Production Function Identification for the Computer and Electronic Products Manufacturing Industry

Alexander Peralta

Indiana University – Kelley School of Business

May 4, 2025

1. Introduction

Commonly, production functions are estimated by economists using data for which companies have been de-identified by the data provider. For example, Olley and Pakes (1996) used de-identified plant-level data from the U.S. Census Annual Survey of Manufacturers, Levinsohn and Petrin (2003) used de-identified plant-level data from the Chilean Annual National Industrial Survey, and Ackerberg, Caves, and Frazer (2015) used de-identified plant-level data from the Colombia Annual Manufacturing Survey. By using de-identified data, economists are limited to analyzing industries at the level of aggregation determined by the data provider.

To overcome this limitation, I propose an alternative approach using quarterly firm-level financial statement data from Compustat Capital-IQ, allowing production function identification at the sub-industry level. Specifically, I implement and compare multiple production function estimation techniques—such as Fixed Effects Models, as well as the Olley-Pakes and Ackerberg-Caves-Frazer methods—offering new insights into the effectiveness of these approaches when applied to financial statement data. In doing so, I provide novel estimates of production functions for the Computer and Electronic Products Manufacturing Industry.

The resulting findings have implications in terms of specialization.

2. Data

Using the SEC Ticker Mapping API, I retrieved a list of all public companies who have filed financial statements with the SEC, as well as the CIK Numbers and SIC Codes associated with these companies. I then filtered this list to only include companies with SIC Codes that correspond to the NAICS Code for the Computer and Electronic Product Manufacturing Industry. Finally, I used Compustat Capital-IQ to collect annual and quarterly financial statement data for all companies in the filtered list for the years 2010 through 2025.

This process described above resulted in two unbalanced panel datasets: one dataset for quarterly financial data, which includes 25,645 observations, and another dataset for annual financial data, which includes 8,169 observations. The dataset for annual financial data includes total revenue (output), employment (labor), and variables that can be used to approximate capital.

3. Identifying Simultaneity Bias and Measurement Error

I start with a Cobb-Douglass production function $Y = zF(K, L)$ where Y, z, and $F(K, L)$ are output, total factor productivity, and an unspecified function of capital and labor, respectively. The production function can be rewritten to explicitly state capital and labor:

$$Y_{it} = zK_{it}^{\alpha}L_{it}^{\beta} \tag{3.1}$$

where the subscript i relates capital and labor to a specific firm and subscript t relates the time period to a specific firm. The production function in (3.1) assumes that capital evolves according to the law of motion for capital accumulation:

$$K_{it} = (1 - \delta_{it})K_{it-1} + I_{it-n} \tag{3.2}$$

where capital expenditure (investment) is calculated as:

$$I_{it-n} = K_{it} - K_{it-n} + Current\ Depreciation_{it} \tag{3.3}$$

That is, investment for firm i at time $t - n$ is equal to the difference between the capital stock of firm i in the current period and the capital stock of firm i, n periods ago, plus depreciation in the current period for firm i. The subscript $t - n$ is used to represent that it takes investment n periods to materialize into capital; that is, I use an n-period lag on investment.

Similar to Levinsohn-Petrin (2003), I approximate capital (K_{it} and K_{it-n}) by taking the sum of inventory, as well as depreciated property, plant, and equipment for a given firm during the appropriate periods.

Because the depreciation rate is $\delta_{it} = \frac{Depreciation_{it} - Depreciation_{it-1}}{K_{it-1}}$, (3.2) can be rewritten as:

$$K_{it} = K_{it-1} - Depreciation_{it} + I_{it-n} \tag{3.4}$$

I calculate 2 measures of capital using (3.4) in Python. The first measure of capital (capital measure 1) does not use a lag on investment, and the second measure of capital (capital measure 2) uses a 3-quarter lag on investment. I denote the regression models using the first and second measures of capital as Model 1 and Model 2, respectively.

Quarterly employment is estimated using a quantile regression forest (QRF), as explained in Appendix A1.

To estimate the production function in (3.2), I take the logarithm of both sides of the equation:

$$\log(Y_{it}) = \log(z) + \alpha \log(K_{it}) + \beta \log(L_{it}) \tag{3.6}$$

Variable	**Observations**	**Mean (in millions)**	**St. Dev. (in millions)**	**Min (in millions)**	**Max (in millions)**
Log Output	17,840	4.37	2.37	0.001	11.73
Log Labor	17,840	1.61	0.02	1.55	1.70
Log Capital Measure 1	17,840	5.35	2.63	0.001	12.95
Log Capital Measure 2	17,840	4.92	3.06	0.00	13.05

Table 3.1: Summary Statistics for Equation (3.6) Parameters

Table 3.1 presents summary statistics for the variables in Equation (3.6). The data suggest that larger firms are underrepresented in the sample. The mean of log output is 4.37, with a standard deviation of 2.37, which corresponds to an average annual revenue of approximately \$79.04 million ($e^{4.37} \approx 79.04$). The large standard deviation indicates substantial variation in firm size, but the average implies that relatively smaller firms dominate the sample distribution. In terms of production function identification, the underrepresentation of the largest firms may limit the identification of scale effects in the production function, as estimates would be driven primarily by smaller firms. For example, if larger firms exhibit increasing returns to scale due to greater capital intensity or organizational efficiency, their absence in the sample may lead to an underestimation of output elasticities or mask nonlinearities in the production function. As a result, the estimated coefficients may primarily reflect the technology and input-output relationships of smaller firms, potentially misrepresenting the true production structure of the industry as a whole.

Running a regression using (3.5) would require data on total factor productivity. Because I do not have data on total factor productivity, the regression specification for (3.5) can be modified by absorbing total factor productivity in the error term:

$$y_{it} = \beta_0 + \beta_1 k_{it} + \beta_2 l_{it} + \epsilon_{it} \tag{3.7}$$

Here, all variables are in logarithmic form, and ϵ_{it} represents the error term, which captures unobserved productivity shocks and measurement error. In equation (3.7), I impose no additional assumptions about the error term. Subsequently, total factor productivity can be computed using the Solow Residual:

$$TFP_{it} = \epsilon_{it} = \exp\left(y_{it} - \beta_0 + \beta_1 k_{it} + \beta_2 n_{it}\right) \tag{3.8}$$

To identify simultaneity bias and measurement error for the simple production function (3.6), I ran an ordinary least squares regression in Python using equation (3.7). The summary statistics for this regression are presented in Table 3.2.

	Model 1	**Model 2**
Const	-8.89***	-43.21***
Capital	0.81*** (0.00)	0.50*** (0.00)
Labor	5.46*** (0.30)	27.95*** (0.47)
Number of Observations	17,840	17,840

R-Squared	0.89	0.68
F-Stat	69113.99	18572.33
SSR	11450.09	32500.97

Table 3.2: Regression Summary for Equation (3.7)

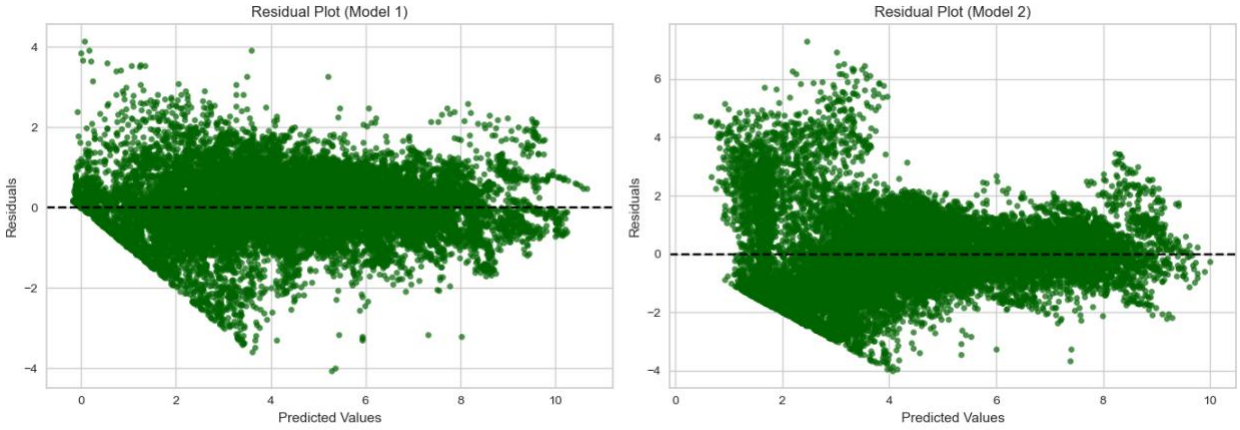

Figure 3.1: Scatter Plot of Predicted Values vs Residual Values

As outlined in Levinsohn-Petrin (2003), the coefficient on labor in both models overestimates the actual effect of labor on output, and the coefficients on capital underestimate the actual effect of capital on output. This suggests that capital and labor are positively correlated, and that labor is more strongly correlated with unobserved productivity shocks than is capital; that is, the estimates may suffer from simultaneity bias.

The upward bias on the coefficient for capital also indicates the presence of measurement error for our measure of capital. That is, based on Collard-Wexler and De Loecker's (2022) findings on capital mismeasurement in firm-level data, the coefficient on labor is picking up "the true capital variation through the positive correlation of labor and (true) capital and thus overestimate the impact of labor on output variation."

In addition to simultaneity bias and measurement error, the simple models also suffer from unobserved heterogeneity across units. Specifically, between-unit variation accounts for 96.84% of total variance.[1] This high proportion suggests that most of the variation in the output is explained by persistent differences across firms rather than within-unit (i.e., over-time) variation, highlighting the need for models that appropriately account for fixed effects or latent heterogeneity.

4. Examining Methods to Correct the Production Function for Endogeneity

[1] For reference, between-unit variation is calculated as $\sum_{i=1}^{G}(\bar{y}_i - \bar{y})^2 * n_i$.

4.1. Fixed Effects Models

In his 1969 paper, Irving Hoch noted that using least squares estimates of Cobb-Douglas production function parameters will lead to simultaneity bias, as the inputs are correlated with the error term. That is, given the output formula and the input demand formula as follows:

$$x_{0i} = k_0 + \sum_q a_q x_{qi} + u_i \tag{4.1.1}$$

$$x_{qi} = c_q + x_{0i} + v_{qi} \tag{4.1.2}$$

Substituting the output equation into the input demand equation:

$$x_{qi} = c_q + k_0 + \sum_q a_q x_{qi} + u_i + v_{qi} \tag{4.1.3}$$

shows that x_{qi} is a function of both x_{qi} and u_i, which implies that input demand is selected based on the error term and that the specification in (4.1.1) is endogenous.

To avoid simultaneity bias, Hoch suggested the use of "combined time-series and cross-section data" to include firm and time fixed effects in production function estimation. Firm fixed effects reflect differences in technical efficiency, while time fixed effects reflect "change in technical efficiency over time and differences in weather between" and potentially "changes in both relative prices and the general price level":

$$mpq = \sum_i \sum_t \left(x_{pit} - \bar{x}_{pi} - \bar{x}_{qt} + \bar{\bar{x}}_p\right)\left(x_{qit} - \bar{x}_{qi} - \bar{x}_{qt} + \bar{\bar{x}}_q\right) \tag{4.1.4}$$

where $\bar{x}_{p_i}$ removes time-invariant firm differences to show how the firm deviates from its own average, $\bar{x}_{p_i}$ removes time-invariant year differences to show how the firm deviates from its own average and the year's average, and $\bar{\bar{x}}_p$ adjusts for the removal of both firm and time averages. This transformation isolates the variation in $x_{p_{it}}$ that is not due to either the firm's average level or the year's average level, which provides a consistent and unbiased estimate of the firm's production function under certain conditions.

To estimate the production function while addressing simultaneity bias arising from unobserved firm-specific and time-specific heterogeneity, I implement a fixed effects panel regression model following the approach proposed by Hoch. The model specification is as follows:

$$\log(Y_{it}) = \alpha_i + \delta_t + \beta_1 \log(K_{it}) + \beta_2 \log(L_{it}) + \epsilon_{it} \tag{4.1.5}$$

where Y_{it} is total revenue for firm i at time t, K_{it} is capital, L_{it} is labor, α_i denotes firm fixed effects, and δ_t captures time fixed effects. Table 4.1.1 presents the regression results for (4.1.5).

	Model 1	**Model 2**
Constant	2.1414***	0.7822***

Capital	0.4806*** (0.0229)	0.1039*** (0.0151)
Labor	6.02*** (0.9882)	10.431*** (1.6775)
R-Squared	0.5634	0.3237
F-Stat	338.24	9.7543
Number of Observations	17,838	17,838

Table 4.1.1: Fixed Effects Regression Results

Relative to the ordinary least squares regression results from Table 3.2, time and firm fixed effects appears to have reduced measurement error for Model 2, while measurement error for Model 1 appears to be slightly worse. However, the results for the fixed effects regression still indicate the presence of measurement error, as the coefficient estimates for capital and labor tend to overstate the contribution of labor to output, and the coefficient for capital is biased downwards. There are several reasons why this may be the case.

First, both Olley-Pakes (1996) and Griliches and Mairess (1995) show that "firms with a large capital stock are more likely to remain in business and tolerate lower productivity levels", which may "introduce a negative bias in β_K". A negative coefficient for the capital coefficient could be explained by the fact that "capital is a fixed factor of production, and, therefore, the variation left in the time series is essentially noise." That is, because firms are generally able to hire and fire workers faster than they are able to operationalize and sell-off capital, large firms are likely to hold on to capital during business cycle contractions, during which output may be below trend.

Second, fixed effects models assume that firm-specific productivity is constant over time. However, this assumption is often violated, as firms experiencing positive productivity shocks are more likely to increase inputs and survive, introducing a selection problem that fixed effects cannot fully address. If this selection is not orthogonal to input decisions, fixed effect estimators may yield biased coefficients. Based on our regression results for the fixed effects model, the upward bias of the labor coefficient reflect the fact that firms adjust labor more rapidly in response to shocks than they do capital, which is typically more fixed in the short run. As a result, observed changes in labor may be confounded with transitory productivity shocks, inflating its estimated effect.

Finally, while the QRF model provides a flexible and accurate approach for estimating employment from annual firm characteristics, its predictions exhibit limited within-firm variation due to the quantile-based structure. As a result, labor input values are clustered and discretized, reducing the effective signal available for identifying

labor elasticities in fixed effects regressions. This contributes to the attenuation and instability of the estimated labor coefficient and may exacerbate the effects of measurement error.

4.2. The Olley-Pakes Model

In estimating the parameters of the production function for the telecommunications industry, Olley-Pakes (1996) noted that total factor productivity is not observable by the economist, and that "exit and input demand decisions" are made by companies based on total factor productivity, which generated the "simultaneity problem" and "the issue of how to handle attrition from, and additions to, the data." The simultaneity problem would cause total factor productivity to be serially correlated with input choices, causing an upward bias in productivity differences as estimated via ordinary least squares. The problem of firm entry and exit may cause the sample to over-represent surviving firms with high-capital and low-productivity, as firms with higher capital are more likely to survive even when they have low productivity. This creates a negative correlation between capital and unobserved productivity, which leads to a downward bias in the estimated capital coefficient in OLS or models that don't control for this selection.

To address these issues, Olley and Pakes propose a two-stage estimation method that uses investment as a proxy for productivity and models firm survival as part of the decision process.

To begin, Olley-Pakes express unobservable total factor productivity as a function of observable factors to control for productivity in estimation:

$$\omega_t = h_t(i_t, a_t, k_t) \tag{4.2.1}$$

This can be substituted into the production function, resulting in the following semiparametric regression model:

$$y_{it} = \beta_l l_{it} + \phi(i_{it}, a_{it}, k_{it}) + \eta_{it} \tag{4.2.2}$$

Because the semiparametric regression model does not separately identify the effects of capital and age on output (since they also influence investment), Olley and Pakes recover the coefficients on capital and age in a second stage by leveraging information from the survival decision. They estimate the probability that a firm remains in the market—i.e., the selection equation—and use it to control for selection bias caused by endogenous exit.

Specifically, the probability of survival is modeled as a function of observable state variables, which allows the researcher to condition on the selection probability (i.e., the propensity score) to recover a second index of the unobserved productivity term. For given values of β_α and β_k, the second index can be conditioned upon using the nonlinear term generated from the first-stage estimation:

$$\phi(i_{it}, a_{it}, k_{it}) = \beta_0 + \beta_a a_{it} + \beta_k k_{it} + h_t(i_{it}, a_{it}, k_{it}) \tag{4.2.3}$$

To recover the coefficients on capital and age, which could not be separately identified in the first stage, Olley-Pakes introduce a second-stage equation that controls for selection bias due to endogenous firm exit. This equation isolates the contributions of capital and age to output in period $t+1$, after controlling for the effect of labor on output. Specifically, the authors regress $y_{t+1} - \beta_l l_{t+1}$ on k_{t+1} and a_{t+1}, along with a control function $g(\cdot)$ that accounts for the correlation between unobserved productivity and firm survival. The control function depends on the estimated survival probability (i.e., the propensity score) and the nonlinear productivity term from the first stage, adjusted for capital and age.

Substituting the survival probability P_t and ϕ_t into $g(\cdot)$:

$$y_{t+1} - \beta_l l_{t+1} = \beta_a a_{t+1} + \beta_k k_{t+1} + g(P_t, \phi_t - \beta_a a_t - \beta_k k_t + \xi_{t+1} + \eta_{t+1} \quad (4.2.4)$$

where ξ_{t+1} is the innovation to productivity and η_{t+1} is a measurement error or output shock. This approach ensures that the capital coefficient is consistently estimated even when productivity is unobserved and selection into the sample depends on it.

To estimate the first stage of the Olley-Pakes model, I specified a flexible approximation of the unobserved productivity term $\phi(i_{it}, k_{it})$ using a third-degree polynomial in investment and capital. The labor input was included linearly, as it is assumed to be an adjustable variable correlated contemporaneously with productivity. I then estimated a linear regression of total revenue on this flexible function of capital and investment, along with labor. The fitted values from this regression served as the first-stage estimate of the composite productivity term $\phi(i_{it}, k_{it})$.

	Model 2	**Std Error**	**P>|z|**
$const$	-2.8301		
$Investment$	1.5384	0.038	0.000
$Capital$	-0.2686	0.029	0.000
$Investment^2$	-0.0286	0.012	0.016
$Investment * Capital$	-0.3725	0.011	0.000
$Capital^2$	0.2667	0.009	0.000
$Investment^3$	0.0013	0.001	0.279
$Investment^2 * Capital$	-0.0009	0.002	0.565

$Investment * Capital^2$	0.0311	0.002	0.000
$Capital^3$	0.0203	0.001	0.000
$Labor$	3.4394	0.259	0.000
R-Squared	0.881		
F-Stat	8747		
Number of Observations	10700		

Following the first-stage estimation, I extracted the predicted values from the regression to get an estimate the firm-specific productivity term, $\hat{\phi}_{it}$. To account for selection bias from endogenous firm exit, I estimated a Probit model of firm survival on the same set of regressors used in the first stage. This allowed me to model the probability that a firm remains in the sample as a function of its observed state variables and productivity. The predicted survival probabilities from this model were saved and later used in the second-stage regression to correct for potential selection bias in the estimation of capital and age coefficients.

To implement the second stage of the Olley-Pakes model, I first dropped any firm-period observations with missing values for next-period output, capital, estimated productivity, or survival probability. I then constructed a second-degree polynomial in the first-stage productivity estimate $\hat{\phi}_{it}$ and the predicted survival probability to approximate the selection correction term. This flexible control function captures the correlation between unobserved productivity and the likelihood of firm survival. I included next-period capital as a separate regressor and estimated a linear regression of next-period output on this full set of variables. The coefficient on capital recovered from this regression provides a consistent estimate of the capital elasticity, accounting for both simultaneity and selection biases.

	Model 2	**Std Error**	**P>|z|**
const	7.0917		
ϕ	0.7922	0.099	0.000
Survival Probability	-15.1014	8.107	0.001
ϕ^2	-0.0044	0.002	0.000

$\phi * \ Survival\ Probability$	0.1527	0.103	0.285
$Survival\ Probability^2$	7.9870	4.390	0.001
Capital	0.0901	0.008	0.000
R-Squared	0.881		
F-Stat	1.348e+04		
Number of Observations	9979		
Durbin-Watson	0.509		

Based on the second stage regression results, the Olley-Pakes method when applied to the data substantially reduced measurement error in capital and labor estimates relative to the ordinary least squares regression. The proxy for unobserved productivity enters positively and significantly, confirming its relevance in explaining output. The inclusion of higher-order terms interactions with survival probability suggests diminishing returns to productivity and some nonlinearity in the selection mechanism. The survival probability itself is significantly negative, while its squared term is positive, indicating that firms more likely to exit tend to be less productive, but the effect is nonlinear. The coefficient on capital is estimated at 0.0901 and is statistically significant. However, the downward bias of capital suggests that our model still suffers from measurement error. There are several reasons why this may be the case.

First, to test whether investment demand can be expressed as a function of capital, age, and productivity, I implemented a robustness check proposed in their original paper. This assumption is central to identifying the labor coefficient β_l in the first-stage regression. If this assumption is violated, the second-stage disturbance term may contain residual correlation with labor, particularly with l_{t+1}, since labor and unobserved productivity are likely to be correlated. To test this, I extended the second-stage regression by including l_{t+1} (i.e., Total Employment in period $t+1$) as an additional regressor. Using a parametric specification, I approximated the control function $g(P_t, \phi_t - \beta_a a_t - \beta_k k_t)$ with a second-degree polynomial in the survival probability and the first-stage productivity index. The coefficient on labor was estimated at -168.77, with a p-value less than 1e-81, indicating extremely strong statistical significance. This result suggests that l_{t+1} is significantly correlated with the second-stage error term, implying that the original estimate of β_l may be biased. Thus, the investment demand function may not be fully captured by capital, age, and productivity alone, and the strict monotonicity condition of the investment demand equation may not hold in this context.

Second, investment may not serve as a strong or valid proxy for unobserved productivity in our sample. The identification strategy in Olley-Pakes relies critically on the assumption that investment is a strictly increasing function of productivity, conditional on capital and age. However, if investment is lumpy, subject to reporting error, or poorly measured, this relationship may be weak or non-monotonic. Additionally, if a large number of firms exhibit zero or near-zero investment in a given period, the required invertibility condition may not hold, thereby undermining the reliability of the control function. In such cases, the simultaneity correction introduced by the OP framework may introduce more noise than it resolves, resulting in greater bias than a naïve OLS estimate.

Finally, firm age is not available in our dataset. Age is used in the OP model both as a determinant of investment and survival, and as a control in the productivity inversion step. Excluding it may bias the control function $\phi(i_t, k_t, a_t)$ and the survival equation, leading to misestimation of the unobserved productivity term and undermining the second-stage corrections for selection bias. Without this variable, the model may confound the effects of capital and productivity with omitted age-related heterogeneity, weakening identification and reducing the effectiveness of the simultaneity and selection corrections.

With the labor coefficient β_l obtained from the first-stage regression and the capital coefficient β_k recovered from the second stage, I computed firm-level total factor productivity as the residual component of output unexplained by observed inputs. Specifically, productivity was calculated as the exponent of the difference between a firm's observed revenue and the predicted contribution of labor and capital:

$$\widehat{TFP_{it}} = \exp\left(y_{it} - \beta_l l_{it} - \beta_k k_{it} \right) \tag{4.2.5}$$

There are a few things to note in 4.1.1. First, aggregate total factor productivity was low immediately and several years after the 2008-2009 great recession. This is likely due to the fact that companies in the Computer Manufacturing and Electronic Products industry were still recovering from the .com crisis. Many firms had overinvested in infrastructure and technologies during the tech boom, leading to a period of adjustment and underutilized capacity. Additionally, in the wake of the financial crisis, credit constraints, reduced investment, and slow diffusion of productivity-enhancing technologies likely contributed to the prolonged stagnation in TFP. Second, aggregate total factor productivity begam to grow at a much higher rate after 2017. This period coincides with increased domestic investment in semiconductor and computing infrastructure, as well as rising demand for high-performance computing, cloud services, and AI-related technologies. Additionally, policy developments—including early discussions of reshoring critical supply chains and the anticipation of legislation like the CHIPS Act—likely contributed to renewed capital formation and operational efficiency. The acceleration in TFP growth also aligns with a broader pivot away from low-margin commodity hardware toward more specialized and scalable technologies, signaling a move up the global value chain and a partial reversal of the offshoring trends that characterized the industry in the early 2010s.

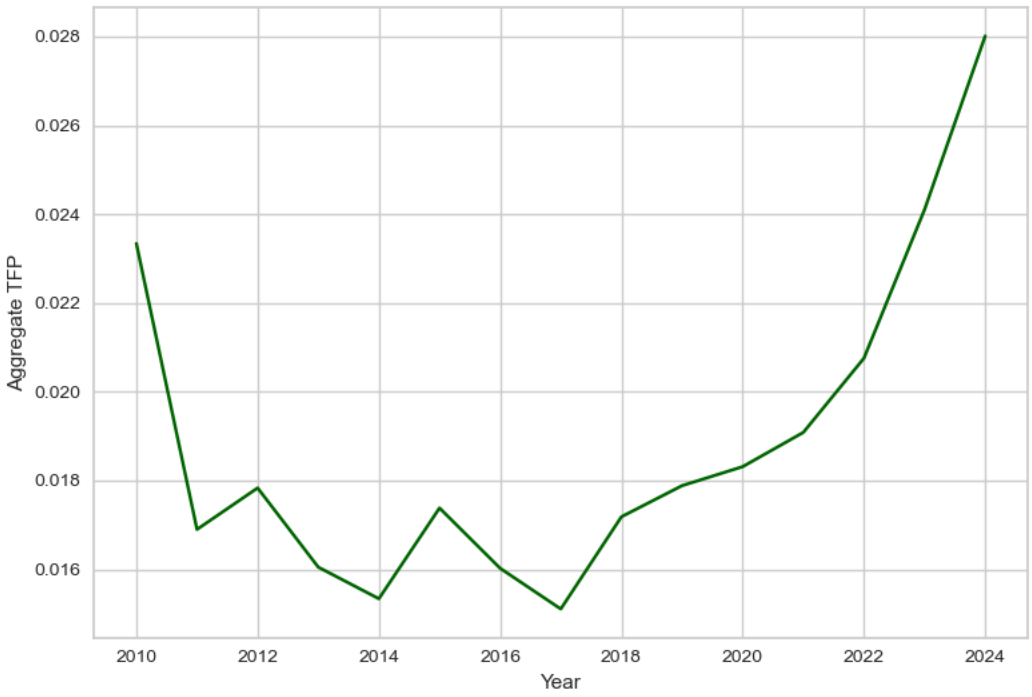

Figure 4.1.1: Industry-Level Aggregate TFP Over Time

4.3. Ackerberg-Caves-Frazer Model with Copula Endogeneity Correction

Ackerberg, Caves, Frazer (ACF) (2015) argue that production function estimation techniques such as the Olley-Pakes (OP) and Levinsohn-Petrin (LP) methods suffer from functional dependence. Specifically, the OP and LP methods rely on inverting the firm's input demand function to proxy for unobserved productivity, and estimate the coefficient for labor by assuming that labor is chosen after productivity is observed and is not a function of the proxy variables. However, if labor is also functionally determined by capital, materials, and time, there is no remaining variation in labor input that can be used to identify the coefficient for labor.

Identification of the coefficient for labor requires that the residual variation in labor, after conditioning on k_{it}, m_{it}, and t is non-zero:

$$E[(l_{it} - E[l_{it} \mid k_{it}, m_{it}, t])^2] > 0 \tag{4.3.1}$$

If this condition fails, then l_{it} contains no independent variation and identification fails.

To resolve the identification problem, ACF invert the conditional input demand functions to control for unobserved productivity. Starting with the value-added production function:

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + \varepsilon_{it} \tag{4.3.2}$$

ACF assume that investment is chosen in period $t - 1$, and that labor can be chosen before, during, or after a productivity shock. Further, input demand is assumed to be unobservable, and can be chosen before or during

the time at which labor is chosen. Finally, $\tilde{f}_t(k_{it}, l_{it}, \omega_{it})$ is strictly increasing in ω_{it}, which is necessary for the inversion of input demand. The invertibility proof is as follows.

Inverting input demand $\omega_{it} = \tilde{f}_t^{-1}(k_{it}, l_{it}, \omega_{it})$ and substituting into the production function:

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \tilde{f}_t^{-1}(k_{it}, l_{it}, \omega_{it}) + \varepsilon_{it} = \tilde{\phi}(k_{it}, l_{it}, \omega_{it}) + \varepsilon_{it} \tag{4.3.3}$$

The inverted function for productivity is nonparametrically estimated. Therefore, the coefficients for the inputs cannot be identified in the first stage, and must be subsumed into $\tilde{\phi}(k_{it}, l_{it}, \omega_{it}) = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \omega_{it}$, resulting in the following first-stage moment condition:

$$E[\varepsilon_{it} | I_{it}] = E[y_{it} - \Phi_t(k_{it}, l_{it}, m_{it}) | I_{it}] = 0 \tag{4.3.4}$$

Using the estimate of $\Phi_t(k_{it}, l_{it}, m_{it})$ from the first-stage, estimates the production function can be obtained using the following second stage conditional moment condition:

$$\begin{aligned} E[\xi_{it} + \varepsilon_{it} | I_{it-1}] = E[y_{it} - \beta_0 - \beta_k k_{it} - \beta_l l_{it} \\ - g(\Phi_{t-1}(k_{it-1}, l_{it-1}, m_{it-1}) - \beta_0 - \beta_k k_{it-1} - \beta_l l_{it-1} | I_{it-1}] = 0 \end{aligned} \tag{4.3.5}$$

Because identification of the labor coefficient β_l is delayed until the second stage, an additional moment condition is required. Assuming that productivity is unobserved and follows a first-order Markov process $\omega_{it} = \rho\, \omega_{it-1} + \xi_{it}$, the first stage proceeds by nonparametrically estimating the composite function $\Phi(\cdot)$ through a high-order polynomial regression of y_{it} on k_{it}, l_{it}, m_{it}. In the second stage, a set of four moment conditions is used in a GMM framework to identify the three structural parameters $(\beta_0, \beta_k, \beta_l)$ and the persistence parameter ρ.

Estimation in ACF proceeds as follows. Using the following Leontief-derived value-added production function:

$$y_{it} = \beta_0 + K_{it}^{\beta_k} + L_{it}^{\beta_l} + e^{\omega_{it}} + e^{\varepsilon_{it}} \tag{4.3.6}$$

y_{it} is regressed on k_{it}, l_{it}, and m_{it} in a first-stage ordinary least squares regression. In the second stage, estimation is based on the four moment conditions. To reduce the dimensionality of the parameter search in the second stage, ACF concentrate out the intercept β_0 and productivity persistence parameter ρ from the GMM procedure, which reduces to reduce the dimensionality of the GMM optimization and isolate the identification of β_k and β_l. For each candidate pair (β_k, β_l), they construct an estimate of the productivity innovation ξ_{it} by regressing the implied productivity term $\omega_{it} = \Phi_t - \beta_k k_{it} - \beta_l l_{it}$ on its lagged value. The residuals from this regression are then used to define two moment conditions, requiring that ξ_{it} be uncorrelated with k_{it} and l_{it-1}. The parameters β_k and β_l are chosen to minimize the deviation of these sample moments from zero.

The estimate of ρ is given by the slope coefficient of the regression of the implied productivity ω_{it} on its lagged value ω_{it-1}, where each is constructed as:

$$\omega_{it} = \Phi_t(k_{it}, l_{it}, m_{it}) - \beta_k k_{it} - \beta_l l_{it} \quad (4.3.6)$$

$$\omega_{it-1} = \Phi_{t-1}(k_{it-1}, l_{it-1}, m_{it-1}) - \beta_k k_{it-1} - \beta_l l_{it-1} \quad (4.3.7)$$

This regression assumes $\omega_{it} = \beta_0 + \rho\, \omega_{it-1} + \xi_{it}$, and so the slope coefficient on ω_{it-1} corresponds to the estimate of ρ, while the residuals ξ_{it} are used to construct the moment conditions for GMM.

To implement the ACF model, I follow the standard two-stage estimation procedure but replace the nonparametric inversion of the input demand function with the two-stage Gaussian copula (2SCOPE) approach proposed by Yang et. al. (2022). This method offers several advantages. First, 2SCOPE does not require an instrumental variable (IV) to control for endogeneity, which avoids the restrictiveness of IV assumptions. Because "2SCOPE requires neither IVs nor the assumption of exclusion restriction", the assumptions of 2SCOPE are relaxed, in general, relative to traditional copula-based endogeneity correction methods (Yang et. al., 2022). Second, copula endogeneity correction is flexible in terms of controlling for endogenous parameters, and is particularly useful for handling endogeneity associated with unobserved shocks to an unobserved variable such as productivity. Therefore, replacing the nonparametric inversion with 2SCOPE allows the ACF estimator to remain consistent even when endogenous regressors are normally distributed and exogenous variables are correlated with the control function. This improves identification and addresses the functional dependence issues highlighted in Olley-Pakes and Levinsohn-Petrin.

In the rest of this section, I explain how I replace the nonparametric inversion of the input demand function with 2SCOPE and present the results from a regression using this adapted production function.

According to Sklar's theorem, the joint cumulative distribution function (CDF) of the endogenous regressor m_{it} (materials input) and the structural error ξ_{it} (the innovation in productivity) can be decomposed into their marginal CDFs and a copula function:

$$F(m_{it}, \xi_{it}) = C\big(H(m_{it}), G(\xi_{it})\big) = C\big(U_m, U_\xi\big) \quad (4.3.7)$$

where $U_{m_{it}} = H(m_{it})$ and $U_{\xi_{it}} = G(\xi_{it})$ are uniformly distributed. This decomposition enables the dependence between m_{it} and ξ_{it} to be modeled flexibly, without requiring functional invertibility. This dependence structure can be captured using the Gaussian copula from Park and Gupta (2012):

$$F(m_{it}, \xi_{it}) = C(U_{m_{it}}, U_{\xi_{it}}) = \Psi_\rho\left(\Phi^{-1}(U_{m_{it}}), \Phi^{-1}(U_{\xi_{it}})\right) \quad (4.3.8)$$
$$= \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} \int_{-\infty}^{\phi^{-1}(U_{m_{it}})} \int_{-\infty}^{\phi^{-1}(U_\xi)} exp\left[\frac{-(s^2 - 2\rho * 2 * t + t^2}{2(1-\rho^2)}\right] dsdt$$

where $\Phi^{-1}(\cdot)$ is the inverse standard normal distribution and Ψ_ρ is the bivariate normal CDF with correlation ρ. The copula parameter ρ directly quantifies the endogeneity of m_{it}; a non-zero ρ indicates dependence between the input and the unobserved shock. Under the assumption that $\xi_{it} \sim N(0, \sigma_\xi^2)$, Park and Gupta show that the structural error can be decomposed as:

$$\xi_t = \sigma_\xi \xi_t^* = \sigma_\xi \rho m_{it}^* + \sigma_\xi \sqrt{1-\rho^2}\, \omega t \tag{4.3.9}$$

where $\sigma_\xi \rho m_{it}^*$ captures the correlation between ξ_t and the endogenous regressor, and $\sigma_\xi \sqrt{1-\rho^2}\, \omega t$ is a new independent error term. Substituting this into the production function leads to:

$$Y_{it} = \mu + m_{it}\alpha + W_t\beta + \sigma_\xi \rho m_{it}^* + \sigma_\xi \sqrt{1-\rho^2}\, \omega t \tag{4.3.10}$$

In this framework, I fit a Gaussian copula to the joint distribution of materials (i.e., cost of goods sold) and the residuals from a first-stage output regression. The resulting scalar control function summarizes the dependence between input choices and productivity shocks, and is included in the second-stage production function regression to correct for endogeneity and consistently identify the coefficients on capital and labor.

	Stage 1	**Stage 2**
const	5.9871***	-194.3558***
Capital Measure 2	0.0882*** (0.003)	0.0986*** (2.886)
Total Employment	-3.6368*** (0.269)	2.4721** (1.061)
Cost of Goods Sold	1.0131*** (0.004)	-
Control Term	-	388.4511*** (6.525)
R-Squared	0.923	0.919
F-Stat	6.377e+04	2331
Number of Observations	15,952	15,952
Durbin-Watson	0.261	0.255

Table 4.3.1: ACF Regression Results

Table 4.3.1 presents the two-stage regression results for the ACF model with copula endogeneity correction. In the first stage, the coefficient for cost of goods sold (1.0131) being near-one coefficient suggests that COGS is very tightly related to output, making it a strong proxy. Further, the negative coefficient for the estimate of labor indicates functional dependence problems, as the labor input might be too closely related to the nonparametric control function to separately identify its coefficient at this stage.

In the second stage, the coefficient for labor reverses signs from Stage 1 and is less precise, indicating that, relative to the OP model, the ACF model better isolates the causal impact of labor on output and does a better job of correcting for simultaneity in the measure of labor. In addition, while the estimate for capital is still low, the coefficient remains stable across OP and ACF models, supporting the robustness of its estimated contribution. Finally, the coefficient on the control term (388.4511) is highly significant, suggesting that a substantial portion of output variation is driven by unobserved productivity. This strengthens the case for the use of copula endogeneity correction, as input choices appear to be highly endogenous.

5. Discussion

Table 5.1 includes a comparison of the regression results for the 4 identification models.

	OLS	**FE**	**OP Stage 1**	**OP Stage 2**	**ACF Stage 1**	**ACF Stage 2**
const	-43.21***	0.78***	-2.8301***	7.09***	5.99***	-194.36***
Capital	0.50*** (0.00)	0.10*** (0.015)	-	0.09*** (0.01)	0.09*** (0.003)	0.099*** (2.89)
Labor	27.95*** (0.47)	10.43*** (1.68)	3.44 (2.59)	-	-3.64*** (0.27)	2.47** (1.06)
R-Squared	0.68	0.32	0.88	0.88	0.92	0.92
F-Stat	18,572.33	9.75	8,747	13,480	63,770	2,331
Num Obs.	17,840	17,838	10,700	9,979	15,952	15,952

Table 5.1: Comparing Regression Results for the Identification Models

6. Limitations

Our research faces several limitations that affect the interpretation and generalizability of the findings. First, data heterogeneity and standardization issues pose significant challenges. Firms report capital investments—such as property, plant, and equipment (PPE), research and development (R&D), and intangible assets—differently, making cross-company comparisons difficult. Employment metrics also vary, with some firms reporting only full-time employees while others include contractors, requiring assumptions that may introduce measurement error.

Endogeneity and model specification issues further complicate the study. While fixed effects control for unobserved firm-level heterogeneity, they cannot fully address reverse causality—for example, whether productivity gains drive workforce reductions or whether layoffs force efficiency improvements. Omitted

variable bias is another concern, as unmeasured factors like management quality or industry-specific shocks may influence the results. Multicollinearity among predictors, such as capital expenditures and R&D spending, inflates the variance in coefficient estimates, making it harder to isolate individual effects.

Selection bias is another critical limitation. By focusing only on surviving firms, I may overstate productivity gains, as failed companies—which might have experienced declining efficiency—are excluded. The study's specific time frame (e.g., post-2010) also limits generalizability, as technological and economic conditions vary across different eras. Furthermore, measurement error in key variables, such as employment (which may not account for outsourcing or automation) and capital stock (especially intangible assets like software and patents), introduce noise into the analysis.

Finally, the reliance on observational data means I cannot make definitive causal claims. Unobserved technological shifts, such as rapid AI adoption, may confound the relationship between workforce changes and productivity. Despite these limitations, I mitigate some concerns by using robust standard errors and lagged variables. However, the findings should be interpreted as suggestive rather than conclusive, highlighting the need for future research with richer datasets or quasi-experimental methods to strengthen causal inference.

APPENDIX A: Estimating Quarterly Employment Data

Because quarterly financial statement data from Compustat Capital-IQ does not include data for the number of employees for any company, I adapted code created by McBride, Ellis (2024) to estimate quarterly data for the number of employees based on annual data using a QRF. I chose to use a quantile regression forest as our data is highly dimensional and skewed, leading to poor estimates of employment for very large companies when using less advanced regression techniques, such as multiple linear regression.

To be precise, I ran three preliminary multiple linear regression models to estimate annual employment:

$$\begin{aligned} &Employment_{it} \\ &= \beta_0 + \beta_1 Receivables\ (Trade)_{it} \\ &+ \beta_2 \text{Liabilities and Stockholders Equity (Total)}_{\text{it}} \\ &+ \beta_3 Stock\ \text{Compensation Expense}_{it} \end{aligned} \tag{A.1}$$

$$\begin{aligned} &Employment_{it} = \beta_0 + \beta_1 \text{Receivables (Total)}_{\text{it}} \\ &+ \beta_2 Sales/\text{Turnover (Net)}_{\text{it}} \\ &+ \beta_3 \text{Selling, General and Administrative Expenses}_{\text{it}} \end{aligned} \tag{A.2}$$

$$\begin{aligned} &Employment_{it} = \beta_0 + \beta_1 Cash_{it} \\ &+ \beta_2 \text{Inventories (Total)}_{\text{it}} \\ &+ \beta_3 \text{Common/Ordinary Equity (Total)}_{\text{it}} \end{aligned} \tag{A.3}$$

Several problems invalidate the results of these regression models. First, the model suffers from heteroscedasticity, as indicated by the funnel shape of the residual plots. In addition to heteroscedasticity, the variance inflation factor for Liabilities and Stockholders' Equity – Total, Receivables – Total, Sales/Turnover (Net), and Selling, General and Administrative Expenses are greater than 5, which indicates the presence of multicollinearity in Regression Models 1 and 2. Further, the curvature of the residuals indicates that the relation between employment and the independent (predictor) variables is not linear for any of the three models.

	Model 1	**Model 2**	**Model 3**
Constant	-1.50***	-1.56***	-0.99***
Receivables – Trade	0.35***		
Stock Compensation Expense	-0.14***		
Liabilities and Stockholders' Equity – Total	0.29***		
Receivables – Total		0.21***	
Sales/Turnover (Net)		0.46***	

Selling, General and Administrative Expenses		-0.14***	
Cash			0.15***
Inventories – Total			0.10***
Common/Ordinary Equity – Total			0.26***

R-squared	0.77	0.80	0.63
R-squared Adj.	0.77	0.80	0.63
F-stat	4370.50	5100.92	2048.06
F p-value	0.000	0.000	0.000
Observations	3872	3874	3634
SSR	917.10	813.09	1426.75
ESS	3108.74	3215.11	2414.92
MSE Res	0.24	0.21	0.39

Table 1: Summary Table for Employment – Multiple Linear Regression

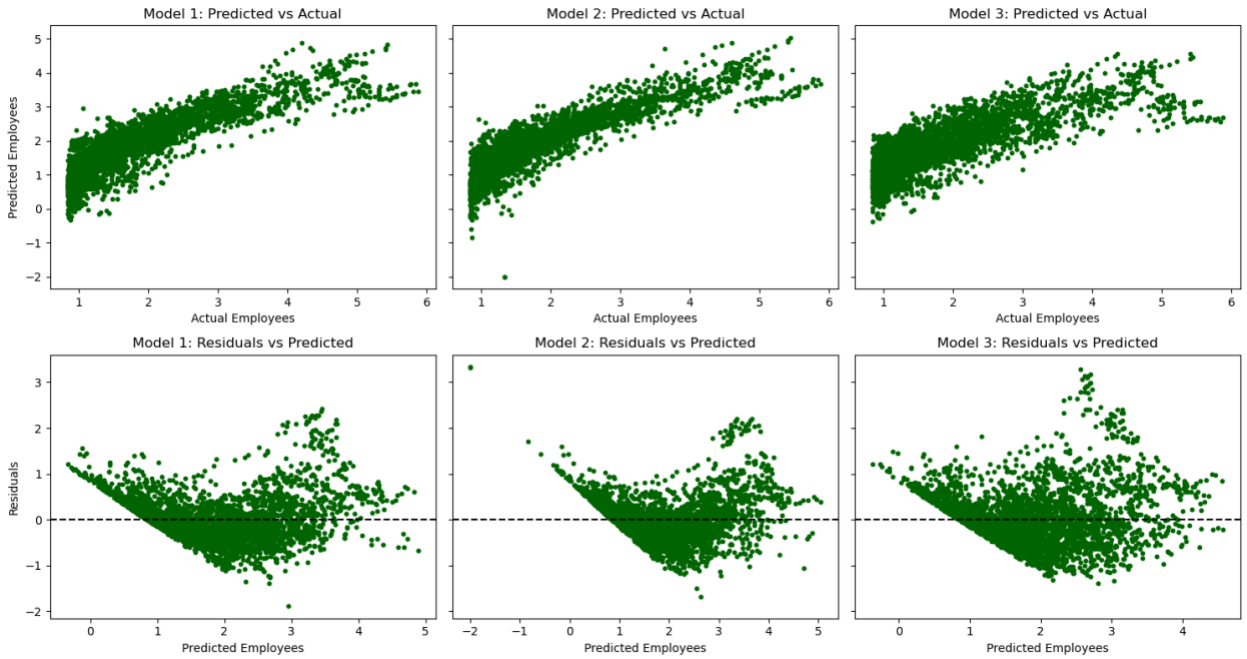

Figure A1: Predicted vs Actual values for Employment – Multiple Linear Regression

Figure 2: Residual Plot for Employment – Multiple Linear Regression

By training a QRF model to estimate quarterly employment, our model will be more accurate for firms across the size spectrum, including those at the upper and lower tails of the distribution. Unlike traditional linear regressions, which only model the conditional mean and assume constant variance, QRFs estimate the full conditional distribution of employment. This allows for more robust predictions, particularly for firms with atypical financial structures or outlier characteristics. Additionally, QRFs naturally handle non-linear relationships and multicollinearity without the need for manual feature engineering or transformation, making them well-suited for high-dimensional and skewed financial datasets such as ours.

To further improve interval validity, I apply the conformal quantile regression (CQR) method proposed by Romano et al. (2019), which uses a separate calibration set to adjust the estimated quantile intervals to achieve finite-sample coverage guarantees. Specifically, I compute conformity scores based on the maximum deviation between the observed values and the predicted interval bounds and inflate the original intervals accordingly. This adjustment ensures that the final prediction intervals attain the desired marginal coverage (e.g., 90%) regardless of the underlying distribution or model miscalibration.

To capture heterogeneity in prediction accuracy across the employment distribution, I train separate QRF models for each decile of annual employment, allowing for more precise estimation in the tails. The resulting models outperform traditional linear regressions in both accuracy and reliability, particularly for large firms where linear methods tend to underestimate employment due to non-linearity and multicollinearity among predictors.

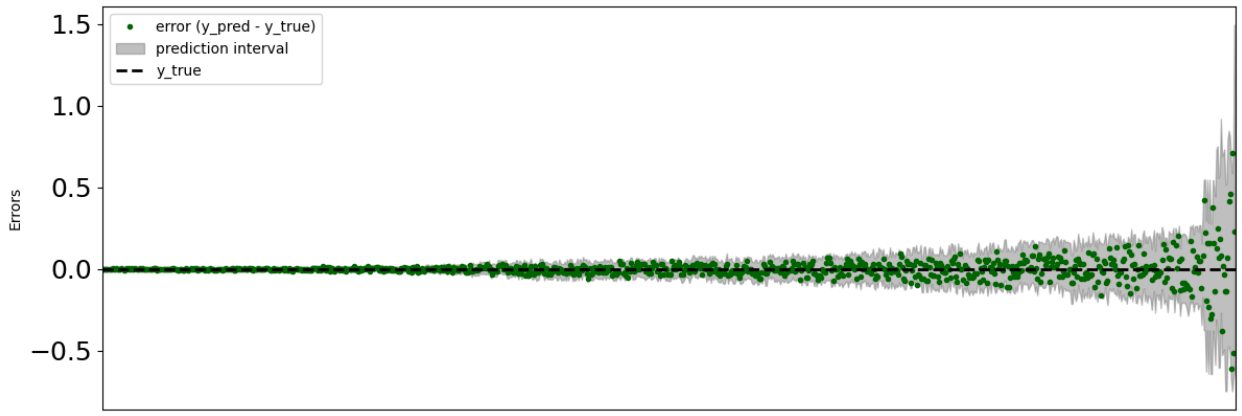

Figure A2: Error Plot Sorted by Prediction Interval Width

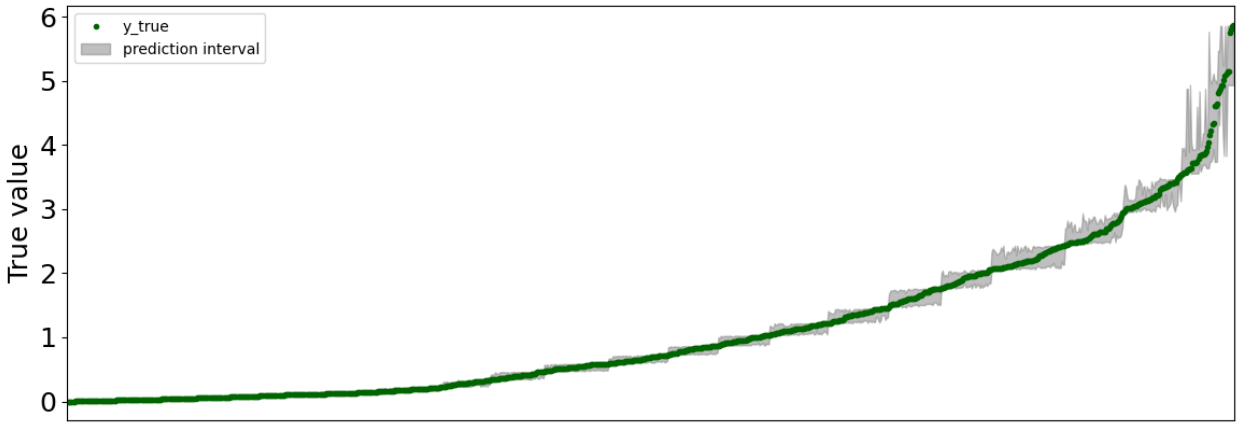

Figure A3: Prediction Intervals Across the Employment Distribution

Figure A2 displays the prediction errors (i.e., predicted minus true employment) alongside conformal prediction intervals sorted by interval width. The error points remain centered around zero, indicating that the model is unbiased across most of the distribution. However, the width of the prediction intervals increases substantially in the right tail, suggesting greater uncertainty in employment predictions for larger firms. This aligns with the observed heteroskedasticity in the earlier residual plots and highlights the model's adaptive behavior of generating wider intervals where prediction risk is higher to maintain valid coverage.

Figure A3 shows the true number of employees sorted in ascending order, overlaid with their corresponding prediction intervals. The conformal intervals remain tight and consistent throughout the lower and middle ranges of the distribution but expand considerably in the upper tail. This expansion illustrates how the model responds to increasing variance and uncertainty associated with large firms — a region where simpler models like OLS typically fail to generalize. The widening intervals provide meaningful uncertainty quantification while maintaining full coverage, albeit at the cost of wider intervals in the extremes.

The conformal prediction intervals produced by our Quantile Regression Forests achieve full empirical coverage, with 100% of ground truth employment values falling within their respective intervals. However, the resulting Mean Winkler Interval Score (MWIS) of 54.624 suggests that the intervals are relatively wide. This reflects a tradeoff between **validity** (coverage) and **efficiency** (interval tightness), which is inherent in conformal methods. While our intervals may be conservative, they ensure robustness across the entire conditional distribution, particularly in the presence of heteroskedasticity and distributional shift.

To assess the efficiency of the conformal prediction intervals produced by our QRF model, I examined the distribution of interval widths. The median interval width was 0.74, indicating that for most firms, the prediction intervals were relatively tight. However, the 90th percentile interval width increased to 11.03, the 95th percentile to 98.02, and the 99th percentile to 238.20. This widespread reflects the model's adaptive response to varying levels of uncertainty across firms.

As shown in Table A2, intervals remain narrow for most firms but expand dramatically in the upper tail of the employment distribution — particularly for firms with over 100 employees. These larger intervals preserve 100% empirical coverage, but at the cost of wider intervals in regions with higher variance or fewer comparable observations. This tradeoff is captured by the Mean Winkler Interval Score (MWIS) of 54.6, which balances interval width and accuracy. While conservative, these intervals offer robust uncertainty quantification in the presence of heteroskedasticity and outliers.

Percentile	Interval Width
0.10	0.042610
0.25	0.178450
0.50	0.735300
0.75	3.260725
0.90	11.034000
0.95	98.020000
0.99	238.200000

To evaluate whether quarterly data was drawn from the same distribution as the annual training data, I trained a binary Random Forest Classifier to distinguish between the two. The classifier achieved a cross-validated accuracy of 99.99%, indicating a strong distributional shift between the datasets. This finding suggests that the financial feature patterns in the quarterly data differ substantially from those in the annual data, which may limit the generalizability of the QRF model and introduce bias in predicted employment estimates.

Although a classifier accuracy of 99.99% is unusually high, this result does not indicate a problematic distributional shift in the context of the study. The classifier was designed to distinguish between quarterly observations based on the firm-year to which they belonged, using financial statement variables as predictors. However, since quarterly employment figures were constructed by interpolating annual employment data, often using financial variables that are smooth and seasonally stable, the classifier was effectively distinguishing between pre-engineered groupings with highly predictable structure. In other words, the classification task was not learning patterns from independently varying labels but from a target constructed directly from financial inputs. Therefore, the near-perfect accuracy reflects the strong internal consistency of the data rather than overfitting or leakage. Since the primary goal was to estimate conditional distributions of employment within this fixed structure, not to generalize to new or temporally distant contexts, the high classifier performance does not compromise the validity of the quantile regression forest estimates.

APPENDIX B: Cobb-Douglass Production Function

Throughout the paper, I used a Cobb-Douglass production function in our regression analyses. In this Appendix, I assess whether the assumptions of the Cobb-Douglass production function hold given the results from the ordinary least squares regression in Table 3.2.

The Cobb-Douglass production function makes the following assumptions:

1. Constant Returns to Scale: $Y = zF(\alpha K, \alpha L)$
2. Positive Marginal Product of Capital and Labor.
3. Diminishing Returns to Capital and Labor.

The positive coefficients for capital (0.81 and 0.50) and labor (5.46 and 27.95) in both models shows that the marginal product of capital and the marginal product of labor is positive. Summary statistics for the marginal products of capital and labor support the theoretical assumption of diminishing marginal returns. The mean marginal product of capital is 0.46, with a maximum of 2.70 and a minimum of 0.06. Similarly, the mean marginal product of labor is 0.61, with a maximum of 3.30 and a minimum of 0.17. The range between the mean and maximum values, combined with the declining trend observed in firm-level MPK and MPN, suggests that as firms accumulate more capital or labor, the unit increase in output from an additional unit of capital or labor decreases. Finally, based on a scatterplot of the relation between capital and output and labor and output, capital and labor appear to exhibit constant returns to scale.

Variable	Mean (in millions)	St. Dev. (in millions)	Min (in millions)	Max (in millions)
Marginal Product of Capital	0.41	0.28	0.0002	21.02
Marginal Product of Labor	3.24	13.52	0.68	630.49

Table A1: Summary Statistics for the Marginal Product of Capital and Labor

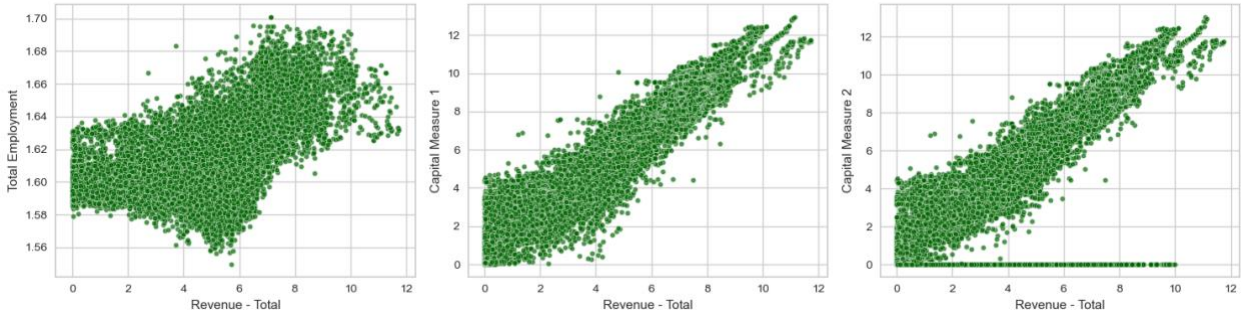

Table A2: Returns to Scale of Capital and Labor