.135)
year-quarter FE	Y	Y
industry FE (4-digit NAICS)	Y	Y
state FE	Y	Y
R^2	0.403	0.502
N	14,290,000	1,786,000

Notes: The omitted category for legal form of organization is sole proprietorship. Robust standard errors are in parentheses. Stars denote statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. N values are rounded for disclosure avoidance. The difference in the number of observations (N) between this table and Table 5 is attributable to missing values in some covariates.

4.6 Discrete entry choice and conditional employment outcome

OLS regressions used earlier treat zero employment outcomes as genuine outcome values (since these zeroes indeed reflect zero employment generation), but these zeros reflect underlying non-entry—the outcome of a latent decision process (see Section 2). As a result, OLS does not capture the potential non-linearity in the entry decision or account for selection into positive employment. Entrants may differ systematically from non-entrants along unobserved dimensions, introducing potential selection bias. To explore these issues further, we estimate models that separately capture the entry decision and the employment outcome conditional on entry.

Following the conceptual framework, we assume the entry and maximum employment outcomes are generated by the system

$$\begin{aligned} e_{it} &= I(V_{it} > 0) = I(\eta' r_{it} + u_{it} > 0), \quad u_{it} \sim N(0, \sigma_u^2), \\ m_{it} &= e_{it} \times (\phi' x_{it} + \epsilon_{it}), \quad \epsilon_{it} \sim N(0, \sigma_\epsilon^2), \end{aligned} \tag{11}$$

where $V_{it} = \eta' r_{it} + u_{it}$ and $Corr(u_{it}, \epsilon_{it}) = \rho$.

If $\rho = 0$, unobserved factors influencing the entry decision are uncorrelated with those affecting employment conditional on entry. In this case, consistent estimates of the coefficient vectors η and ϕ can be obtained using a two-part model; for instance, a Probit model for the entry probability and a linear model for employment conditional on entry (see, e.g., Belotti, Deb, Manning, and Norton (2015)). However, if $\rho \neq 0$, unobserved heterogeneity (e.g., ability or ambition) may influence both entry and employment, potentially biasing OLS estimates of η.

To account for potential selection on unobservables, we apply Heckman's maximum likelihood estimation procedure, which allows for $\rho \neq 0$. Although the Heckman model identifies parameters via functional form, identification is stronger when there exists at least one variable that affects entry but not post-entry employment (an exclusion restriction), such that r_{it} includes variables excluded from x_{it}.

Such exclusion restrictions can be motivated by the structure of the entry decision. Let the latent net value of entry be, $V_{it} = W_{it} - \kappa_{it}$, where W_{it} is the gross value of operating and κ_{it} is a sunk entry cost. Variables that shift κ_{it} but do not directly shift the determinants of operating value or scale can therefore serve as exclusion restrictions: they affect the extensive-margin entry decision but, conditional on entry and controls, do not affect initial employment m_{it}. In our setting,

this requires that the excluded variables operate through fixed entry frictions rather than through post-entry demand or production conditions (and are not correlated with unobserved determinants of employment conditional on entry).[27]

To construct plausible exclusionary restrictions, we use data on procedural entry costs from Arizona State University's Doing Business in North America (DBNA) database. This dataset reports time and monetary costs for the legal steps required to start a domestic limited liability company (LLC), including name registration, filing documents, fulfilling publication requirements, and securing state-level business registration. These costs vary geographically and are sourced from local and state agencies. As they pertain to initial business formation but not operational scale, they provide a natural set of instruments for the entry decision.[28] A limitation of the DBNA data is that it pertains specifically to LLCs and is available only for a set of 83 relatively large U.S. cities spanning all 50 states. However, our microdata includes a large number of LLC applications in these cities, enabling robust estimation for this subset.

4.6.1 Two-part models

The two-part model results in Tables 7 provide a more nuanced view of how expectations relate to both the probability of becoming an employer and the size of realized employment conditional on entry. The first part estimates a Probit model for entry, and the second part estimates generalized linear models (GLMs) for realized employment using identity links with both Gaussian and Gamma error distributions. The Gaussian specification corresponds to the OLS regression and the Gamma specification is implemented for robustness, given the skewed distribution of realized initial employment conditional on entry.

Across all specifications, expected maximum employment ($\tilde{\mu}_{it}$) is a highly significant predictor of both entry and post-entry employment size. In the full sample (all applications – left panel), the Probit coefficient on expectations is 0.035, while the GLM coefficients range from 0.503 (Gamma) to

[27] A potential concern is that entry frictions could induce short-term hiring to complete startup procedures (e.g., legal, administrative, or consulting support). However, to the extent that m_{it} is measured from payroll-based employer records that exclude non-payroll administrative or contracting support, this mechanical channel does not affect the observed intensive-margin outcome.

[28] The steps/procedures include: (1) reserving/registering the name of the LLC, (2) choosing/assigning a registered agent, (3) filing the articles of incorporation/organization/formation, (4) completing state LLC publication requirements, (5) filing the initial statement of information, (6) creating an LLC operating agreement, (7) obtaining a state identification number, and (8) fulfilling additional county/city-level requirements. See the DBNA website for methodological details.

0.675 (Gaussian). This implies that realized employment increases with expectations, but at a less-than-proportional rate—consistent with systematic overprediction at the margin. In the positive-expectations subsample (right panel), the GLM coefficients rise to 0.763–0.887, suggesting stronger alignment between expectations and realizations among applicants who explicitly anticipated hiring.

The two-part model results help reconcile the average tendency of entrants to underpredict employment with the positive relationship between expected and actual employment. The GLM estimates show that while applicants with higher expectations do tend to hire more, the increase in realized employment is less than proportional to the increase in expected employment. As a result, forecast errors rise with expected employment at the margin: applicants expecting to hire more workers tend to overpredict more. This pattern implies that forecast errors among entrants are not uniform, but vary with the scale of expectations, helping explain how underprediction can be common on average among entrants even as overprediction becomes more pronounced at higher expectation levels (Panel B of Figure 2).

Other covariates provide additional insights. Compared to sole proprietors, applicants registering as LLCs, corporations, or partnerships are significantly more likely to enter and, conditional on entry, expect and achieve higher employment. In the full sample, LLCs and corporations are associated with 1.7–2.1 more employees (Gaussian GLM). These findings are broadly consistent with more formal organizational types pursuing more ambitious or better-resourced plans.

Prior EIN holders are more likely to enter, but exhibit smaller or no significant differences in post-entry size. In the positive-expectation sample, the association with entry is even negative, and their conditional employment is slightly lower. One interpretation is that applicants with prior EINs may be more experienced but not necessarily more expansive in early-stage hiring.

A higher local college share is associated with both a higher likelihood of entry and greater realized employment, whereas a higher nonwhite share is linked to lower levels of both. Market structure variables show mixed effects. Higher firm density and employment concentration (top 4 share) are associated with more entry but lower post-entry size. In contrast, greater employment dispersion is related to lower entry and higher conditional employment.

The COVID-19 shock is associated with slightly lower entry probabilities but has no significant connection with conditional employment. GDP growth effects vary: in the full sample, it is negatively associated with realized employment under both GLM specifications, while in the

Table 7: Two-part model estimation for actual maximum employment, m_{it} — All applications (left) vs. Positive expectations $\left(\tilde{\mu}_{it} > 0\right)$ (right)

	All applications			Positive expectations ($\tilde{\mu}_{it} > 0$)			
	A. Probit	B. GLM		A. Probit	B. GLM		
		Gaussian	Gamma		Gaussian		Gamma
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\tilde{\mu}_{it}$	0.035***	0.675***	0.503***	0.005***	0.763***	0.766***	0.887***
	(0.000)	(0.004)	(0.002)	(0.000)	(0.005)	(0.005)	(0.004)
partnership	0.269***	2.038***	1.337***	-0.115***	1.351***	1.346***	0.802***
	(0.005)	(0.078)	(0.062)	(0.009)	(0.105)	(0.106)	(0.062)
corporation	1.091***	1.689***	0.751***	0.416***	1.580***	1.604***	0.684***
	(0.002)	(0.027)	(0.016)	(0.005)	(0.035)	(0.035)	(0.019)
llc	0.579***	2.050***	0.896***	0.106***	1.634***	1.630***	0.716***
	(0.002)	(0.026)	(0.016)	(0.004)	(0.034)	(0.035)	(0.018)
other	0.474***	1.668***	0.956***	-0.139***	0.892***	0.863***	0.469***
	(0.005)	(0.090)	(0.074)	(0.011)	(0.132)	(0.130)	(0.089)
prior EIN	0.197***	-0.577*	-0.212	-0.132**	-0.384	-0.367	-0.351**
	(0.024)	(0.244)	(0.163)	(0.041)	(0.455)	(0.464)	(0.133)
college share	0.245***	2.532***	1.273***	0.346***	1.635***	1.601***	0.707***
	(0.004)	(0.074)	(0.049)	(0.009)	(0.101)	(0.103)	(0.059)
nonwhite share	-0.465***	-0.478***	0.048	-0.629***	-0.673***	-0.615***	-0.095*
	(0.003)	(0.048)	(0.035)	(0.006)	(0.068)	(0.071)	(0.040)
ln(median age)	-0.037***	-0.343***	-0.211***	-0.021**	-0.185*	-0.188*	-0.132**
	(0.004)	(0.052)	(0.034)	(0.007)	(0.073)	(0.074)	(0.040)
ln(firm density)	0.141***	-0.258***	-0.113***	0.136***	-0.193***	-0.210***	-0.032
	(0.002)	(0.023)	(0.015)	(0.003)	(0.032)	(0.032)	(0.018)
dhs(avg. firm emp.)	-0.031***	0.619***	0.373***	-0.036***	0.551***	0.519***	0.220***
	(0.002)	(0.024)	(0.018)	(0.003)	(0.033)	(0.034)	(0.019)
top4 emp. share	0.277***	-0.888***	-0.203***	0.312***	-0.748***	-0.754***	-0.222***
	(0.005)	(0.068)	(0.045)	(0.009)	(0.091)	(0.092)	(0.051)
cv(firm emp.)	-0.011***	-0.033***	-0.030***	-0.012***	-0.036***	-0.034***	-0.018***
	(0.000)	(0.003)	(0.002)	(0.000)	(0.005)	(0.005)	(0.003)
cv(labor prod.)	-0.008***	0.035	0.011	-0.009***	0.000	0.002	-0.021
	(0.001)	(0.023)	(0.015)	(0.003)	(0.030)	(0.031)	(0.017)
local gdp growth	0.003	-0.375*	-0.632***	-0.038*		-0.221	
	(0.010)	(0.171)	(0.144)	(0.018)		(0.226)	
local covid-shock	-0.019***	-0.004	-0.006	-0.020***		-0.035	
	(0.001)	(0.017)	(0.011)	(0.002)		(0.024)	
const.	-3.096***	0.734	1.703***	-1.954***	0.678	0.528	0.836*
	(0.232)	(0.402)	(0.382)	(0.205)	(0.457)	(0.458)	(0.400)
year-quarter FE	Y	Y	Y	Y	Y	Y	Y
industry FE (4-digit NAICS)	Y	Y	Y	Y	Y	Y	Y
state FE	Y	Y	Y	Y	Y	Y	Y
Log Likelihood		-6506000	-5475000		-2734000	-2664000	-2266000
N		14,290,000	14,290,000		1,837,000	1,786,000	1,837,000

Notes: Columns (1) and (4) estimate Probit models for $\Pr(m_{it} > 0)$. Columns (2)–(3) and (5)–(7) estimate GLMs for $E[m_{it} \mid m_{it} > 0]$, assuming Gaussian (2,5,6) and Gamma (3,7) distributions. In the right panel, the MLE with a Gamma specification using the full set of controls analogous to column (6) failed to converge, so only the reported Gamma specification (7) is shown. The omitted legal form category is sole proprietorship. Standard errors in parentheses (robust in right-panel columns (4)–(7)). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Log likelihood is the total log likelihood of Probit and GLM components. N and log likelihood values are rounded for disclosure avoidance.

positive-expectation sample the coefficient is negative but insignificant.

Overall, the two-part model results show that expectations strongly correlate with realized outcomes, but forecast errors remain substantial and systematic. Year-specific bi-variate two-part model estimates (Table 8) show stable expectation-outcome relationships from 2017–2021, but with changing magnitudes over time. In the full sample, Probit and GLM coefficients on $\tilde{\mu}_{it}$ decline over time, possibly reflecting rising uncertainty during the pandemic. In the positive-expectation subsample, Probit coefficients remain small but significant, while GLM coefficients are larger, though they also decline. Overall, expectations remain predictive, but their association with realizations weakens after COVID.

4.6.2 Heckman models

Table 9 presents Heckman maximum likelihood estimates for LLC applicants, separately for the full sample and the subset with positive expected employment. The selection equation includes *ltime*, the log of local procedural days required to establish an LLC, as an exclusion restriction.[29]

In both samples, expected maximum employment ($\tilde{\mu}_{it}$) is a strong and statistically significant predictor of both entry and conditional realized employment. As in the two-part models, outcome equation coefficients are well below one, confirming systematic overprediction: realized employment rises with expectations but less than one-for-one.

In regressions with all covariates, the coefficients on $\tilde{\mu}_{it}$ in the selection equation also differ across samples: 0.033 in the full sample and 0.007 in the positive-expectation subsample. This reflects reduced variation in entry intent when conditioning on positive expectations. Together, these results indicate that expectations remain predictive of outcomes in the Heckman framework, as in the case of the two-part models.

The exclusion restriction *ltime* is negative and significant in the full sample, suggesting that greater procedural burdens deter entry. It is positive but not statistically insignificant in the positive-expectation sample. One interpretation is that administrative burdens may matter more for marginal applicants than for those with clearer hiring plans.

Selection correction estimates point to negative selection in both samples. The estimated corre-

[29] This variable is the logarithm of the estimated number of days required to complete LLC startup tasks in a given city; see DBNA website.

Table 8: Two-part model estimation by year

A. All Applications

Variable	2017	2018	2019	2020	2021
			Probit		
$\tilde{\mu}_{it}$	0.051***	0.052***	0.052***	0.047***	0.042***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
const.	-1.411***	-1.436***	-1.469***	-1.595***	-1.617***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
			GLM		
$\tilde{\mu}_{it}$	0.812***	0.806***	0.789***	0.768***	0.759***
	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)
const.	3.931***	3.749***	3.693***	3.632***	3.626***
	(0.027)	(0.023)	(0.022)	(0.022)	(0.021)
Log Likelihood	-1202000	-1771000	-1680000	-1672000	-1991000
N	2,003,000	3,084,000	3,096,000	3,843,000	4,791,000

B. Positive Expectations ($\tilde{\mu}_{it} > 0$)

Variable	2017	2018	2019	2020	2021
			Probit		
$\tilde{\mu}_{it}$	0.006***	0.006***	0.007***	0.006***	0.006***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
const.	-0.478***	-0.514***	-0.570***	-0.731***	-0.763***
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
			GLM		
$\tilde{\mu}_{it}$	0.873***	0.864***	0.852***	0.827***	0.813***
	(0.012)	(0.010)	(0.010)	(0.011)	(0.010)
const.	2.823***	2.713***	2.540***	2.547***	2.609***
	(0.050)	(0.043)	(0.042)	(0.045)	(0.042)
Log Likelihood	-506500	-730500	-663000	-609200	-727500
N	283,000	421,000	403,000	436,000	538,000

Notes: Panel A reports estimates for all applications; Panel B restricts to applications with positive expected employment ($\tilde{\mu}_{it} > 0$). Each panel includes a Probit model for $\Pr(m_{it} > 0)$ and a GLM with Gaussian specification for $E[m_{it} \mid m_{it} > 0]$. Log likelihood is the total log likelihood of both stages. Standard errors in parentheses. *** $p < 0.01$. N and Log Likelihood values are rounded for disclosure avoidance.

Table 9: Heckman maximum-likelihood estimation for the sample of LLC applications

	A. Selection Equation				B. Outcome Equation			
	All		Pos. Exp. ($\tilde{\mu}_{it} > 0$)		All		Pos. Exp. ($\tilde{\mu}_{it} > 0$)	
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\tilde{\mu}_{it}$	0.043***	0.033***	0.013***	0.007***	0.634***	0.563***	0.765***	0.709***
	(0.001)	(0.001)	(0.001)	(0.000)	(0.016)	(0.015)	(0.016)	(0.016)
prior EIN		0.378***		-0.707***		-0.898		-2.306
		(0.088)		(0.204)		(1.071)		(1.675)
college share		0.267***		0.416***		2.730***		2.010***
		(0.013)		(0.027)		(0.274)		(0.415)
nonwhite share		-0.499***		-0.680***		1.287***		0.256
		(0.010)		(0.021)		(0.213)		(0.333)
ln(med. age)		-0.067***		-0.033		0.075		0.106
		(0.012)		(0.023)		(0.228)		(0.346)
ln(firm dens.)		0.119***		0.118***		-1.012***		-0.920***
		(0.007)		(0.015)		(0.141)		(0.225)
dhs(emp.)		0.010		-0.007		0.056		0.096
		(0.007)		(0.015)		(0.139)		(0.208)
top4 emp. share		0.252***		0.401***		-3.004***		-2.427**
		(0.032)		(0.066)		(0.607)		(0.905)
cv(emp.)		-0.009***		-0.014***		0.094***		0.033
		(0.001)		(0.002)		(0.020)		(0.031)
cv(labor prod.)		-0.006		-0.284*		0.157		0.220
		(0.006)		(0.011)		(0.124)		(0.175)
gdp growth		0.006		-0.043		0.637		0.509
		(0.006)		(0.048)		(0.527)		(0.705)
covid shock		-0.023***		-0.148*		0.023		0.020
		(0.004)		(0.007)		(0.066)		(0.103)
ltime	-0.006***	-0.011***	-0.005	-0.010				
	(0.002)	(0.003)	(0.004)	(0.006)				
const.	-1.711***	-1.706***	-0.956***	-5.825	12.240***	2.560	8.234***	-1.579
	(0.003)	(0.488)	(0.006)	(4.705)	(0.497)	(2.068)	(0.857)	(2.548)
year-quarter FE		Y		Y		Y		Y
industry FE (4-digit NAICS)		Y		Y		Y		Y
state FE		Y		Y		Y		Y
Log Likelihood					-667,000	-554,500	-228,400	-193,800
N					1,870,000	198,000	1,630,000	178,000
$\hat{\rho}$		-0.352***		-0.149***				
$\hat{\sigma}$		9.667***		9.282***				
$\hat{\lambda} = \hat{\sigma}\hat{\rho}$		-3.398***		-1.383***				

Notes: Heckman MLE estimation results. Panel A is based on a Probit model for $\Pr(m_{it} > 0)$. Panel B estimates an OLS regression for $E[m_{it} \mid m_{it} > 0]$. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. N and log likelihood values are rounded for disclosure avoidance.

lation between unobservables in the selection and outcome equations ($\hat{\rho}$) is negative and statistically significant, as is the implied selection term ($\hat{\lambda}$). These results suggest that unobserved characteristics that make applicants more likely to enter tend to lower the conditional realized employment.

Several mechanisms could underlie a negative selection. Applicants with lower opportunity costs or limited outside options may enter more readily but scale at smaller levels, while those with more ambitious plans face greater barriers to entry, consistent with the distinction between subsistence and transformational entrepreneurs in Schoar (2010). These dynamics reinforce the importance of accounting for ex-ante heterogeneity when analyzing early-stage employment outcomes. We also note that coefficient patterns for covariates largely align with the two-part model results for LLCs (Tables A.3 and A.4) that do not explicitly account for selection based on unobservables.

Though informative, the Heckman specification rests on key assumptions. The exclusion restriction must be uncorrelated with unobserved drivers of post-entry outcomes, which may not hold if *ltime* proxies for broader planning or regulatory quality. Moreover, the model's reliance on joint normality and functional form assumptions can influence the magnitude of selection effects. Nonetheless, the findings are generally consistent with the two-part model results: expectations are informative but imperfect, systematically overstating realized scale.

4.7 Expectation errors and near-term outcomes

Table 10 presents OLS regression results examining how initial expectation errors, δ_{it}, relate to two near-term outcomes measured three years (twelve quarters) after application: (i) an indicator for whether the business has positive employment, and (ii) the corresponding employment level. Applicants that never materialized as employer businesses or that entered and exited before the end of the third year are assigned zero employment. The analysis includes all applications submitted between 2017 and 2019 for which three-year outcomes are observable.

All regressions include the realized initial employment m_{it}, ensuring that comparisons are made holding actual initial size fixed. Consistent with canonical models of firm dynamics (Jovanovic, 1982; Hopenhayn, 1992), realized initial size strongly predicts both 3-year employer status and employment. However, forecast errors remain predictive even after conditioning on realized size.

In the full sample of applicants, greater initial overestimation is positively associated with both employer status and employment three years later. Among applicants who initially reported positive

Table 10: Expectation errors and near-term (3-year) outcomes

	All Applications		Positive Exp. ($\tilde{\mu}_{it} > 0$)	
Variable	3-yr. pos. emp.	3-yr. emp.	3-yr. pos. emp.	3-yr. emp.
δ_{it}	0.001***	0.082***	-0.004***	0.085***
	(0.000)	(0.005)	(0.000)	(0.007)
m_{it}	0.025***	0.676***	0.017***	0.660***
	(0.000)	(0.007)	(0.000)	(0.009)
partnership	0.016***	0.104***	0.010**	0.537**
	(0.000)	(0.018)	(0.003)	(0.170)
corporation	0.119***	0.268***	0.145***	0.688***
	(0.000)	(0.009)	(0.002)	(0.028)
llc	0.049***	0.188***	0.080***	0.659***
	(0.000)	(0.004)	(0.002)	(0.023)
other	0.041***	0.026	0.039***	0.191
	(0.001)	(0.029)	(0.004)	(0.191)
prior EIN	0.011	-0.068	-0.032	-0.133
	(0.008)	(0.068)	(0.017)	(0.193)
college share	0.022***	0.319***	0.072***	0.939***
	(0.001)	(0.024)	(0.004)	(0.127)
nonwhite share	-0.034***	-0.111***	-0.099***	-0.510***
	(0.000)	(0.010)	(0.003)	(0.006)
ln(median age)	-0.001	-0.018	-0.002	-0.032
	(0.001)	(0.013)	(0.003)	(0.064)
ln(firm density)	0.014***	0.0006***	0.022***	-0.065***
	(0.000)	(0.006)	(0.001)	(0.031)
dhs(avg. firm emp.)	-0.003***	0.064***	-0.004*	0.274***
	(0.000)	(0.007)	(0.001)	(0.038)
top4 emp. share	0.026***	0.013	0.059***	-0.035
	(0.001)	(0.018)	(0.004)	(0.099)
cv(firm emp.)	-0.001***	-0.008***	-0.003***	-0.022***
	(0.000)	(0.001)	(0.000)	(0.004)
cv(labor prod.)	-0.0007*	0.006	-0.001	0.007
	(0.000)	(0.006)	(0.001)	(0.030)
local gdp growth	0.002	0.052	-0.023	-0.018
	(0.006)	(0.116)	(0.026)	(0.513)
const.	0.110***	0.061	0.286***	-0.496
	(0.004)	(0.062)	(0.015)	(0.306)
year-quarter FE	Y	Y	Y	Y
industry FE (4-digit NAICS)	Y	Y	Y	Y
state FE	Y	Y	Y	Y
R^2	0.167	0.173	0.171	0.189
N	4,338,000	4,338,000	604,000	604,000

Notes: Each column shows a regression of the corresponding outcome on initial expectations and control variables. Columns (1) and (3) pertain to linear probability models (LPM) of 3-year employer status indicator (i.e. indicator of positive employment), and columns (2) and (4) are based on OLS models of 3-year employment level (including zero employment). Omitted category for legal form of organization is sole proprietorship. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. N values are rounded for disclosure avoidance.

employment expectations, overestimation is negatively associated with the employer status by the end of the third year but positively associated with the employment levels.

From the standpoint of theory, these findings go beyond the predictions of canonical rational-expectations frameworks. In the Jovanovic (1982) model, which abstracts from selection at the entry margin, all entrants share a common prior and posterior beliefs are determined solely by realized productivity. Once early outcomes are controlled, forecast errors should be irrelevant for subsequent performance—a similar implication holds in Hopenhayn (1992).[30] However, when entrants hold heterogeneous priors that contain additional information, forecast errors may retain incremental predictive content even after conditioning on initial outcome.

A broader class of rational-expectations models allows *ex-ante* heterogeneity and selection at the entry margin. In occupational-choice models (e.g., Lucas, 1978; Kihlstrom and Laffont, 1979; Jovanovic, 1994), differences in ability, risk preferences, or managerial skill determine who becomes an entrepreneur and how large firms are in equilibrium. Such frameworks can generate a positive association between pre-entry characteristics and subsequent outcomes, consistent with part of the evidence. But they do not feature subjective priors or explicit forecasts, and thus cannot explain systematically optimistic expectations or forecast errors that remain predictive once early outcomes are controlled for.

Finally, behavioral perspectives offer complementary explanations. Evidence on entrepreneurial optimism and overconfidence (e.g., Koellinger et al., 2007; Astebro et al., 2014) highlights how subjective beliefs may deviate systematically from rational posteriors. Such deviations can shape outcomes. For instance, overoptimism may reduce responsiveness to early negative signals or undermine credibility with financiers, employees, or customers, lowering the likelihood of employer status. At the same time, ambitious expectations may spur resource mobilization and result in higher employment among those businesses that survive.

Given the need for additional elements for canonical models to explain the observed patterns, the next section offers one parsimonious model that can account for the key empirical findings.

[30] In Hopenhayn (1992), entrants have common priors (equal to the mean of the initial productivity distribution of entrants) and know their initial productivity draw upon entry.

5 Heterogeneous priors, learning, and selection before entry

The model builds on the classic entrepreneurial learning framework of Jovanovic (1982), with each additional element introduced only as needed to match the observed patterns in the data. The goal is to explain four core findings: (i) business applicants systematically overestimate first-year employment; (ii) conditional on both expected and realized entry, applicants underestimate first-year employment; (iii) in the Mincer–Zarnowitz regression (8), the coefficient on expected first-year employment lies strictly between 0 and 1; and (iv) the first-year expectation error is positively correlated with subsequent employment, even after conditioning on realized first-year employment.

5.1 Model setup

As in Jovanovic (1982), each potential entrant i has a latent business idea quality $\theta_i \sim \mathcal{N}(\bar{\theta}, \sigma_\theta^2)$. Quality maps to realized (first-year) employment with some noise $m_{i1} = \theta_i + \varepsilon_{i1}, \varepsilon_{i1} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. *Before* deciding to apply for business, the agent observes an informative signal $s_{i0} = \theta_i + \varepsilon_{i0},\ \varepsilon_{i0} \sim \mathcal{N}(0, \sigma_0^2)$, and holds a heterogeneous prior (mean shift) ϕ_i, independent of $(\theta_i, \varepsilon_{i0}, \varepsilon_{i1})$, with $\phi_i \sim \mathcal{N}(0, \sigma_\phi^2)$. The shift ϕ_i can be interpreted as idiosyncratic optimism/pessimism or subjective evaluation of common information. Equivalently, the prior can be represented as $\theta_i \sim \mathcal{N}(\bar{\theta} + \phi_i, \sigma_\theta^2)$.

Given the prior mean, $(\bar{\theta} + \phi_i)$, and the observed s_{i0}, the posterior mean (the agent's forecast of θ_i and thus of expected m_{i1}) is

$$\tilde{\mu}_{i1} = \mathbb{E}[\theta_i \mid s_{i0}, \phi_i] = (1 - \kappa_0)(\bar{\theta} + \phi_i) + \kappa_0 s_{i0}, \qquad \kappa_0 = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_0^2}.$$

The cross-sectional variance of forecasts is $\sigma_{\tilde{\mu}_1}^2 = \mathrm{Var}(\tilde{\mu}_{i1}) = \kappa_0 \sigma_\theta^2 + (1 - \kappa_0)^2 \sigma_\phi^2$.

Suppose that there is a one-time cost of preparing and filing an application, and similarly, of actual entry (operational launch).[31] Under a monotone mapping from employment to net value, these costs induce thresholds c_A and c_E such that a potential entrant applies if $\tilde{\mu}_{i1} > c_A$ and enters if $m_{i1} > c_E$. We write $\tilde{a}_{i1} = \mathbf{1}\{\tilde{\mu}_{i1} > c_A\}$ and $e_{i1} = \mathbf{1}\{m_{i1} > c_E\}$. For $t \geq 2$, realized employment follows $m_{it} = \theta_i + \varepsilon_{it}$, with $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ i.i.d. across t and independent of all else.

[31] While the application itself may have a nominal cost, costs in the application stage broadly include those associated with planning and preparation for the business, such as estimating demand, seeking and securing financing, understanding relevant regulations, socializing the business idea and obtaining advice, searching for potential employees and suppliers.

5.2 Main implications

The model yields the following implications, with detailed proofs provided in Appendix C.

(i) Overall overestimation: Because business applications involve fixed costs, only agents whose expected first-year employment exceeds a threshold choose to apply. Suppressing the subscript i for notational simplicity, the expected forecast error among applicants is

$$\mathbb{E}\big[\tilde{\mu}_1 - m_1 \,\big|\, \tilde{\mu}_1 > c_A\big] = \frac{(1-\kappa_0)^2\,\sigma_\phi^2}{\sigma_{\tilde{\mu}_1}}\,\lambda\!\left(\frac{c_A - \bar{\theta}}{\sigma_{\tilde{\mu}_1}}\right) \;>\; 0,$$

where $\lambda(\cdot) = \varphi(\cdot)/\big(1-\Phi(\cdot)\big)$ is the inverse Mills ratio of the standard normal distribution. In other words, selecting on high $\tilde{\mu}_1$ picks individuals whose high forecasts partly come from ϕ which does not persist into m_1, so forecasts sit above realizations on average.

(ii) Underestimation conditional on expected *and* actual entry: Among actual entrants, average forecast error is negative since truncating on realized outcomes shifts the conditional mean of m_1 upward,

$$\mathbb{E}[\tilde{\mu}_1 - m_1 \mid m_1 > c_E] = -\,\frac{(1-\kappa_0)\sigma_\theta^2 + \sigma_\varepsilon^2}{\sqrt{\sigma_\theta^2+\sigma_\varepsilon^2}}\,\lambda\!\left(\frac{c_E - \bar{\theta}}{\sqrt{\sigma_\theta^2+\sigma_\varepsilon^2}}\right) \;<\; 0.$$

Conditioning on $\{\tilde{\mu}_1 > c_A,\ m_1 > c_E\}$ preserves strict negativity if the overall overestimation is not too large: $\mathbb{E}[\tilde{\mu}_1 - m_1 \mid \tilde{\mu}_1 > c_A,\ m_1 > c_E] < 0$.[32]

For applicants who expected to enter and did so, success corresponds to a positive period-1 shock that pushes realizations above forecasts. If this selection effect outweighs aggregate overestimation, the selected sample displays underestimation on average.

(iii) Mincer-Zarnowitz slope $\in (0,1)$: In the regression of m_1 on $\tilde{\mu}_1$, the population slope is

$$\beta_{\text{MZ}} = \frac{\kappa_0\,\sigma_\theta^2}{\kappa_0\sigma_\theta^2 + (1-\kappa_0)^2\sigma_\phi^2} \;\in\; (0,1),$$

provided $\sigma_0^2, \sigma_\phi^2 \in (0,+\infty)$. Forecasts are predictive for outcomes but not one-for-one on the margin. Only the informative part co-moves with realized employment and enters the covariance (numerator), while both the informative and heterogeneous-prior components inflate the variance of forecasts (denominator). Note that $\beta_{\text{MZ}} = 0$ when $\sigma_0^2 = +\infty$ and $\beta_{\text{MZ}} = 1$ when $\sigma_\phi^2 = 0$: in the absence of any informative signal prior to application, the coefficient equals zero; in the absence of uninformative heterogeneous-prior, it equals one. β_{MZ} does not change when conditional on

[32] Appendix C provides an explicit expression of the condition.

the set of applicants $A = \{\tilde{\mu}_1 > c_A\}$, or the set of applicants with positive expected employment $B = \{\tilde{\mu}_1 > c_E\}$, since the numerator and the denominator get scaled by the same truncation effect. This result also holds in the Heckman model setting.[33]

(iv) Positive partial correlation between expectation error and subsequent employment: Consider the regression

$$m_{it} = \beta_0 + \beta_1 m_{i1} + \beta_2 \delta_{i1} + u_{it}, \ t \geq 2. \tag{12}$$

where u_{it} is an error term. The coefficient on the expectation error (δ_{i1}) is given by

$$\beta_2 = \frac{\kappa_0 \, \sigma_\theta^2 \, \sigma_\varepsilon^2}{\left(\sigma_\theta^2 + \sigma_\varepsilon^2\right) \mathrm{Var}(\tilde{\mu}_1) \ - \ \left(\kappa_0 \, \sigma_\theta^2\right)^2} \ > \ 0.$$

Holding the first-year employment, m_{i1}, fixed, δ_{i1} isolates incremental information about θ coming from the applicant's signal before application. If no informative signal is available prior to application (i.e., $\sigma_0 = \infty$), then $\beta_2 = 0$. This conclusion still holds when conditioning on the set of applicants $A = \{i : \tilde{\mu}_{i1} > c_A\}$, or on the subset with positive expected employment $B = \{i : \tilde{\mu}_{i1} > c_E\}$.

Table 11: Implications of the model's elements

	Model Element			
	Baseline	Informative signal (s_0)	Heterogeneous prior (ϕ)	Both (s_0, ϕ)
Overall overestimation	N	N	Y	Y
Conditional underestimation	Y	Y	Y	Y
Mincer-Zarnowitz slope $\in (0, 1)$	N (not identified)	N ($= 1$)	N ($= 0$)	Y
$\beta_2 > 0$ in regression (12)	N (not identified)	Y	N	Y

Table 11 summarizes how each additional model ingredient (relative to the baseline Jovanovic model) contributes to the results. The baseline Jovanovic framework does not generate overestimation, nor does it identifies the Mincer–Zarnowitz slope or the coefficient on the forecast-error in the near-term employment regression. Incorporating the *informative signal* links forecasts to the true business quality, delivering $\beta_2 > 0$. Introducing *heterogeneous priors* accounts for overall

[33] In the second-stage regression of the Heckman Model, the conditional expectation of m_{i1} in the selected sample $e_{i1} = 1$ is

$$E[m_{i1} \mid \tilde{\mu}_{i1}, e_{i1} = 1] = \alpha + \beta_{MZ} \tilde{\mu}_{i1} + \rho \sigma_\epsilon \lambda(r_i \eta),$$

where $\lambda(\cdot)$ is the inverse Mills ratio and ρ is the correlation between ε_i (the error term in employment) and u_i (the error term in firms' net present value, and thus in the entry decision). β_{MZ} has the same form as in the OLS regression and remains strictly between 0 and 1 when both the informative and heterogeneous-prior components are present.

overestimation. Combining *both* elements produces a Mincer-Zarnowitz slope in (0,1). We note that this model is not the only one consistent with the core findings, and further generalizations of the model are possible.[34]

5.3 Entry distortions from heterogeneous priors

We use *entry distortions* to mean mis-sorting at the application (or expected employer-entry) margin: decisions deviate from the ones implied by underlying business quality because expectations contain a noisy, uninformative idiosyncratic component (ϕ). In other words, business quality is no longer the sole driver of application and expected-entry decisions.

Define the *excess* application/expected-entry rate as the probability that an agent would not apply/expect entry absent noisy priors ($\phi = 0$) but does so when $\phi \neq 0$:

$$p_j^{\text{exc}} = \Pr\big(\tilde{\mu}_1^0 \leq c_j,\ \tilde{\mu}_1^0 + (1-\kappa_0)\phi > c_j\big) = \int_{-\infty}^{c_j} \Phi\left(\frac{u - c_j}{(1-\kappa_0)\sigma_\phi}\right) f_{\tilde{\mu}_1^0}(u)\, du, \tag{13}$$

where $\tilde{\mu}_1^0 \sim \mathcal{N}(\bar{\theta},\ \kappa_0\sigma_\theta^2)$ is expected first-year employment absent noisy priors, and $f_{\tilde{\mu}_1^0}$ denotes its density. The subscript $j \in \{A, E\}$ denotes application (A) or expected entry (E), respectively.

Similarly, define *forgone* application/expected entry as the probability that an agent would apply/expect entry absent noisy priors ($\phi = 0$) but does not when $\phi \neq 0$:

$$p_j^{\text{for}} = \Pr\big(\tilde{\mu}_1^0 > c_j,\ \tilde{\mu}_1^0 + (1-\kappa_0)\phi \leq c_j\big) = \int_{c_j}^{+\infty} \Phi\left(\frac{c_j - u}{(1-\kappa_0)\sigma_\phi}\right) f_{\tilde{\mu}_1^0}(u)\, du, \tag{14}$$

The resulting *ϕ-driven entry distortion* is the sum of these two components:

$$p_j^{\text{dis}} = p_j^{\text{exc}} + p_j^{\text{for}}.$$

The following result characterizes how entry distortions respond to changes in prior dispersion (see Appendix C for a proof).

Proposition 1. *Both p_j^{exc} and p_j^{for} are strictly increasing in the dispersion of noisy priors, σ_ϕ. Hence, p_j^{dis} is also strictly increasing in σ_ϕ.*

[34] For instance, one can allow ϕ to be generally correlated with θ: $\text{Cov}(\theta, \phi) = c \geq 0$. Write $\phi = (c/\sigma_\theta^2)\theta + u, u \perp \theta, \text{Var}(u) = \sigma_\phi^2 - (c^2/\sigma_\theta^2)$. When $c = 0$ (the case studied here), ϕ is purely idiosyncratic; when $c > 0$, it embeds an informative component proportional to θ. Agents also observe $s_0 = \theta + \varepsilon_0$, $\varepsilon_0 \sim \mathcal{N}(0, \sigma_0^2)$, and form $\tilde{\mu}_1 = E[\theta \mid s_0, \phi] = \bar{\theta} + a(s_0 - \bar{\theta}) + b\phi$, where $a = (\sigma_\theta^2\sigma_\phi^2 - c^2)/D, b = (c\sigma_0^2)/D$, and $D = (\sigma_\theta^2 + \sigma_0^2)\sigma_\phi^2 - c^2$. Thus ϕ combines a rational component correlated with θ and a residual noise term u. Intuitively, s_0 represents structured, observable information (e.g., prior experience), while ϕ captures heterogeneous priors that may include tacit or network-based knowledge (when $c > 0$) with idiosyncratic behavioral elements, such as optimism or pessimism (through u).

Interpreted through the lens of the model, the decline in β_{MZ} during the period 2017-2021 (Table 8) is consistent with greater dispersion in priors relative to underlying quality and/or a noisier pre-application signal.[35] If σ_θ and σ_0 are stable over time, a decline in β_{MZ} then corresponds to a higher σ_ϕ and therefore larger entry distortions.

Entry distortions can lead to misallocation in the aggregate economy. On one margin, noisy signals can attract financing or policy support to ventures that ultimately do not become employers or operate at small scale. On the other margin, conservative or tightly calibrated forecasts can be underweighted, limiting resources for projects that, conditional on entry, would scale more. This misallocation can shift entrant cohorts away from higher-growth firms and slow the reallocation of labor and capital toward more productive ones. Even if learning and selection eventually may correct these errors, the transition may feature excess experimentation and delayed expansion among eventual winners.

6 Conclusion

This paper provides new evidence on entrepreneurial expectations around the time of business formation by linking business application-stage forecasts of employer entry and first-year employment to subsequent realizations in large-scale administrative microdata. This linkage makes it possible to quantify expectation errors at both the extensive margin (entry) and the intensive margin (conditional initial scale), and to characterize how these errors vary across organizational forms, local conditions, and sectors.

Expectation errors near the entry margin are statistically and economically meaningful. Errors are also highly heterogeneous, even among applicants who both expect and achieve entry. Expectations are biased and inefficient overall. At the same time, they are informative: expected employment is positively correlated with both the probability of entry and realized employment, and forecast errors predict near-term employment outcomes. Taken together, these patterns suggest a role for heterogeneous priors, learning, and selection at the entry margin—bridging canonical firm-dynamics frameworks with theories of entrepreneurial choice and behavioral approaches.

Looking ahead, future work should identify the mechanisms of expectation formation and evalu-

[35] In the model, β_{MZ} decreases as either $\sigma_\phi/\sigma_\theta$ or σ_0/σ_θ increases.

ate how different types of misalignment affect business performance and where policy interventions could be effective. Expectations can be linked to financing and innovation strategies, hiring practices, and the regulatory environment, providing a richer account of how early frictions shape entrepreneurial trajectories. It is also important to understand how different types of uncertainty (e.g., policy, demand, or costs) affect expectation formation. Finally, the timeliness and high frequency of the BFS data enable studying expectations as early indicators of startup scale and job creation, motivating closer examination of their time-series relationship with realized outcomes.

References

Asplund, M. and V. Nocke (2006). Firm turnover in imperfectly competitive markets. *Review of Economic Studies 73*(2), 295–327.

Astebro, T., H. Herz, R. Nanda, and R. A. Weber (2014). Seeking the roots of entrepreneurship: Insights from behavioral economics. *Journal of Economic Perspectives 28*(3), 49–70.

Astebro, T., S. Jeffrey, and G. K. Adomdza (2007). Inventor perseverance after being told to quit: The role of cognitive biases. *Journal of Behavioral Decision Making 20*(3), 253–272.

Bachmann, R. and C. Bayer (2014). Investment dispersion and the business cycle. *American Economic Review 104*(4), 1392–1416.

Bachmann, R., S. Elstner, and E. R. Sims (2013). Uncertainty and economic activity: Evidence from business survey data. *American Economic Journal: Macroeconomics 5*(2), 217–249.

Bachmann, R., G. Topa, and W. van der Klaauw (2022). *Handbook of Economic Expectations*. Elsevier.

Baron, R. A. (2004). The cognitive perspective: A valuable tool for answering entrepreneurship's basic "why" questions. *Journal of Business Venturing 19*(2), 221–239.

Bayard, K., E. Dinlersoz, T. Dunne, J. Haltiwanger, J. Miranda, and D. Stevens (2018). Early-stage business formation: An analysis of applications for employer identification numbers. Technical Report CES 18-39, U.S. Census Bureau Center for Economic Studies.

Belotti, F., P. Deb, W. G. Manning, and E. C. Norton (2015). Twopm: Two-part models. *The Stata Journal 15*(1), 3–20.

Bloom, N. (2007). Uncertainty and investment dynamics. *Review of Economic Studies 74*(2), 391–415.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica 77*(3), 623–685.

Camerer, C. and D. Lovallo (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review 89*(1), 306–318.

Campbell, J. R. and H. A. Hopenhayn (2005). Market size matters. *The Journal of Industrial Economics 53*(1), 1–25.

Carlino, G. A. and W. R. Kerr (2009). Localized knowledge spillovers: Evidence from the patent data. *Journal of Urban Economics 65*(1), 14–29.

Coibion, O., Y. Gorodnichenko, and S. Kumar (2018). How do firms form their expectations? new survey evidence. *American Economic Review 108*(9), 2671–2713.

Datta, S., M. Iskandar-Datta, and V. Sharma (2011). Product market pricing power, industry concentration and analysts' earnings forecasts. *Journal of Banking & Finance 35*(6), 1352–1366.

Davis, S., J. Haltiwanger, C. Krizan, R. Jarmin, J. Miranda, A. Nucci, and K. Sandusky (2009). Measuring the dynamics of young and small businesses: Integrating the employer and non-employer businesses. In *in Producer Dynamics: New Evidence from Micro Data*, pp. 329–366. University of Chicago Press.

Dingel, J. I. and B. Neiman (2020). How many jobs can be done at home? *Journal of Public Economics 189*, 104235.

Dinlersoz, E., T. Dunne, J. Haltiwanger, and V. Penciakova (2023). The local origins of business formation. CES Working Paper No. 23-34.

Dinlersoz, E., Z. Kroff, A. Luque, and V. Novik (2025). The modes of entry: Employer versus non-employer business formation. Unpublished Manuscript (in progress).

Dinlersoz, E. M. (2004). Firm organization and the structure of retail markets. *Journal of Economics & Management Strategy 13*(2), 207–240.

Ericson, R. and A. Pakes (1995). Markov-perfect industry dynamics: A framework for empirical work. *Review of Economic Studies 62*(1), 53–82.

Fairlie, R. W. et al. (2012). Immigrant entrepreneurs and small business owners, and their access to financial capital. *Small Business Administration 396*, 1–46.

Fairlie, R. W., Z. Kroff, J. Miranda, and N. Zolas (2023). *The promise and peril of entrepreneurship: Job creation and survival among US startups*. MIT Press.

Gennaioli, N., R. La Porta, F. Lopez-de Silanes, and A. Shleifer (2013). Human capital and regional development. *Quarterly Journal of Economics 128*(1), 105–164.

Goldschlag, N. and J. Miranda (2020). Business dynamics statistics of high tech industries. *Journal of Economics & Management Strategy 29*(1), 3–30.

Guiso, L. and G. Parigi (1999). Investment and demand uncertainty. *The Quarterly Journal of Economics 114*(1), 185–227.

Guzman, J. and S. Stern (2020). The state of american entrepreneurship: New estimates of the quantity and quality of entrepreneurship for 32 us states, 1988–2014. *American Economic Journal: Economic Policy 12*(4), 212–243.

Hall, B. H. (2010). Innovation and productivity. *NBER Working Paper No. 17178*.

Hall, R. E. and S. E. Woodward (2010). The burden of the nondiversifiable risk of entrepreneurship. *American Economic Review 100*(3), 1163–1194.

Hecker, D. E. (2005). High-technology employment: a naics-based update. *Monthly Labor Review 128*, 57–72.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica 60*(5), 1127–1150.

Hurst, E. and B. W. Pugsley (2011). What do small businesses do? *Brookings Papers on Economic Activity 43*(2), 73–142.

Jovanovic, B. (1982). Selection and the evolution of industry. *Econometrica 50*(3), 649–670.

Jovanovic, B. (1994). Firm formation with heterogeneous management and labor skills. *Small Business Economics 6*(3), 185–191.

Kihlstrom, R. E. and J.-J. Laffont (1979). A general equilibrium entrepreneurial theory of firm formation based on risk aversion. *Journal of Political Economy 87*(4), 719–748.

Koellinger, P., M. Minniti, and C. Schade (2007). Entrepreneurial overconfidence and the performance of new ventures. *Journal of Economic Psychology 28*(4), 502–527.

Lafontaine, F. and K. Shaw (2016). Serial entrepreneurship: Learning by doing? *Journal of Labor Economics 34*(S2), S217–S254.

Lazear, E. P. (2005). Entrepreneurship. *Journal of Labor Economics 23*(4), 649–680.

Leduc, S. and Z. Liu (2016). Uncertainty shocks are aggregate demand shocks. *Journal of Monetary Economics 82*, 20–35.

Levine, R. and Y. Rubinstein (2017). Smart and illicit: who becomes an entrepreneur and do they earn more? *The Quarterly Journal of Economics 132*(2), 963–1018.

Lucas, R. E. (1978). On the size distribution of business firms. *The Bell Journal of Economics 9*(2), 508–523.

Manski, C. F. (2004). Measuring expectations. *Econometrica 72*(5), 1329–1376.

Manso, G. (2016). Experimentation and the returns to entrepreneurship. *The Review of Financial Studies 29*(9), 2319–2340.

Mincer, J. A. and V. Zarnowitz (1969). The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pp. 3–46. NBER.

Moretti, E. (2004a). Human capital externalities in cities. In *Handbook of regional and urban economics*, Volume 4, pp. 2243–2291. Elsevier.

Moretti, E. (2004b). Workers' education, spillovers, and productivity: Evidence from plant-level production functions. *American Economic Review 94*(3), 656–690.

Moskowitz, T. J. and A. Vissing-Jørgensen (2002). The returns to entrepreneurial investment: A private equity premium puzzle? *American Economic Review 92*(4), 745–778.

Nocke, V. (2006). A gap for me: Entrepreneurs and entry. *Journal of the European Economic Association 5*(4), 929–956.

Papanikolaou, D. and L. D. Schmidt (2022). Working remotely and the supply-side impact of covid-19. *The Review of Asset Pricing Studies 12*(1), 53–111.

Parker, S. C. (2009). *The Economics of Entrepreneurship*. Cambridge University Press.

Puri, M. and D. T. Robinson (2007). Optimism and economic choice. *Journal of Financial Economics 86*(1), 71–99.

Schoar, A. (2010). The divide between subsistence and transformational entrepreneurship. *Innovation policy and the economy 10*(1), 57–81.

Sterk, V., P. Sedláček, and B. Pugsley (2021). The nature of firm growth. *American Economic Review 111*(2), 547–579.

U.S. Small Business Administration (2025). Choose a business structure. https://www.sba.gov/business-guide/launch-your-business/choose-business-structure. Accessed 31 July 2025.

Wu, B. and A. M. Knott (2006). Entrepreneurial risk and market entry. *Management Science 52*(9), 1315–1330.

Appendix A Additional tables and figures

Table A.1: Summary statistics for expectation errors by sector

Industry	A. All applications			B. Positive Exp. ($\tilde{\mu}_{it} > 0$)		
	$\tilde{\mu}_{it}$	m_{it}	δ_{it}	$\tilde{\mu}_{it}$	m_{it}	δ_{it}
Accommodation & Food Services	2.142	2.120	0.022	6.994	4.689	2.305
	(0.007)	(0.008)	(0.008)	(0.019)	(0.020)	(0.023)
Administrative & Support & Waste Management	0.609	0.358	0.252	4.352	1.436	2.916
	(0.003)	(0.003)	(0.004)	(0.021)	(0.015)	(0.022)
Arts, Entertainment, & Recreation	0.644	0.442	0.202	5.070	1.958	3.112
	(0.005)	(0.005)	(0.006)	(0.036)	(0.028)	(0.038)
Construction	0.445	0.337	0.108	3.293	1.321	1.972
	(0.002)	(0.002)	(0.002)	(0.011)	(0.009)	(0.012)
Educational Services	0.539	0.352	0.188	4.308	1.591	2.717
	(0.006)	(0.005)	(0.006)	(0.038)	(0.028)	(0.039)
Finance & Insurance	0.286	0.165	0.122	3.903	1.201	2.701
	(0.003)	(0.001)	(0.003)	(0.039)	(0.013)	(0.040)
Health Care & Social Assistance	0.986	0.664	0.322	4.977	1.983	2.995
	(0.005)	(0.004)	(0.005)	(0.021)	(0.015)	(0.021)
Information	0.563	0.248	0.315	4.776	1.055	3.721
	(0.006)	(0.004)	(0.006)	(0.044)	(0.022)	(0.046)
Manufacturing	0.965	0.632	0.333	6.251	2.467	3.785
	(0.009)	(0.007)	(0.010)	(0.052)	(0.036)	(0.052)
Mining, Quarrying, & Oil & Gas Extraction	0.968	0.858	0.111	5.717	2.906	2.811
	(0.028)	(0.030)	(0.034)	(0.149)	(0.128)	(0.161)
Other Services (except Public Administration)	0.399	0.269	0.131	3.442	1.305	2.136
	(0.002)	(0.002)	(0.002)	(0.013)	(0.010)	(0.014)
Professional, Scientific, & Technical Services	0.356	0.273	0.083	3.277	1.358	1.919
	(0.002)	(0.001)	(0.002)	(0.013)	(0.010)	(0.014)
Real Estate & Rental & Leasing	0.174	0.138	0.036	2.531	0.858	1.672
	(0.001)	(0.001)	(0.002)	(0.018)	(0.012)	(0.019)
Retail Trade	0.325	0.218	0.107	3.581	1.466	2.115
	(0.001)	(0.001)	(0.001)	(0.012)	(0.010)	(0.013)
Transportation & Warehousing	0.386	0.257	0.129	3.272	0.975	2.298
	(0.002)	(0.003)	(0.003)	(0.016)	(0.014)	(0.018)
Utilities	0.450	0.279	0.172	5.801	1.814	3.988
	(0.027)	(0.020)	(0.029)	(0.314)	(0.179)	(0.299)
Wholesale Trade	0.431	0.286	0.145	3.673	1.223	2.451
	(0.004)	(0.003)	(0.004)	(0.030)	(0.020)	(0.030)

Notes: Each row corresponds to a 2-digit NAICS industry – excludes Agriculture and Management of Companies & Enterprises. "All" refers to all applications, and "Positive Exp." refers to applications that report positive expected maximum employment. Standard errors are in parentheses. $\delta_{it} = \tilde{\mu}_{it} - m_{it}$.

Table A.2: The definition and sources of covariates

Covariate	**Definition / Construction**	**Data Source**
legal form: sole proprietorship, partnership, corporation, LLC, other	Dummy variables indicating legal form of organization reported on IRS Form SS-4 at application. Sole proprietorship is the omitted category.	BFS microdata (IRS Form SS-4)
prior EIN	Indicator for whether the applicant has previously obtained an EIN.	BFS microdata (IRS Form SS-4)
college share	Share of population aged 25+ with a bachelor's degree or higher in the applicant's census tract, lagged two years.	ACS 5-year estimates
nonwhite share	Share of population in the applicant's census tract identifying as nonwhite, lagged two years.	ACS 5-year estimates
log(median age)	Natural log of median age in the applicant's census tract, lagged two years.	ACS 5-year estimates
log(firm density)	Natural log of number of firms per 1,000 county residents in the applicant's industry–county cell, lagged one year.	LBD
dhs(avg. firm emp.)	Davis–Haltiwanger–Schuh (DHS) transformation of average employment in the applicant's industry–county cell, lagged one year: $$DHS(\overline{\text{Emp}})_{j,c,t-1} = \frac{\overline{\text{Emp}}_{j,c,t-1} - \overline{\text{Emp}}}{0.5(\overline{\text{Emp}}_{j,c,t-1} + \overline{\text{Emp}})}$$	LBD
top4 emp. share	Employment share of the largest four firms in the applicant's industry–county cell, lagged one year.	LBD
cv(firm emp.)	Coefficient of variation (std. dev. / mean) of firm employment in the applicant's industry–county cell, lagged one year.	LBD
cv(labor prod.)	Coefficient of variation of firm labor productivity (revenue per employee) in the applicant's industry–county cell, lagged one year.	LBD
local gdp growth	County-level GDP growth in the quarter prior to application.	BEA
local covid shock	$(1 - \text{WFHshare}_{j(i)}) \times \frac{\text{NewCases}_{c(i),t}}{\text{Pop}_{c(i),2019}}$, where $\text{WFHshare}_{j(i)}$ is the industry share of jobs teleworkable (Dingel and Neiman, 2020); $\text{NewCases}_{c(i),t}$ is new COVID-19 cases in county c and quarter t; $\text{Pop}_{c(i),2019}$ is pre-pandemic population.	WFHshare: Dingel and Neiman (2020); COVID cases: New York Times; Population: Census; Industry–county mapping from BFS
ltime	Natural log of the number of local procedural days required to legally establish a domestic limited liability company (LLC).	Arizona State University's Doing Business in North America (DBNA) database

Notes: The first column lists the covariates used in the OLS regression in equation (10), the two-part model and the Heckman model in equation (11); the second column describes the construction of each variable, and the third column reports the corresponding data source.

Table A.3: Two-part model estimation – sample of LLC applications

	A. Probit	B. GLM	
		Gaussian	Gamma
Variable	(1)	(2)	(3)
$\tilde{\mu}_{it}$	0.032***	0.631***	0.494***
	(0.001)	(0.014)	(0.008)
prior EIN	0.378***	0.127	1.090
	(0.088)	(1.045)	(1.170)
college share	0.266***	3.477***	1.495***
	(0.013)	(0.268)	(0.163)
nonwhite share	-0.500***	-0.166	0.379**
	(0.010)	(0.191)	(0.131)
ln(median age)	-0.067***	-0.117	-0.017
	(0.012)	(0.226)	(0.138)
ln(firm density)	0.118***	-0.685***	-0.411***
	(0.007)	(0.137)	(0.090)
dhs(avg. firm emp.)	0.009	0.083	-0.079
	(0.007)	(0.137)	(0.087)
cv(firm emp.)	-0.009***	0.069***	0.047***
	(0.001)	(0.019)	(0.014)
top4 emp. share	0.248***	-2.330***	-1.068**
	(0.032)	(0.597)	(0.379)
cv(labor prod.)	-0.007	0.135	0.039
	(0.006)	(0.123)	(0.071)
local gdp growth	0.006	0.641	-0.917*
	(0.026)	(0.520)	(0.397)
local covid shock	-0.024***	-0.040	-0.070
	(0.004)	(0.065)	(0.038)
const.	-1.725***	-4.902***	-2.409**
	(0.476)	(1.331)	(0.841)
year-quarter FE	Y	Y	Y
industry FE (4-digit NAICS)	Y	Y	Y
state FE	Y	Y	Y
Log Likelihood		-554,600	-466,400
N		1,630,000	1,630,000

Notes: Panel A provides estimates of the Probit model for $\Pr(m_{it} > 0)$. Panel B contains estimates of the GLM for $E[m_{it} \mid m_{it} > 0]$ using Gaussian and Gamma links. Log likelihood is the total log likelihood of Probit and GLM models. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. N and log likelihood values are rounded for disclosure avoidance.

Table A.4: Two-part model estimation – sample of LLC applications with positive expectations

	A. Probit	B. GLM	
		Gaussian	Gamma
Variable	(1)	(2)	(3)
$\tilde{\mu}_{it}$	0.007***	0.716***	0.857***
	(0.000)	(0.016)	(0.012)
prior EIN	-0.707***	-3.066	-1.040***
	(0.204)	(1.655)	(0.189)
college share	0.416***	2.440***	0.613***
	(0.027)	(0.408)	(0.183)
nonwhite share	-0.681***	-0.479	0.085
	(0.021)	(0.306)	(0.145)
ln(median age)	-0.033	0.064	0.266
	(0.023)	(0.345)	(0.156)
ln(firm density)	0.118***	-0.685***	-0.411***
	(0.007)	(0.137)	(0.090)
dhs(avg. firm emp.)	0.009	0.083	-0.079
	(0.007)	(0.138)	(0.088)
cv(firm emp.)	-0.009***	0.069***	0.047***
	(0.001)	(0.019)	(0.014)
top4 emp. share	0.248***	-2.330***	-1.068***
	(0.032)	(0.597)	(0.379)
cv(labor prod.)	-0.029**	0.190	0.060
	(0.012)	(0.170)	(0.087)
local gdp growth	0.006	0.641	-0.917
	(0.026)	(0.520)	(0.397)
local covid shock	-0.015*	-0.050	-0.063
	(0.008)	(0.119)	(0.057)
const.	0.194	-2.407	-0.539
	(0.188)	(2.545)	(1.332)
year-quarter FE	Y	Y	Y
industry FE (4-digit NAICS)	Y	Y	Y
state FE	Y	Y	Y
Log Likelihood		-193,800	-161,000
N		178,000	178,000

Notes: Panel A provides estimates of the Probit model for $\Pr(m_{it} > 0)$. Panel B contains estimates of the GLM for $E[m_{it} \mid m_{it} > 0]$ using Gaussian and Gamma links. Log likelihood is the total log likelihood of Probit and GLM models. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. N and log likelihood values are rounded for disclosure avoidance.

Appendix B Robustness to measurement frequency

To assess the sensitivity of the estimates of the average expectation error to perceived employment measurement frequency by the potential entrant, we estimate theoretical upper bounds for the expected value of the maximum of τ random variables where $\tau = 12, 24$, corresponding to monthly and approximately bi-weekly measurement frequencies for realized employment.

First, we use a two-part model to estimate the probability of entry, $P(E_{it} = 1)$ and the average realized employment (conditional on entry), $\hat{\mu}_{it}$, over the 4 quarter horizon as a function of observables. We also use an OLS regression to estimate the standard deviation of realized quarterly employment across quarters conditional on entry, again using application characteristics and fixed effects as predictors. We then obtain the application-level predicted average quarterly employment, $\hat{\mu}_{it}$, and the predicted standard deviation of quarterly employment, $\hat{\sigma}_{it}$.

We also estimate the average pairwise correlation between two consecutive quarterly employment levels conditional on entry, which yields $\hat{\rho}_e = 0.87$ – indicating that the correlation between two consecutive quarterly employment levels is quite high. As an alternative, we use a higher (assumed) value $\hat{\rho}_e = 0.95$ to capture the likely higher average correlation between two consecutive monthly or bi-weekly employment levels.

The estimated theoretical upper bound for the expected value of the maximum of τ Gaussian (not necessarily independent) random variables is then given by

$$\hat{U}(E[M_{it}|E_{it} = 1]) = \hat{\mu}_{it} + \hat{\sigma}_{it} \times \sqrt{2(1 - \hat{\rho}_e)\tau}. \tag{15}$$

A similar estimated upper bound can also be calculated for the case of Gamma random variables as

$$\hat{U}(E[M_{it}|E_{it} = 1]) = \frac{2(\hat{\sigma}_{it}/\hat{\mu}_{it})(1 - \hat{\rho}_e)ln(\tau)}{1 - \tau^{-\hat{\mu}^2_{it}/\hat{\sigma}_{it}}}. \tag{16}$$

We then construct, for each application, an estimate of the theoretical upper bound of the expected value of realized maximum employment (M_{it}) in the 4 quarters following application

$$\begin{aligned} \hat{U}(E[M_{it}]) &= \hat{P}(E_{it} = 1) \times \hat{U}(E[M_{it}|E_{it} = 1]) + (1 - \hat{P}(E_{it} = 1)) \times 0 \\ &= \hat{P}(E_{it} = 1) \times \hat{U}(E[M_{it}|E_{it} = 1]). \end{aligned} \tag{17}$$

Using (17), we construct estimates of the average expectation error $\hat{\bar{\delta}}_\tau$ for $\tau = 12, 24$. For comparison, we also calculate the average expectation errors based on *i.i.d.* assumption ($\hat{\rho}_e = 0$) for quarterly employments, $\hat{\bar{\delta}}^{iid}_\tau$ for each case $\tau = 12, 24$. We also report the average predicted expectation error based on the estimated maximum employment using the two-part model, $\hat{\bar{\delta}}^p$, for comparison with the average (raw) expectation error.

The results in Tables B.1 and B.2 indicate that estimated average expectation error using the theoretical upper bounds are significantly greater than zero statistically for both cases, $\rho_e = 0.87, 0.95$ and only becomes negative when the extreme assumption of *i.i.d* quarterly employments is imposed. These results support the finding of overestimation with quarterly measurement, given the fact that these estimated upper bounds are conservative (i.e., larger than the actual theoretical expected value of the maximum employment).

Table B.1: Expectation errors based on theoretical upper bounds for maximum employment — All applications

	A. Gaussian		B. Gamma	
Statistic	$\hat{\rho}_e = 0.87$	$\hat{\rho}_e = 0.95$	$\hat{\rho}_e = 0.87$	$\hat{\rho}_e = 0.95$
$\hat{\bar{\delta}}$	0.144	0.144	0.144	0.144
	(0.0009)	(0.0009)	(0.0009)	(0.0009)
$\hat{\bar{\delta}}_p$	0.144	0.144	0.144	0.144
	(0.0006)	(0.0006)	(0.0005)	(0.0005)
$\hat{\bar{\delta}}_{12}$	0.045	0.119	0.025	0.160
	(0.0006)	(0.0007)	(0.0006)	(0.0007)
$\hat{\bar{\delta}}_{24}$	0.019	0.103	0.022	0.142
	(0.0006)	(0.0006)	(0.0006)	(0.0007)
$\hat{\bar{\delta}}_{12}^{iid}$	-0.305	-0.305	-1.439	-1.439
	(0.0006)	(0.0006)	(0.0007)	(0.0007)
$\hat{\bar{\delta}}_{24}^{iid}$	-0.376	-0.376	-1.799	-1.799
	(0.0006)	(0.0006)	(0.0008)	(0.0008)

Notes: The table shows mean expectation errors and associated standard errors (in parentheses). Results are reported for two values of the quarterly employment correlation, $\hat{\rho}_e = 0.87$ and $\hat{\rho}_e = 0.95$, under both Gaussian and Gamma forecast error assumptions. The value $\hat{\rho}_e = 0.87$ corresponds to the average observed correlation between consecutive quarterly employment levels within a firm, while $\hat{\rho}_e = 0.95$ reflects an assumed correlation for monthly employment.

Table B.2: Expectation errors based on theoretical upper bounds for maximum employment — Applications with positive exp. ($\tilde{\mu}_{it} > 0$)

	A. Gaussian		B. Gamma	
Statistic	$\hat{\rho}_e = 0.87$	$\hat{\rho}_e = 0.95$	$\hat{\rho}_e = 0.87$	$\hat{\rho}_e = 0.95$
$\hat{\bar{\delta}}$	2.436	2.436	2.436	2.436
	(0.006)	(0.006)	(0.006)	(0.006)
$\hat{\bar{\delta}}_p$	2.437	2.437	2.433	2.433
	(0.004)	(0.004)	(0.004)	(0.004)
$\hat{\bar{\delta}}_{12}$	2.136	2.402	2.153	2.583
	(0.004)	(0.004)	(0.004)	(0.004)
$\hat{\bar{\delta}}_{24}$	2.044	2.345	2.047	2.542
	(0.004)	(0.004)	(0.004)	(0.004)
$\hat{\bar{\delta}}_{12}^{iid}$	0.891	0.891	-2.521	-2.521
	(0.004)	(0.004)	(0.004)	(0.004)
$\hat{\bar{\delta}}_{24}^{iid}$	0.636	0.636	-3.341	-3.341
	(0.004)	(0.004)	(0.005)	(0.005)

Notes: The table shows mean expectation errors and associated standard errors (in parentheses). Results are reported for two values of the quarterly employment correlation, $\hat{\rho}_e = 0.87$ and $\hat{\rho}_e = 0.95$, under both Gaussian and Gamma forecast error assumptions. The value $\hat{\rho}_e = 0.87$ corresponds to the average observed correlation between consecutive quarterly employment levels within a firm, while $\hat{\rho}_e = 0.95$ reflects an assumed correlation for monthly employment.

Appendix C Proofs

(i) Overall overestimation

Proof. Conditional on the expectation of the maximum first-year employment, we have

$$\mathbb{E}[\tilde{\mu}_1 - m_1 \mid \tilde{\mu}_1] = \tilde{\mu}_1 - \mathbb{E}[\theta \mid \tilde{\mu}_1].$$

Since $(\theta, \varepsilon_0, \phi)$ are jointly Gaussian and independent across components up to means, $(\theta, \tilde{\mu}_1)$ is jointly Gaussian. Hence the conditional expectation of θ given $\tilde{\mu}_1$ is affine in $\tilde{\mu}_1$:

$$\mathbb{E}[\theta \mid \tilde{\mu}_1] = \bar{\theta} + \frac{\mathrm{Cov}(\theta, \tilde{\mu}_1)}{\mathrm{Var}(\tilde{\mu}_1)} (\tilde{\mu}_1 - \bar{\theta}).$$

We compute the necessary moments. First,

$$\mathbb{E}[\tilde{\mu}_1] = (1-\kappa_0)\bar{\theta} + \kappa_0 \mathbb{E}[\theta + \varepsilon_0] + (1-\kappa_0)\mathbb{E}[\phi] = \bar{\theta}.$$

Next, using independence and centering of ε_0, ϕ,

$$\mathrm{Cov}(\theta, \tilde{\mu}_1) = \mathrm{Cov}(\theta,\ \kappa_0\theta) = \kappa_0 \,\mathrm{Var}(\theta) = \kappa_0 \sigma_\theta^2.$$

For the variance,

$$\mathrm{Var}(\tilde{\mu}_1) = \mathrm{Var}(\kappa_0(\theta + \varepsilon_0) + (1-\kappa_0)\phi) = \kappa_0^2 \mathrm{Var}(\theta + \varepsilon_0) + (1-\kappa_0)^2 \mathrm{Var}(\phi).$$

Since θ and ε_0 are independent, $\mathrm{Var}(\theta + \varepsilon_0) = \sigma_\theta^2 + \sigma_0^2$. With the definition of κ_0,

$$\kappa_0^2(\sigma_\theta^2 + \sigma_0^2) = \kappa_0 \sigma_\theta^2,$$

so

$$\sigma_{\tilde{\mu}_1}^2 = \mathrm{Var}(\tilde{\mu}_1) = \kappa_0 \sigma_\theta^2 + (1-\kappa_0)^2 \sigma_\phi^2.$$

Therefore,

$$\mathbb{E}[\theta \mid \tilde{\mu}_1] = \bar{\theta} + \frac{\kappa_0 \sigma_\theta^2}{\sigma_{\tilde{\mu}_1}^2} (\tilde{\mu}_1 - \bar{\theta}),$$

and subtracting from $\tilde{\mu}_1$ yields

$$\tilde{\mu}_1 - \mathbb{E}[\theta \mid \tilde{\mu}_1] = \left(1 - \frac{\kappa_0 \sigma_\theta^2}{\sigma_{\tilde{\mu}_1}^2}\right)(\tilde{\mu}_1 - \bar{\theta}) = \frac{\sigma_{\tilde{\mu}_1}^2 - \kappa_0 \sigma_\theta^2}{\sigma_{\tilde{\mu}_1}^2} (\tilde{\mu}_1 - \bar{\theta}).$$

Using $\sigma_{\tilde{\mu}_1}^2 - \kappa_0 \sigma_\theta^2 = (1-\kappa_0)^2 \sigma_\phi^2$ from the variance decomposition above gives the final equality.

Since $\tilde{\mu}_1 \sim \mathcal{N}(\bar{\theta}, \sigma_{\tilde{\mu}_1}^2)$, the truncated-normal mean is

$$\mathbb{E}[\tilde{\mu}_1 \mid \tilde{\mu}_1 > c_A] = \bar{\theta} + \sigma_{\tilde{\mu}_1} \lambda\!\left(\frac{c_A - \bar{\theta}}{\sigma_{\tilde{\mu}_1}}\right).$$

Combining,

$$\mathbb{E}[\tilde{\mu}_1 - m_1 \mid \tilde{\mu}_1 > c_A] = \frac{(1-\kappa_0)^2 \sigma_\phi^2}{\sigma_{\tilde{\mu}_1}} \lambda\!\left(\frac{c_A - \bar{\theta}}{\sigma_{\tilde{\mu}_1}}\right).$$

Since $(1-\kappa_0)^2 \sigma_\phi^2 > 0$, $\sigma_{\tilde{\mu}_1} > 0$, and $\lambda(\cdot) > 0$, the expression is strictly positive.

(ii) Underestimation among those with positive expected *and* actual entry

Proof. Conditional on the expectation $\tilde{\mu}_1$ and truncating on realized entry, we have

$$\mathbb{E}[\tilde{\mu}_1 - m_1 \mid \tilde{\mu}_1,\ m_1 > c_E] = \tilde{\mu}_1 - \mathbb{E}[m_1 \mid \tilde{\mu}_1,\ m_1 > c_E].$$

Since $(\tilde{\mu}_1, m_1)$ is jointly Gaussian, the conditional law of m_1 given $\tilde{\mu}_1$ is normal with

$$\mathbb{E}[m_1 \mid \tilde{\mu}_1] = b + a\,\tilde{\mu}_1, \qquad \mathrm{Var}(m_1 \mid \tilde{\mu}_1) = \sigma_y^2,$$

where

$$a = \frac{\mathrm{Cov}(m_1, \tilde{\mu}_1)}{\mathrm{Var}(\tilde{\mu}_1)} = \frac{\kappa_0 \sigma_\theta^2}{\sigma_{\tilde{\mu}_1}^2} \in (0,1), \qquad b = \mathbb{E}[m_1] - a\,\mathbb{E}[\tilde{\mu}_1] = (1-a)\bar{\theta},$$

and

$$\sigma_{\tilde{\mu}_1}^2 = \mathrm{Var}(\tilde{\mu}_1) = \kappa_0 \sigma_\theta^2 + (1-\kappa_0)^2 \sigma_\phi^2, \qquad \sigma_y^2 = \mathrm{Var}(m_1) - a^2\,\sigma_{\tilde{\mu}_1}^2 = \sigma_\theta^2 + \sigma_\varepsilon^2 - \frac{(\kappa_0 \sigma_\theta^2)^2}{\sigma_{\tilde{\mu}_1}^2} > 0.$$

By the truncated-normal mean formula,

$$\mathbb{E}[m_1 \mid \tilde{\mu}_1,\ m_1 > c_E] = b + a\,\tilde{\mu}_1 + \sigma_y\,\lambda\!\left(\frac{c_E - b - a\,\tilde{\mu}_1}{\sigma_y}\right).$$

Therefore

$$\mathbb{E}[\tilde{\mu}_1 - m_1 \mid \tilde{\mu}_1,\ m_1 > c_E] = (1-a)\,(\tilde{\mu}_1 - \bar{\theta})\ -\ \sigma_y\,\lambda\!\left(\frac{c_E - b - a\,\tilde{\mu}_1}{\sigma_y}\right).$$

The second term is strictly positive, hence for each fixed $\tilde{\mu}_1$,

$$\mathbb{E}[\tilde{\mu}_1 - m_1 \mid \tilde{\mu}_1,\ m_1 > c_E] < (1-a)\,(\tilde{\mu}_1 - \bar{\theta}).$$

We now average over the joint selection set $\{\tilde{\mu}_1 > c_A,\ m_1 > c_E\}$. Using the law of iterated expectations

$$\mathbb{E}[\tilde{\mu}_1 - m_1 \mid \tilde{\mu}_1 > c_A,\ m_1 > c_E] = \mathbb{E}\Big[\ \mathbb{E}[\tilde{\mu}_1 - m_1 \mid \tilde{\mu}_1,\ m_1 > c_E]\ \Big|\ \tilde{\mu}_1 > c_A,\ m_1 > c_E\Big].$$

Applying the pointwise bound yields

$$\mathbb{E}[\tilde{\mu}_1 - m_1 \mid \tilde{\mu}_1 > c_A,\ m_1 > c_E] < (1-a)\,\mathbb{E}\big[\tilde{\mu}_1 - \bar{\theta} \mid \tilde{\mu}_1 > c_A,\ m_1 > c_E\big] - \mathbb{E}\big[\sigma_y\,\lambda(\cdot) \mid \tilde{\mu}_1 > c_A,\ m_1 > c_E\big].$$

Since $\lambda(\cdot) > 0$ and $\sigma_y > 0$, the last term is strictly positive. Hence the conditional mean is strictly negative whenever

$$(1-a)\,\mathbb{E}\big[\tilde{\mu}_1 - \bar{\theta} \mid \tilde{\mu}_1 > c_A,\ m_1 > c_E\big] < \mathbb{E}\big[\sigma_y\,\lambda(\cdot) \mid \tilde{\mu}_1 > c_A,\ m_1 > c_E\big].$$

A simple sufficient condition is

$$c_E\ \geq\ b + a\,\mathbb{E}[\tilde{\mu}_1 \mid \tilde{\mu}_1 > c_A], \qquad c_E\ \geq\ \bar{\theta},$$

under which the truncation shift $\sigma_y\,\lambda\big(\frac{c_E - b - a\,\tilde{\mu}_1}{\sigma_y}\big)$ is uniformly bounded away from zero on a set of positive probability while $(1-a)(\tilde{\mu}_1 - \bar{\theta})$ remains bounded, implying

$$\mathbb{E}[\tilde{\mu}_1 - m_1 \mid \tilde{\mu}_1 > c_A,\ m_1 > c_E]\ <\ 0.$$

(iii) Mincer-Zarnowitz slope $\in (0,1)$

Proof. The Mincer-Zarnowitz slope can be expressed as

$$\beta_{MZ} = \frac{\mathrm{Cov}(m_1, \tilde{\mu}_1)}{\mathrm{Var}(\tilde{\mu}_1)}.$$

By independence and centering of $\varepsilon_0, \varepsilon_1, \phi$,

$$\mathrm{Cov}(m_1, \tilde{\mu}_1) = \mathrm{Cov}(\theta + \varepsilon_1,\ \kappa_0(\theta + \varepsilon_0) + (1-\kappa_0)\phi) = \kappa_0\,\mathrm{Var}(\theta) = \kappa_0 \sigma_\theta^2.$$

For the variance of $\tilde{\mu}_1$,

$$\mathrm{Var}(\tilde{\mu}_1) = \mathrm{Var}\big(\kappa_0(\theta + \varepsilon_0)\big) + \mathrm{Var}\big((1-\kappa_0)\phi\big) = \kappa_0^2(\sigma_\theta^2 + \sigma_0^2) + (1-\kappa_0)^2\sigma_\phi^2 = \kappa_0 \sigma_\theta^2 + (1-\kappa_0)^2\sigma_\phi^2,$$

where the last equality uses $\kappa_0 = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_0^2}$. Therefore

$$\beta_{\mathrm{MZ}} = \frac{\kappa_0 \sigma_\theta^2}{\kappa_0 \sigma_\theta^2 + (1-\kappa_0)^2 \sigma_\phi^2}.$$

If $\sigma_0 = +\infty$, then $\kappa_0 = 0$ and $\beta_{\text{MZ}} = 0$. If $\sigma_\phi^2 = 0$, the denominator equals the numerator and $\beta_{\text{MZ}} = 1$. If $\sigma_0 < +\infty$ and $\sigma_\phi^2 > 0$, then the denominator strictly exceeds the numerator, hence $0 < \beta_{\text{MZ}} < 1$. Note that σ_ε^2 drops out because it affects m_1 only through an innovation independent of $\tilde{\mu}_1$, hence it does not enter the covariance.

Now, we prove β_{MZ} does not change when conditional on the set of applicants $A = \{\tilde{\mu}_1 > c_A\}$, or the set of applicants with positive expected employment $B = \{\tilde{\mu}_1 > c_E\}$.

Since $E[m_1 \mid \tilde{\mu}_1] = b + a\,\tilde{\mu}_1$, define $\eta = m_1 - (b + a\,\tilde{\mu}_1)$. Then, $E[\eta \mid \tilde{\mu_1}] = 0$. Because A and B are selection events measurable with respect to $\tilde{\mu}_1$, we have $E[\eta \mid A] = E[\eta \mid B] = 0$ and $\text{Cov}(\eta, \tilde{\mu}_1 \mid A) = \text{Cov}(\eta, \tilde{\mu}_1 \mid B) = 0$. Hence,

$$\text{Cov}(m_1, \tilde{\mu}_1 \mid A) = \text{Cov}\big(b + a\tilde{\mu}_1 + \eta,\ \tilde{\mu}_1 \mid A\big) = a\,\text{Var}(\tilde{\mu}_1 \mid A),$$

$$\text{Cov}(m_1, \tilde{\mu}_1 \mid B) = \text{Cov}\big(b + a\tilde{\mu}_1 + \eta,\ \tilde{\mu}_1 \mid B\big) = a\,\text{Var}(\tilde{\mu}_1 \mid B).$$

So the Mincer–Zarnowitz slope among applicants and among applicants with positive expected employment equals the population slope:

$$\beta_{\text{MZ}}^{(A)} = \frac{\text{Cov}(m_1, \tilde{\mu}_1 \mid A)}{\text{Var}(\tilde{\mu}_1 \mid A)} = a = \frac{\text{Cov}(m_1, \tilde{\mu}_1)}{\text{Var}(\tilde{\mu}_1)} = \beta_{\text{MZ}},$$

$$\beta_{\text{MZ}}^{(B)} = \frac{\text{Cov}(m_1, \tilde{\mu}_1 \mid B)}{\text{Var}(\tilde{\mu}_1 \mid B)} = a = \frac{\text{Cov}(m_1, \tilde{\mu}_1)}{\text{Var}(\tilde{\mu}_1)} = \beta_{\text{MZ}}.$$

(iv) Predictive content of initial forecast errors (unconditional and conditional on S).

Proof. Fix any $t \geq 2$ and define

$$m_t = \theta + \varepsilon_t, \qquad m_1 = \theta + \varepsilon_1, \qquad \delta_1 = \tilde{\mu}_1 - m_1,$$

where ε_t is mean-zero and independent of $(\theta, \varepsilon_0, \varepsilon_1, \phi)$. Assume $(\theta, \tilde{\mu}_1, m_1)$ is jointly normal and $\kappa_0 > 0$. Let S be any event measurable with respect to $\tilde{\mu}_1$ (e.g. $S = \{\tilde{\mu}_1 > c\}$).

(iv-a) Unconditional result: $\beta_2 > 0$ in the regression of m_t on (m_1, δ_1). Consider the population regression

$$m_t = \beta_0 + \beta_1 m_1 + \beta_2 \delta_1 + u_t.$$

By the Frisch–Waugh–Lovell (FWL) theorem,

$$\beta_2 = \frac{\text{Cov}(m_t, \delta_{1,\perp})}{\text{Var}(\delta_{1,\perp})}, \qquad \delta_{1,\perp} = \delta_1 - \mathbb{E}[\delta_1 \mid m_1].$$

Since $\delta_1 = \tilde{\mu}_1 - m_1$ and m_1 is measurable w.r.t. $\sigma(m_1)$,

$$\delta_{1,\perp} = \tilde{\mu}_1 - \mathbb{E}[\tilde{\mu}_1 \mid m_1].$$

Moreover, $\varepsilon_t \perp (\tilde{\mu}_1, m_1)$ implies

$$\text{Cov}(m_t, \delta_{1,\perp}) = \text{Cov}(\theta + \varepsilon_t, \delta_{1,\perp}) = \text{Cov}(\theta, \delta_{1,\perp}).$$

By joint normality, the conditional expectation of θ is affine:

$$\mathbb{E}[\theta \mid \tilde{\mu}_1, m_1] = \alpha + \pi\,\tilde{\mu}_1 + \rho\, m_1.$$

Using the projection property and the fact that $\delta_{1,\perp}$ is measurable w.r.t. $(\tilde{\mu}_1, m_1)$,

$$\text{Cov}(\theta, \delta_{1,\perp}) = \text{Cov}\big(\mathbb{E}[\theta \mid \tilde{\mu}_1, m_1], \delta_{1,\perp}\big) = \text{Cov}(\alpha + \pi\tilde{\mu}_1 + \rho m_1, \delta_{1,\perp}).$$

The constant drops out. By definition of $\delta_{1,\perp}$ as a conditional-mean residual,

$$\text{Cov}(m_1, \delta_{1,\perp}) = 0.$$

Also, since $\tilde{\mu}_1 = \mathbb{E}[\tilde{\mu}_1 \mid m_1] + \delta_{1,\perp}$ and $\mathrm{Cov}(\mathbb{E}[\tilde{\mu}_1 \mid m_1], \delta_{1,\perp}) = 0$, we have

$$\mathrm{Cov}(\tilde{\mu}_1, \delta_{1,\perp}) = \mathrm{Var}(\delta_{1,\perp}).$$

Therefore,

$$\mathrm{Cov}(\theta, \delta_{1,\perp}) = \pi \,\mathrm{Var}(\delta_{1,\perp}), \qquad \Rightarrow \qquad \beta_2 = \pi.$$

Closed form for π. Let $X = (\tilde{\mu}_1, m_1)^\top$. For jointly normal variables, $(\pi, \rho) = \mathrm{Cov}(\theta, X)\,\mathrm{Var}(X)^{-1}$. Using

$$\mathrm{Cov}(\theta, \tilde{\mu}_1) = \kappa_0 \sigma_\theta^2, \quad \mathrm{Cov}(\theta, m_1) = \sigma_\theta^2, \quad \mathrm{Cov}(\tilde{\mu}_1, m_1) = \kappa_0 \sigma_\theta^2, \quad \mathrm{Var}(m_1) = \sigma_\theta^2 + \sigma_\varepsilon^2,$$

and writing $\mathrm{Var}(\tilde{\mu}_1)$ for the variance of $\tilde{\mu}_1$, a 2×2 inversion yields

$$\pi = \frac{\kappa_0 \, \sigma_\theta^2 \, \sigma_\varepsilon^2}{(\sigma_\theta^2 + \sigma_\varepsilon^2)\,\mathrm{Var}(\tilde{\mu}_1) - (\kappa_0 \sigma_\theta^2)^2} \; > \; 0,$$

where the denominator is positive whenever $\mathrm{Var}(\tilde{\mu}_1 \mid m_1) > 0$ (i.e., $\tilde{\mu}_1$ is not perfectly collinear with m_1). Hence $\beta_2 > 0$ for every $t \geq 2$.

(iv-b) Conditional result: selection on $\tilde{\mu}_1$ leaves the result unchanged. Let S denote a selection event that depends only on $\tilde{\mu}_1$, e.g. $S \in \{A, B\}$ with A and B as defined in part (iii) above. Consider the within-S (population) regression

$$m_t = \beta_{0,S} + \beta_{1,S} m_1 + \beta_{2,S} \delta_1 + u_{t,S}.$$

By FWL applied within S,

$$\beta_{2,S} = \frac{\mathrm{Cov}(m_t, \delta_{1,S}^\perp \mid S)}{\mathrm{Var}(\delta_{1,S}^\perp \mid S)}, \qquad \delta_{1,S}^\perp = \delta_1 - \mathbb{E}[\delta_1 \mid m_1, S] = \tilde{\mu}_1 - \mathbb{E}[\tilde{\mu}_1 \mid m_1, S].$$

Because $\varepsilon_t \perp (\tilde{\mu}_1, m_1)$ and S depends only on $\tilde{\mu}_1$, conditioning on S does not affect this independence, so $\varepsilon_t \perp (\tilde{\mu}_1, m_1) \mid S$. Thus

$$\mathrm{Cov}(m_t, \delta_{1,S}^\perp \mid S) = \mathrm{Cov}(\theta + \varepsilon_t, \delta_{1,S}^\perp \mid S) = \mathrm{Cov}(\theta, \delta_{1,S}^\perp \mid S).$$

Let $Z = (\tilde{\mu}_1, m_1, S)$. Since $\delta_{1,S}^\perp$ is measurable w.r.t. Z and $\theta - \mathbb{E}[\theta \mid Z]$ is orthogonal to all Z-measurable random variables,

$$\mathrm{Cov}(\theta, \delta_{1,S}^\perp \mid S) = \mathrm{Cov}\big(\mathbb{E}[\theta \mid \tilde{\mu}_1, m_1, S], \delta_{1,S}^\perp \mid S\big).$$

Moreover, because S is measurable w.r.t. $\tilde{\mu}_1$, conditioning on $(\tilde{\mu}_1, m_1)$ already reveals whether S occurs, hence

$$\mathbb{E}[\theta \mid \tilde{\mu}_1, m_1, S] = \mathbb{E}[\theta \mid \tilde{\mu}_1, m_1] = \alpha + \pi\, \tilde{\mu}_1 + \rho\, m_1,$$

with the same (π, ρ) as in part (iv-a). Therefore,

$$\mathrm{Cov}(\theta, \delta_{1,S}^\perp \mid S) = \mathrm{Cov}(\alpha + \pi\tilde{\mu}_1 + \rho m_1, \delta_{1,S}^\perp \mid S).$$

The constant drops out. By the definition of $\delta_{1,S}^\perp$ as a conditional-mean residual,

$$\mathrm{Cov}(m_1, \delta_{1,S}^\perp \mid S) = 0.$$

Also, since $\tilde{\mu}_1 = \mathbb{E}[\tilde{\mu}_1 \mid m_1, S] + \delta_{1,S}^\perp$ and $\mathrm{Cov}(\mathbb{E}[\tilde{\mu}_1 \mid m_1, S], \delta_{1,S}^\perp \mid S) = 0$, we have

$$\mathrm{Cov}(\tilde{\mu}_1, \delta_{1,S}^\perp \mid S) = \mathrm{Var}(\delta_{1,S}^\perp \mid S).$$

Hence,

$$\mathrm{Cov}(\theta, \delta_{1,S}^\perp \mid S) = \pi \,\mathrm{Var}(\delta_{1,S}^\perp \mid S), \qquad \Rightarrow \qquad \beta_{2,S} = \pi.$$

Since $\pi > 0$ from part (iv-a), we conclude $\beta_{2,S} > 0$ for every $t \geq 2$ and every selection event S measurable w.r.t. $\tilde{\mu}_1$, provided $\mathrm{Var}(\delta_{1,S}^\perp \mid S) > 0$ (nondegeneracy).

(v) Proof for Proposition 1

(v-a) p_j^{exc} strictly increases in σ_ϕ whenever $\Pr(\tilde{\mu}_1^0 < c_j) > 0$. In the model representation,

$$\tilde{\mu}_1 = \tilde{\mu}_1^0 + (1-\kappa_0)\phi, \qquad \tilde{\mu}_1^0 \sim \mathcal{N}(\bar{\theta},\ \kappa_0\sigma_\theta^2), \qquad \phi \sim \mathcal{N}(0, \sigma_\phi^2),$$

with $\tilde{\mu}_1^0 \perp \phi$ and $\kappa_0 = \sigma_\theta^2/(\sigma_\theta^2 + \sigma_0^2) \in (0,1)$. Note that

$$p_j^{\text{exc}} = \Pr\big(\tilde{\mu}_1^0 \le c_j,\ \tilde{\mu}_1 > c_j\big) = \Pr\big(\tilde{\mu}_1^0 \le c_j,\ \tilde{\mu}_1^0 + (1-\kappa_0)\phi > c_j\big). \tag{18}$$

Conditioning on $\tilde{\mu}_1^0 = u \le c_j$ yields

$$\Pr\big(\tilde{\mu}_1 > c_j \mid \tilde{\mu}_1^0 = u\big) = \Pr\left(\phi > \frac{c_j - u}{1-\kappa_0}\right) = 1 - \Phi\left(\frac{c_j - u}{(1-\kappa_0)\sigma_\phi}\right) = \Phi\left(\frac{u - c_j}{(1-\kappa_0)\sigma_\phi}\right).$$

Therefore,

$$p_j^{\text{exc}} = \int_{-\infty}^{c_j} \Phi\left(\frac{u - c_j}{(1-\kappa_0)\sigma_\phi}\right) f_{\tilde{\mu}_1^0}(u)\, du, \qquad \tilde{\mu}_1^0 \sim \mathcal{N}(\bar{\theta},\ \kappa_0\sigma_\theta^2). \tag{19}$$

Fix any $u < c_j$. The integrand in (19) is $g(\sigma_\phi; u) = \Phi\Big(\frac{(u-c_j)}{((1-\kappa_0)\sigma_\phi)}\Big)$. Since $u - c_j < 0$, the map $\sigma_\phi \mapsto (u-c_j)/((1-\kappa_0)\sigma_\phi)$ is strictly increasing (toward 0), and Φ is strictly increasing, hence $g(\sigma_\phi; u)$ is strictly increasing in σ_ϕ for all $u < c_j$. Because $f_{\tilde{\mu}_1^0}(u) \ge 0$, integrating over $u \in (-\infty, c_j)$ implies $\partial p_j^{\text{exc}}/\partial\sigma_\phi > 0$ whenever $\Pr(\tilde{\mu}_1^0 < c_j) > 0$.

(v-b) p_j^{for} strictly increases in σ_ϕ whenever $\Pr(\tilde{\mu}_1^0 > c_j) > 0$. Consider the definition of p_j^{for}

$$p_j^{\text{for}} \ = \ \Pr\big(\tilde{\mu}_1^0 > c_j,\ \tilde{\mu}_1 \le c_j\big) = \Pr\big(\tilde{\mu}_1^0 > c_j,\ \tilde{\mu}_1^0 + (1-\kappa_0)\phi \le c_j\big). \tag{20}$$

Conditioning on $\tilde{\mu}_1^0 = u > c_j$ yields

$$\Pr\big(\tilde{\mu}_1 \le c_j \mid \tilde{\mu}_1^0 = u\big) = \Pr\left(\phi \le \frac{c_j - u}{1-\kappa_0}\right) = \Phi\left(\frac{c_j - u}{(1-\kappa_0)\sigma_\phi}\right) = \Phi\left(\frac{c_j - u}{(1-\kappa_0)\sigma_\phi}\right).$$

Therefore,

$$p_j^{\text{for}} = \int_{c_j}^{\infty} \Phi\left(\frac{c_j - u}{(1-\kappa_0)\sigma_\phi}\right) f_{\tilde{\mu}_1^0}(u)\, du, \qquad \tilde{\mu}_1^0 \sim \mathcal{N}(\bar{\theta},\ \kappa_0\sigma_\theta^2). \tag{21}$$

Fix any $u > c_j$. The integrand in (21) is $h(\sigma_\phi; u) = \Phi\Big(\frac{c_j - u}{(1-\kappa_0)\sigma_\phi}\Big)$. Since $c_j - u < 0$, the map $\sigma_\phi \mapsto (c_j - u)/((1-\kappa_0)\sigma_\phi)$ is strictly increasing (toward 0), and Φ is strictly increasing, hence $h(\sigma_\phi; u)$ is strictly increasing in σ_ϕ for all $u > c_j$. Because $f_{\tilde{\mu}_1^0}(u) \ge 0$, integrating over $u \in (c_j, \infty)$ implies $\partial p_j^{\text{for}}/\partial\sigma_\phi > 0$ whenever $\Pr(\tilde{\mu}_1^0 > c_j) > 0$.

Combining (v-a) and (v-b) completes the proposition.