HOME ASSIGNMENT

Empirical Industrial Organisation — ECONM0013

Student Number: **1864173**

Table of Contents:

Table I: HA_Data Variable Definitions

Var. Name	**Description**
firm	Firm ID i
year	Time period t
L	Log of labour l_{it}
I	Log of investment i_{it}
K	Log of capital k_{it}
A	Age of firm a_{it}
X	Continuation Dummy
Y	Log of output y_{it}

[1.A, 5 Points]. *Report sample statistics (number of observations, mean, median, standard deviation, etc.) for the key variables in the data (yit, lit, kit, iit, and ait) for the full sample, the balanced sub-panel (i.e., those firms that are present in all years), and the exiters (i.e., those firms that are not present in all years). Do these statistics seem different? What do these differences tell you about the types of firms that tend to survive versus those that exit?*

Table 1: Full Sample

Variable	Obs	Mean	Std. dev.	Min	Max
y_{it}	39,569	22.66	3.11	8.96	33.15
l_{it}	39,569	5.03	1.00	0.39	8.88
k_{it}	39,569	8.99	1.87	2.09	14.57
i_{it}	39,569	5.03	1.00	1.13	9.34
a_{it}	39,569	8.54	3.21	1.00	17.00

Table 2: Balanced Sub-panel

Variable	Obs	Mean	Std. dev.	Min	Max
y_{it}	21,800	22.57	3.10	8.96	33.15
l_{it}	21,800	5.01	1.00	0.39	8.88
k_{it}	21,800	9.16	1.80	2.24	14.34
i_{it}	21,800	5.04	1.00	1.13	9.34
a_{it}	21,800	7.32	3.23	1.00	16.00

Table 3: Exiting Firms

Variable	Obs	Mean	Std. dev.	Min	Max
y_{it}	17,769	22.77	3.11	10.63	32.80
l_{it}	17,769	5.05	0.99	1.42	8.86
k_{it}	17,769	8.78	1.93	2.09	14.57
i_{it}	17,769	5.02	0.99	1.34	8.87
a_{it}	17,769	10.03	2.47	1.00	17.00

Tables 1-3 reveal notable differences between exiting and surviving firms:

Log Capital ($\boldsymbol{k_{it}}$): Exiting firms have lower mean log capital (8.78) relative to surviving firms (9.16), with slightly higher variation (std. dev. 1.93 vs. 1.80). Since these are log values, this represents approximately 38% less capital ($e^{(9.16-8.78)} \approx 1.38$) for exiting firms in real terms.

Firm Age ($\boldsymbol{a_{it}}$): Exiting firms have a substantially higher mean age (10.03) relative to surviving firms (7.32), though with lower variation (std. dev. 2.47 vs. 3.23)

Log Output ($\boldsymbol{y_{it}}$): Exiting firms have a higher average log output (22.77) relative to surviving firms (22.57) such that they have approximately 22% higher output ($e^{(22.77-22.57)} \approx 1.22$) than surviving firms in real terms.

Log Labour (l_{it}) & log Investment (i_{it}): Surviving and exiting firms show minimal differences in labour and investment. Exiting firms have approximately a 4% bigger labour force ($e^{(5.05-5.01)} \approx 1.04$) and 2% less investment ($e^{(5.02-5.04)} \approx 0.98$) relative to surviving firms in real terms. These smaller differences suggest that they are less predictive of survival relative to the other variables.

These statistical differences suggest several important characteristics about firm survival dynamics:

- **Capital Investment Advantage**: Surviving firms tend to have higher capital, suggesting that productive capacity and economies of scale are crucial for longevity.
- **Age Paradox**: The higher average age of exiting firms challenges the idea that older firms are more established and resilient. This may be reflective of older firms being complacent and less innovative such that they are less adaptive to changing market conditions relative to younger firms.
- **Efficiency Trade-Off**: Exiting firms show higher output with lower capital, implying that they might operate with higher, but potentially unsustainable capital utilisation rates.

Overall, capital investment appears more critical for survival than firm age, as older firms may struggle to maintain market presence despite their experience.

[1.B, 10 Points]. *Using only the balanced sub-panel compute the total, between, within, and random effects estimators for equation (1). How are they different? Perform a Hausman test of random effects versus fixed effects (i.e., within estimator). What have you learned about firm heterogeneity and about possible measurement error from these results?*

The production function (PF) we wish to estimate is:

$$\textbf{(PF)} \quad \boldsymbol{y_{it} = \alpha_0 + \alpha_A a_{it} + \alpha_L l_{it} + \alpha_K k_{it} + \omega_{it} + e_{it}} \qquad \textbf{(1)}$$

where y_{it} is log of output
a_{it} is age
l_{it} is log of labour
k_{it} is log of capital
ω_{it} is the productivity shock at time t

Using the balanced sub-panel, I estimated equation (1) employing the four estimators which differ in how they treat unobserved firm heterogeneity. In the context of **PF(1)**, firm heterogeneity means that even when the same measured inputs (labour, capital) are used, some firms consistently produce more output than others due to these unobserved firm-specific effects that influence productivity. While robust standard errors would typically be used to account for potential heteroskedasticity and autocorrelation, they are not used here as the Hausman test requires non-robust standard errors and as robust errors are also not typically applied to the between estimator due to the lack of within-panel variation.

Looking at the results in Table 4, we see that Coefficients on age (a_{it}), log of labour (L), and log of capital (K) remain positive and significant across all estimators. However, differences in magnitudes and the constant term across estimators reveal important aspects of firm heterogeneity.

Table 4: Estimates of PF (1)

Variable	Total	Between	Fixed	Random
a_{it}	0.207***	0.217***	0.206***	0.207***
	(0.003)	(0.006)	(0.004)	(0.003)
l_{it}	1.265***	1.519***	1.238***	1.260***
	(0.008)	(0.028)	(0.008)	(0.008)
k_{it}	1.172***	1.266***	1.162***	1.170***
	(0.005)	(0.015)	(0.006)	(0.005)
constant	3.981***	1.773***	4.209***	4.024***
	(0.057)	(0.200)	(0.06)	(0.057)
N	21,800	21,800	21,800	21,800
R^2	0.858	0.848	0.861	

Legend: * p<0.05; ** p<0.01; *** p<0.001

Differences Across Estimators

Total (Pooled OLS) Estimator:

$$y_{it} = \alpha_0 + \alpha_A a_{it} + \alpha_L l_{it} + \alpha_K k_{it} + (\omega_{it} + e_{it})$$

This estimator ignores firm heterogeneity by treating all firm-year observations as independent. The coefficient estimates (0.207 for a_{it}, 1.265 for l_{it}, and 1.172 for k_{it}) are statistically significant, but are likely biased if unobserved firm characteristics correlate with the inputs.

Between Estimator:

$$\bar{y}_i = \alpha_0 + \alpha_A \bar{a}_i + \alpha_L \bar{l}_i + \alpha_K \bar{k}_i + \bar{\omega}_i + \bar{e}_i$$

By using firm level averages, this estimator captures cross-sectional variation. The coefficients, especially for labour (1.519) and capital (1.266), are noticeably higher compared to the FE estimates. The lower intercept (1.773) relative to the others (average around 4.0) indicates that much of the overall output level is explained by between-firm differences in unobserved factors.

Fixed Effects (Within) Estimator:

$$(y_{it} - \bar{y}_i) = \alpha_A(a_{it} - \bar{a}_i) + \alpha_L(l_{it} - \bar{l}_i) + \alpha_K(k_{it} - \bar{k}_i) + (\omega_{it} - \bar{\omega}_i) + (e_{it} - \bar{e}_i)$$

This method removes time-invariant firm heterogeneity through demeaning. Its coefficients (0.206 for a_{it}, 1.238 for l_{it}, and 1.162 for k_{it}) are similar to the total estimator, suggesting that within-firm dynamics are robust even after removing persistent firm traits.

Random Effects Estimator:

$$(y_{it} - \theta\bar{y}_i) = \alpha_A(a_{it} - \theta\bar{a}_i) + \alpha_L(l_{it} - \theta\bar{l}_i) + \alpha_K(k_{it} - \theta\bar{k}_i) + (\omega_{it} - \theta\bar{\omega}_i) + (e_{it} - \theta\bar{e}_i)$$

The RE estimates (0.207 for a_{it}, 1.260 for l_{it}, and 1.170 for k_{it}) are similar to the total and FE estimates, as this estimator combines within- and between-firm variation. However, it relies on the assumption that unobserved firm effects are uncorrelated with the regressors.

Hausman Test and Implications:

The Hausman test results in Table 5contrasts the FE and RE estimates. The small differences between the FE and RE coefficients (e.g., a difference of -0.001 for a_{it} and -0.022 for l_{it}) are statistically significant as indicated by a Chi-Squared statistic of 154.01 (df = 3, p = 0.0000). This means we reject the null hypothesis that firm-specific effects are uncorrelated with the regressors such that the random effects estimator is not consistent and appropriate.

Table 5: Hausman Test Results

	---Coefficients---			
Variable	(b) Fixed Effect	(B) Random Effect	(b-B) Difference	$\sqrt{\text{diag}(V_b - V_B)}$
a_{it}	0.206	0.207	-0.001	0.002
l_{it}	1.238	1.260	-0.022	0.002
k_{it}	1.162	1.170	-0.008	0.003

Test Summary	Chi-Squared Statistic	Chi-Squared d.f.	Prob. > Chi-Squared
Cross Section Random	154.01	3	0.0000

Conclusions:

Firm Heterogeneity: The rejection of the RE model via the Hausman test confirms that unobserved, time-invariant firm characteristics (such as management style or work culture) significantly influence productivity. This supports the use of FE estimation when such heterogeneity is present.

Measurement Error: While statistically significant, the relatively small magnitude of differences between fixed and random effects estimators in Table 5 (differences of -0.001 for a_{it}, -0.022 for l_{it}, and -0.008 for l_{it}) suggests that measurement error is present but not severely biasing estimates. However, differences in the constant terms and the between estimates hint that cross-sectional measurement issues may still impact interpretation.

[1.C, 10 Points]. *Using the balanced sub-panel, compute difference estimators in which you take differences over t of both sides of equation (1). Report results from estimates of the first (i.e., 1 year) differenced model, second (i.e., 2 years) differenced model and third (i.e., 3 years) differenced model. What do these tell you about measurement error? Base your discussion on Golsbee (2000, NBER).*

When we apply differencing to the production function:

$$\textbf{(PF)} \quad \boldsymbol{y_{it} = \alpha_0 + \alpha_A a_{it} + \alpha_L l_{it} + \alpha_K k_{it} + \omega_{it} + e_{it}} \qquad \textbf{(1)}$$

We eliminate the constant term and transform the equation into differences over time periods (1-year, 2-year, and 3-year differences):

$$\Delta_s \boldsymbol{y}_{it} = \boldsymbol{y}_{it} - \boldsymbol{y}_{it-s} = \boldsymbol{\alpha}_A \Delta_s \boldsymbol{a}_{it} + \boldsymbol{\alpha}_L \Delta_s \boldsymbol{l}_{it} + \boldsymbol{\alpha}_K \Delta_s \boldsymbol{k}_{it} + \Delta_s \boldsymbol{\omega}_{it} + \Delta_s \boldsymbol{e}_{it} \quad \textbf{(2)}$$

where the s, the subscript on Δ_s, is the length of the differencing

This transformation helps address endogeneity and exposes patterns related to measurement error.

Empirical Analysis:

Using the balanced sub-panel, I estimated the production function using first, second, and third-year differenced models as shown in Table 6.

In Table 6, the $\Delta_s l_{it}$ coefficients increase from 1.184 to 1.224 and the $\Delta_s k_{it}$ coefficients increase from 1.056 to 1.157, suggesting measurement error in these inputs. This pattern suggests that the 1-year estimates are attenuated due to measurement error in l_{it} and k_{it}, and that longer differences reduce this bias, bringing the estimate closer to the true $\alpha_L$and α_K. This aligns with Goolsbee (2000)'s tax term findings.

In differenced models age simply becomes a constant ($\Delta_s a_{it} = s$) for all firms, while labour and capital retain their meaningful variation across different differencing intervals, allowing their true production relationships to emerge as measurement error is averaged out over longer periods. This aligns with the $\Delta_s a_{it}$ coefficients decreasing (0.303 to 0.203) as the differencing period lengthens which is likely reflective of a real economic effect where age's impact on output weakens over longer horizons rather than bias.

Table 6: Difference Estimator Results

Variable	1-year Differenced Model	2-year Differenced Model	3-year Differenced Model
$\Delta_s a_{it}$	0.303*** (0.012)	0.223*** (0.007)	0.203*** (0.005)
$\Delta_s l_{it}$	1.184*** (0.008)	1.222*** (0.008)	1.224*** (0.009)
$\Delta_s k_{it}$	1.056*** (0.008)	1.132*** (0.007)	1.157*** (0.007)
N	19,620	17,440	15,260
R^2	0.701	0.782	0.823

Legend: * p<0.05; ** p<0.01; *** p<0.001

Measurement Error and Model Fit:

The increase in R² from 0.701 to 0.823 and the decline in standard errors for labour and capital suggest that longer differencing intervals reduce transitory noise and improve model precision. This supports Goolsbee (2000), who argues that longer differences average out temporary measurement errors, enhancing the signal-to-noise ratio and mitigating attenuation bias. As Griliches and Hausman (1986) note, first-differencing can exacerbate bias by amplifying the impact of noise while 2 or 3-year differences dilute it by increasing the variance of the true signal relative to measurement error.

Conclusion:

The consistent increase in the magnitude of input coefficients and the improved model fit as the differencing interval lengthens suggest the presence of significant measurement error in the original variables. First differences exacerbate this issue, while longer differences recover more accurate estimates of $\alpha_L$and α_K highlighting the importance of accounting for measurement error in production functions.

[D, 10 Points]. The following two questions try to measure the importance of endogenous exit and sample selection:

1. *Using the full (unbalanced) panel, compute the pooled and fixed-effect estimators. How do these estimates compare to the pooled and fixed-effect estimates on the balanced panel? What does this tell you about the possible effects of selection in this dataset?*

The pooled and fixed-effect estimators for the balanced and unbalanced panel are shown in Table 7. The estimate coefficients for labour and capital are nearly identical in both the pooled unbalanced and balanced panels, which indicates robustness in these coefficients. However, the age coefficient is slightly lower in the balanced panel. This discrepancy is notable because, as we have already observed in 1.A, exiting firms tend to be older (mean age 10.03) than surviving ones (mean age 7.32) with less variation.

Table 7: Balanced & Unbalanced Pooled & FE Estimates

Variable	Pooled Unbalanced	FE Unbalanced	Pooled Balanced	FE Balanced
a_{it}	0.219*** (0.002)	0.208*** (0.003)	0.207*** (0.003)	0.206*** (0.004)
l_{it}	1.264*** (0.006)	1.233*** (0.006)	1.265*** (0.008)	1.238*** (0.008)
k_{it}	1.155*** (0.004)	1.154*** (0.005)	1.172*** (0.006)	1.162*** (0.006)
constant	4.049*** (0.044)	4.311*** (0.046)	3.981*** (0.061)	4.209*** (0.064)
N	39,569	39,569	21,800	21,800
R²	0.858	0.861	0.858	0.861

Consequently, the balanced panel systematically underrepresents older, exiting firms. This selection bias is likely to lead to an understated effect of firm age on output in the balanced sample relative to the full unbalanced panel.

Overall, while the labour and capital coefficients remain stable, the observed differences in the age coefficient signal that relying solely on the balanced panel might distort our understanding of how firm age influences production, emphasising the importance of using the full unbalanced dataset to capture the complete dynamics. Using Olley & Pakes (1996) CFA framework to control for endogenous exit would mitigate selection bias in the balanced panel, providing more reliable estimates of the production function parameters and better capturing the full dynamics of firm behaviour.

2. *Use a Probit model to estimate the probability that a firm exits in period t + 1 as a function of iit, ait, and kit. (Variable X in the dataset is zero in t if the firm exits in t + 1.) Compute the implied inverse mills ratio (as in a standard endogenous sample selection model) and include it as a regressor in both your pooled and fixed effect regressions above. Does this appear to correct for selection bias?*

To account for potential selection bias from firm exit, I began by estimating a probit regression to estimate the probability that a firm exits the market in period $t + 1$ as a function of i_{it}, a_{it} and k_{it}:

$$\boldsymbol{P(X_{it} = 1 | i_{it}, a_{it}, k_{it}) = \Phi(\beta_0 + \beta_1 a_{it} + \beta_2 i_{it} + \beta_3 k_{it})} \quad \textbf{(3)}$$

Where the dependent variable X is zero in period t if the firm exits in $t+1$ *(*$X_{it}=0$*)*

From this we can predict the probit score (z_{it}) which is the linear combination of the explanatory variables and their estimated coefficients for each firm at time *t*:

$$\boldsymbol{z_{it} = \hat{\beta}_1 a_{it} + \hat{\beta}_2 i_{it} + \hat{\beta}_3 k_{it}} \quad \textbf{(4)}$$

We can then use z_{it} to calculate the inverse mills ratio (IMR) λ_{it} for the firms that survived into period $t+1$ which captures the hazard of exit conditional on survival:

$$\boldsymbol{\lambda_{it} = \frac{\phi(z_{it})}{1-\Phi(z_{it})}} \quad \textbf{(5)}$$

z_{it} is the predicted probit score
$\phi(z_{it})$ is the PDF of the standard normal distribution of z_{it}
$\Phi(z_{it})$ is the predicted probability of exiting (standard normal CDF)
$1-\Phi(z_{it})$ is the predicted probability of survival

We can then include the IMR (λ_{it}) in our original production function:

$$\boldsymbol{y_{it} = \alpha_0 + \alpha_A a_{it} + \alpha_L l_{it} + \alpha_K k_{it} + \delta\lambda_{it} + \omega_{it} + e_{it}} \quad \textbf{(6)}$$

If the coefficient (δ) on the IMR is statistically significant, this is suggestive that selection bias is present and the IMR is helping correct for it.

Results Analysis

Table 8 shows the pooled and FE estimates of the unbalanced panel both with and without the IMR (λ_{it}) included as a regressor. Inclusion of the IMR causes the age coefficient to increase from 0.219 to 0.231 in the pooled unbalanced panel and to increase from 0.208 to 0.216 in the fixed effects model, implying that the effect of age on output was previously underestimated. The coefficients on labour and capital remain stable across specifications, indicating their robustness.

The statistically significant and negative IMR coefficients (approximately −0.231 in the pooled and −0.202 in the FE models) infer that firms with a higher predicted probability of exit tend to have lower output, conditional on observed inputs. This suggests that exit is non-random and correlated with unobserved productivity.

The upward adjustment in the age coefficient after controlling for selection bias initially appears counterintuitive since surviving firms skew younger than exiting firms. However, this result is consistent with the selection mechanism in our data.

When younger firms are disproportionately represented in the surviving sample, the uncorrected estimates understate the true effect of age on productivity. The IMR correction properly adjusts for this sample selection issue, revealing that age has a stronger positive relationship with productivity than uncorrected estimates suggest.

Several factors might account for this. First, unobserved traits, like high productivity in some young firms, could distort the age-survival link, as only high-performing older firms may endure while less productive ones exit. Second, selection effects could differ by industry or firm size, with the overall pattern hiding these variations. Third, age might interact with other factors influencing both survival and productivity, altering its apparent impact. Lastly, the age-productivity relationship could be non-linear, and therefore, inadequately captured by our model.

The upward adjustment confirms that selection bias from endogenous exit is present, though its mechanism is more complex than a simple age-survival relationship would suggest.

Table 8: Pooled & FE Estimates with/without IMR

Variable	Pooled Unbalanced	FE Unbalanced	Pooled Unbalanced IMR	FE Unbalanced IMR
a_{it}	0.219*** (0.002)	0.208*** (0.003)	0.231*** (0.002)	0.216*** (0.004)
l_{it}	1.264*** (0.006)	1.233*** (0.006)	1.265*** (0.006)	1.265*** (0.006)
k_{it}	1.155*** (0.004)	1.154*** (0.005)	1.154*** (0.004)	1.156*** (0.005)
λ_{it} (IMR)			-0.231*** (0.022)	-0.202*** (0.023)
Constant	4.049*** (0.041)	4.311*** (0.043)	3.984*** (0.043)	4.247*** (0.045)
N	39,569	39,569	37,389	37,389
R²	0.858	0.861	0.858	0.86

Legend: * p<0.05; ** p<0.01; *** p<0.001

[E, 20 Points]. *By following the procedure detailed in problem sets 3 and 4, implement the control function approach by Olley & Pakes [i.e., OP] (1996) (both ignoring and correcting for endogenous exit) for the estimation of equation (1). Note that in problem sets 3 and 4 there was not any ait variable. Consider both ait and kit as state variables and treat them symmetrically: estimate both αA and αK only in the second step of the procedure. When correcting for endogenous exit, estimate the probability of survival as a function of iit, ait, and kit (as in question D above).*

Addressing Simultaneity Bias

In production function models like PF(1), simultaneity bias arises as firms choose inputs after observing their productivity shocks such that higher unobserved productivity can lead firms to invest more in capital or hire more labour. This endogeneity creates a correlation between the input choices and the error term in the production function estimation, resulting in biased coefficient estimates when using standard methods like OLS.

The Olley-Pakes (OP) framework addresses this problem by exploiting the relationship between a firm's unobserved productivity and its investment decisions. By inverting the investment function (ID), the framework constructs a control function that proxies for the productivity shock ω_{it}, which can then be substituted into the production function. This essentially purges the simultaneity bias by isolating the influence of unobserved productivity from the observed inputs, leading to more consistent and reliable estimates of the production function parameters.

OP First Stage: Estimating α_L

The production function we wish to estimate is:

$$\textbf{(PF)} \quad \boldsymbol{y_{it} = \alpha_0 + \alpha_A a_{it} + \alpha_L l_{it} + \alpha_K k_{it} + \omega_{it} + e_{it}} \qquad \textbf{(1)}$$

$$\textbf{(ID)} \quad \boldsymbol{i_{it} = f_K(a_{it}, k_{it}, \omega_{it}, r_{it})} \qquad \textbf{(7)}$$

With (ID) representing firm i's demand rules for investments in capital in a dynamic decision model with state variables $a_{it}, k_{it}, \omega_{it}, r_{it}$ and age (a_{it}) and log capital (k_{it}) being treated symmetrically. In the Olley-Pakes (OP) framework, state variables are factors that persist over time and influence current production and affect the firm's investment decisions, which serve as a proxy for unobserved productivity (ω_{it}). Like the original OP approach, lagged labour

(l_{it-1}) is not considered to be a state variable here such that there is no labour adjustment cost and labour is a perfectly flexible input.

The identification of α_L using a control function approach relies on the following OP assumptions:

OP1: *$i_{it} = f_k(a_{it}, k_{it}, \omega_{it}, r_{it})$ is invertible in ω_{it}*

OP2: *For every firm i: $r_{it} = r_t$, There is no cross-sectional variability in unobservables affecting investment other than ω_{it} such that for every firm i: unobserved input prices are $r_{it} = r_t$.*

Under OP1, we assume that investment (i_{it}) is a strictly increasing function of unobserved productivity (ω_{it}) such that (ID) [$i_{it} = f_k(a_{it}, k_{it}, \omega_{it}, r_{it})$] can be inverted to get:

$$\boldsymbol{\omega_{it} = f_K^{-1}(a_{it}, k_{it}, i_{it}, r_{it})} \quad \textbf{(8)}$$

This allows ω_{it} to be proxied by a function of the observable variables $(a_{it}, k_{it}, i_{it}, r_{it})$ which we can then substitute into the (PF):

$$y_{it} = \alpha_0 + \alpha_A a_{it} + \alpha_L l_{it} + \alpha_K k_{it} + f_K^{-1}(a_{it}, k_{it}, i_{it}, r_{it}) + e_{it}$$

$$y_{it} = \alpha_L l_{it} + \underbrace{\{\alpha_0 + \alpha_A a_{it} + \alpha_K k_{it} + f_K^{-1}(a_{it}, k_{it}, i_{it}, r_{it})\}}_{\phi_t(a_{it}, k_{it}, i_{it})} + e_{it}$$

$$\boldsymbol{y_{it} = \alpha_L l_{it} + \phi_t(a_{it}, k_{it}, i_{it}) + e_{it}} \quad \text{(1st Stage Equation)} \qquad \textbf{(9)}$$

Where:

$$\phi_t(a_{it}, k_{it}, i_{it}) \equiv \alpha_0 + \alpha_A a_{it} + \alpha_K k_{it} + \omega_{it} = \alpha_0 + \alpha_A a_{it} + \alpha_K k_{it} + f_K^{-1}(a_{it}, k_{it}, i_{it}, r_{it})$$

which absorbs the effect of ω_{it} and the residual, e_{it}, is assumed to be i.i.d and uncorrelated with a_{it}, k_{it}, i_{it} and r_{it}.

We can approximate $\phi_t(\cdot)$ with a polynomial in a_{it}, k_{it} and i_{it}, and estimate α_L by OLS and obtain fitted values $\widehat{\phi_t}$ and $\widehat{\alpha_L}$ which we estimate to be 1.264 and statistically significant at the 1% level as shown in Table 7.

Second Stage: Estimating α_A and α_K (No Endogenous Exit Control):

The identification of α_A and α_K relies on the following OP assumptions:

> **OP3**: ω_{it} *follows a first-order Markov process:* $Pr\,[\omega_{it}|\,\omega_{it-1},\ldots,\omega_{i0}] = Pr\,[\omega_{it}|\,\omega_{it-1}]$
>
> **OP4:** *It takes "some" time to build up physical capital. Investment i_{it} is chosen at period t, but it is not productive until period $t+1$. Given the coefficient of depreciation of capital δ, the law of motion of capital is $k_{it} = (1-\delta)k_{it-1} + i_{it-1}$*

Since ω_{it} follows a first-order Markov process (OP 3):

$$\omega_{it} = \mathrm{E}\,[\omega_{it}|\,\omega_{it-1}] + \varepsilon_{it}$$

$$\omega_{it} = h(\,\omega_{it-1}) + \varepsilon_{it}$$

And given that $\omega_{it-1} = f_K^{-1}(a_{it-1}, k_{it-1}, i_{it-1}, r_{it-1})$, then:

$$\begin{aligned}\omega_{it} &= h(\,\omega_{it-1}) + \varepsilon_{it}\\ &= h(\,f_K^{-1}(a_{it-1}, k_{it-1}, i_{it-1}, r_{it-1})) + \varepsilon_{it}\\ &= h(\phi_{t-1}(a_{it-1}, k_{it-1}, i_{it-1}) - \alpha_0 - \alpha_A a_{it-1} - \alpha_K k_{it-1}) + \varepsilon_{it}\end{aligned}$$

Given our first stage estimates, we can use polynomial approximation for $h(\cdot)$ and estimate α_A and α_K from the following equation by Non-Linear Least Squares (NLLS):

$$y_{it} - \widehat{\alpha_L} l_{it} = \alpha_0 + \alpha_A a_{it} + \alpha_K k_{it} + h(\hat{\phi}_{t-1} - \alpha_0 - \alpha_A a_{it-1} - \alpha_K k_{it-1}) + \varepsilon_{it} + e_{it}$$

$$\boldsymbol{y_{it} - \widehat{\alpha_L} l_{it} = \alpha_A a_{it} + \alpha_K k_{it} + \tilde{h}(\hat{\phi}_{t-1} - \alpha_A a_{it-1} - \alpha_K k_{it-1}) + \varepsilon_{it} + e_{it}} \qquad \textbf{(10)}$$

Where function $\tilde{h}(\cdot)$ subsumes both occurrences of α_0

Correcting for Endogenous Exit:

Endogenous exit occurs when a firm's decision to leave the market is not random but is correlated with its productivity ω and its input decisions such that the innovation term ξ has different distributions for exiting firms ($\xi_{it}^{d=0}$) and surviving firms ($\xi_{it}^{d=1}$).

This introduces selection bias in the sample as survival isn't random, but rather it's tied to productivity and, indirectly, to state variables like age (a_{it}) and capital (k_{it}). In our sample, we observe that surviving firms tend to have higher capital levels but are generally younger than exiting firms. This suggests a complex selection mechanism where capital investment appears to support survival, but older firms face higher exit probabilities, potentially due to inability to adapt to changing market conditions or technological obsolescence. The higher output-to-capital ratio among exiting firms further indicates they may operate with unsustainable efficiency levels, possibly sacrificing long-term viability for short-term output.

These characteristics create correlations between survival, productivity, and state variables that, if not accounted for, would lead to biased estimates of input elasticities (α_A and α_K). Rather than uniformly upward bias, the selection effects likely differ across inputs: the age coefficient may be underestimated due to the over-representation of younger firms in the surviving sample, while capital coefficients might be overestimated due to the positive correlation between capital levels and survival probability. Correcting for this selection mechanism is therefore essential for obtaining consistent estimates of the production function parameters.

Why α_L is unaffected

Our first-stage estimate of α_L in the OP framework remains unchanged whether there is endogenous exit or not, as it relies on a static OLS regression that treats labour as a flexible input, uncorrelated with the error term, e_{it}. Since labour can be adjusted continuously and does not require long-term investment like capital, exiting firms does not create the same selection bias in the relationship between labour and output such that α_L remains unbiased.

Exit Condition and Productivity Threshold:

The decision of firm i (d_{it}) to leave the market by the end of period t is $d_{it} = 0$, whereas if firm i decides to stay in the market for an extra period, $d_{it} = 1$. A firm exits the market ($d_{it} = 0$) if:

$$\begin{aligned} \omega_{it} &< \omega^*_{it}(a_{it}, k_{it}) \\ h(\,f_K^{-1}(a_{it-1}, k_{it-1}, i_{it-1}, r_{it-1})) + \xi_{it} &< \omega^*_{it}(a_{it}, k_{it}) \\ \xi_{it} &< \omega^*_{it}(a_{it}, k_{it}) - h(\,f_K^{-1}(a_{it-1}, k_{it-1}, i_{it-1}, r_{it-1})). \end{aligned}$$

Where $\omega^*_{it}(a_{it}, k_{it})$ is the productivity threshold below which staying in the market is unprofitable given capital k_{it} and age a_{it}

Given **OP4** ($k_{it} = (1 - \delta)k_{it-1} + i_{it-1}$) and that $a_{it} = a_{it-1} + 1$, capital k_{it} is a function of $k_{it-1}$and i_{it-1} and age a_{it} is a function of a_{it-1} such that:

$$s_{t-1}(a_{it-1}, k_{it-1}, i_{it-1}) = \omega^{*}_{it}(a_{it-1}, k_{it-1}, i_{it-1}) - h(f_K^{-1}(a_{it-1}, k_{it-1}, i_{it-1}, r_{it-1}))$$

Where $\omega^{*}_{it}(a_{it}, k_{it}) = \omega^{*}_{it}(a_{it-1}, k_{it-1}, i_{it-1})$ reflects the dependence on the lagged variables for determining current states and s_{t-1} is the survival threshold. Given this, the **optimal exit decision rule** can then be expressed as:

$$\boldsymbol{d_{it}} = \begin{cases} \mathbf{1} & \boldsymbol{if \;\; \xi_{it} \geq s_{t-1}(a_{it-1}, k_{it-1}, i_{it-1})} \\ \mathbf{0} & \boldsymbol{if \;\; \xi_{it} < s_{t-1}(a_{it-1}, k_{it-1}, i_{it-1})} \end{cases} \qquad \textbf{(11)}$$

This decision rule implies that surviving firms ($d_{it} = 1$) have systematically higher observed productivity (ω_{it}) and productivity innovations (ξ_{it}) than exiting firms ($d_{it} = 0$).

Propensity Score:

The probability a firm survives into period t given past information is:

$$\begin{aligned} \mathbf{Pr}(\boldsymbol{d_{it} = 1 | a_{it-1}, k_{it-1}, i_{it-1}, t-1}) &= \mathbf{Pr}\;[\boldsymbol{\xi_{it} \geq s_{t-1}(a_{it-1}, k_{it-1}, i_{it-1})}] \\ &= \mathbf{1} - \boldsymbol{F_{\xi}[s_{t-1}(a_{it-1}, k_{it-1}, i_{it-1})]} \qquad \textbf{(12)} \\ &= \boldsymbol{P_{it}} \end{aligned}$$

Where P_{it} is the propensity score representing the likelihood of survival based on lagged capital, investment, and age.

Selection Bias in $\boldsymbol{\xi_{it}}$

For surviving firms ($d_{it} = 1$) , the productivity shock term ξ_{it} has a truncated distribution as only firms with productivity above the survival threshold ($\xi_{it} \geq s_{t-1}$) remain. We define the expected value of ξ_{it} given survival as:

$$E\;[\xi_{it} | d_{it} = 1] = \lambda(P_{it}) \qquad \textbf{(13)}$$

Where $\lambda(P_{it})$ is the correction term that accounts for the truncation in the distribution of ξ_{it}. In the case that ξ_{it} follows a normal distribution, the correction term $\lambda(P_{it})$ is the inverse mills ratio:

$$\boldsymbol{\lambda(P_{it}) = \frac{\phi(s_{t-1})}{1 - \Phi(s_{t-1})}} \qquad \textbf{(13)}$$

$\phi(s_{t-1})$ is the value of the standard normal PDF evaluated at the survival threshold s_{t-1}
$\Phi(s_{t-1})$ is the CDF of the standard normal distribution evaluated at s_{t-1}
$1 - \Phi(s_{t-1})$ represents the probability of survival (ie $\xi_{it} \geq s_{t-1}$)

We can therefore write the innovation term of surviving firms as:

$$\boldsymbol{\xi_{it}^{d=1} = \lambda(P_{it}) + \tilde{\xi}_{it}} \qquad \textbf{(14)}$$

Where $\tilde{\xi}_{it}$ is mean independent of the lagged inputs: $E[\tilde{\xi}_{it}|(a_{it-1}, k_{it-1}, i_{it-1})] = 0$

Adjusting for Endogenous Exit:

Without endogenous exit, the second stage equation is:

$$\boldsymbol{y_{it} - \hat{\alpha}_L l_{it} = \alpha_0 + \alpha_A a_{it} + \alpha_K k_{it} + h(\hat{\phi}_{t-1} - \alpha_0 - \alpha_A a_{it-1} - \alpha_K k_{it-1}) + \varepsilon_{it} + e_{it}} \quad \textbf{(10)}$$

Where estimates $\hat{\alpha}_L$ and $\hat{\phi}_{t-1}$ come from the first stage and ε_{it} is assumed to have a zero mean. With endogenous exit, only surviving firms ($d_{it} = 1$) are observed and $\xi_{it}^{d=1} = \lambda(P_{it}) + \tilde{\xi}_{it}$ has a non-zero mean $\lambda(P_{it})$ such that:

$$\begin{aligned} y_{it} - \hat{\alpha}_L l_{it} &= \alpha_0 + \alpha_A a_{it} + \alpha_K k_{it} + h(\hat{\phi}_{t-1} - \alpha_0 - \alpha_A a_{it-1} - \alpha_K k_{it-1}) + \xi_{it}^{d=1} + e_{it} \\ &= \alpha_0 + \alpha_A a_{it} + \alpha_K k_{it} + h\big(\hat{\phi}_{t-1} - \alpha_0 - \alpha_A a_{it-1} - \alpha_K k_{it-1}\big) + \lambda(P_{it}) + \tilde{\xi}_{it} + e_{it} \\ &= \alpha_0 + \alpha_A a_{it} + \alpha_K k_{it} + g(\hat{\phi}_{t-1} - \alpha_0 - \alpha_A a_{it-1} - \alpha_K k_{it-1}, P_{it}) + \tilde{\xi}_{it} + e_{it} \end{aligned}$$

$$\boldsymbol{y_{it} - \hat{\alpha}_L l_{it} = \alpha_A a_{it} + \alpha_K k_{it} + \tilde{g}(\hat{\phi}_{t-1} - \alpha_A a_{it-1} - \alpha_K k_{it-1}, P_{it}) + \tilde{\xi}_{it} + e_{it}} \quad \textbf{(14)}$$

Where function $\tilde{g}(\cdot)$ subsumes both occurrences of α_0

If we don't correct for endogenous exit, the non-zero mean of the productivity shock for surviving firms, $\mathrm{E}[\xi_{it}^{d=1}] = \lambda(P_{it})$, gets absorbed into the estimated coefficients for age α_A and for capital α_K. This causes an upward bias in these estimates as surviving firms, which have higher capital (k_{it}) and older age (a_{it}) also experience larger innovation shocks (ξ_{it}).

The higher output of surviving firms is partly due to their elevated productivity (higher baseline productivity (ω_{it}) and more favourable innovation shocks (ξ_{it})), not just their capital (k_{it}) and age (a_{it}). Without correction, this makes k_{it} and a_{it} seem more important to output than they really are causing us to overestimate their contribution.

Given our first stage estimates $\hat{\alpha}_L$ and $\hat{\phi}_{t-1}$ and by incorporating our estimate of the propensity score, P_{it}, (probability of survival as a function of k_{it}, a_{it}, and i_{it}) we can use polynomial approximations for $\tilde{g}(\cdot)$ and estimate α_A and α_L from (2) by NLLS as shown below in Table 9:

Table 9: OP CFA With & Without Endogenous Exit Control

Variable	OP CFA Not Controlling for Endogenous Exit	OP CFA Controlling for Endogenous Exit
l_{it}	1.264*** (0.006)	1.264*** (0.006)
a_{it}	0.213*** (0.003)	0.19*** (0.004)
k_{it}	1.043*** (0.019)	1.06** (0.02)

Legend: * p<0.05; ** p<0.01; *** p<0.001

[F, 10 Points]. *Compute the standard errors from the second step of the OP estimation method by clustered bootstrap, treating all observations for a single firm as one cluster. You can learn how to implement bootstrap methods with STATA from chapter 13 of Cameron & Trivedi (2009). Why do we need to do this to estimate the standard errors in the second step of the OP estimation procedure?*

In the OP framework, the first stage uses a control function approach to estimate the labour coefficient ($\hat{\alpha}_L$) and the productivity term ($\hat{\phi}_t$) via a partially linear model. The second stage estimation of α_A and α_K via nonlinear least squares (NLLS), uses these first-stage estimates, and treats them as fixed values, ignoring their variability. This introduces downward bias in the second-stage NLLS standard errors as $\hat{\alpha}_L$ and $\hat{\phi}_t$ from the first stage are estimates, not true values, and therefore their variability propagates to the second stage. Consequently, the standard errors do not therefore account for the sampling variability of the first-stage estimates, leading to underestimated standard errors and overly optimistic inference (i.e. confidence intervals and p-values).

Secondly, in a standard NLLS regression, standard errors are also computed under the assumption that all observations are independent and identically distributed (i.i.d), such that there is no autocorrelation, and that the residuals are homoscedastic. Panel data exhibits within-firm autocorrelation due to persistent productivity (ω_{it} follows a Markov process via **OP3**: $\omega_{it} = h(\omega_{it-1}) + \varepsilon_{it}$) and capital dynamics (**OP4** ($k_{it} = (1 - \delta)k_{it-1} + i_{it-1}$)), violating NLLS's independence assumption. This autocorrelation reduces the effective sample size, further biasing standard errors downward.

A clustered bootstrap, resampling entire firms as clusters, addresses both problems. We perform 400 bootstrap replications, sufficient for standard-error estimation (Cameron & Trivedi, Section 13.3.4). By clustering at the firm-level and re-estimating both stages on each bootstrap sample, we allow the first-stage estimation error to propagate and properly account for within-firm autocorrelation. This clustered bootstrap provides more accurate standard errors, protecting against understated inference.

The results in Table 10 highlight the impact of this correction. The bootstrapped standard errors of the OP estimate of the age coefficient increased only marginally over the conventional S.E., largely remaining the same. In contrast, for the capital coefficient, the standard errors decrease markedly when using the bootstrap approach. Without controlling for endogenous exit, the conventional standard error is 0.019 but shrinks to 0.007 with the bootstrap; similarly, with exit control, the SE drops from 0.020 to 0.006.

This counterintuitive shrinkage in the capital coefficient's standard errors likely reflects two factors. Firstly, the conventional NLLS method may overestimate uncertainty because it does not properly account for the true within-firm correlation structure and the fact that the first-stage estimates are treated as fixed. The bootstrap, by fully propagating the estimation error and respecting the firm-level clustering, suggests that the true sampling variability for capital is lower than what the conventional method implies.

Secondly, capital input, being governed by predictable dynamics such as depreciation and investment decisions, may exhibit a more stable relationship with output. The bootstrap method captures this stability more effectively than the conventional NLLS method and therefore yields tighter standard errors for α_K as a result.

In summary, while the OP two-step estimation produces biased standard errors due to ignoring first-stage variability and within-firm autocorrelation, and the bootstrapped standard errors clustered at the firm-level correct for this bias. The adjusted standard errors for the production function parameters are therefore more reliable, although the unexpected reduction in the capital coefficient's standard error indicates that the conventional approach might have inflated its variance estimate, revealing that capital's effect is estimated with greater precision than originally assumed.

Table 10: OP 2nd Stage Estimates With/Without Bootstrap Std. Errors

Variable	OP (No Endogenous Exit Control)	OP (No Endo. Exit Control + **Bootstrap std. err.) 400 reps**	OP (Endogenous Exit Control)	OP (Endo. Exit Control + **Bootstrap std. err.) 400 reps**
a_{it}	0.213*** (0.003)	0.213*** (0.003)	0.19*** (0.004)	0.19*** (0.004)
k_{it}	1.043*** (0.019)	1.043*** (0.007)	1.058** (0.02)	1.058*** (0.006)

Legend: * p<0.05; ** p<0.01; *** p<0.001

[G, 15 Points]. *How do your OP results compare with the previous ones**? Base your discussion on Griliches and Mairesse** (1995, NBER).*

<u>Key Differences and Implications</u>

Log Labour (l_{it}):

Looking at different estimation model results in Table 11, we can see that the OP estimate of α_L (1.264) aligns exactly with Pooled Unbalanced estimate (1.264) and is close to Pooled IMR and FE IMR estimates (1.265) but exceeds the FE estimate (1.233).

Griliches & Mairesse (1995) note that OLS-based methods like Pooled often overestimate labour coefficients due to simultaneity bias with labour correlating with unobserved productivity shocks. FE reduces this bias by accounting for firm-specific effects, as reflected by a lower coefficient (1.233). My OP estimates, despite using investment proxies to address simultaneity, produces a higher estimate (1.264), suggesting that it may not fully correct this bias for labour.

Firm Age ($\boldsymbol{a_{it}}$):

The OP estimate of $\boldsymbol{\alpha_A}$ is 0.213 without controlling for endogenous exit, and decreases to 0.19 with this control, within the prior range of estimates (0.208–0.231). Conversely, the Inverse Mills Ratio (IMR) approach increases $\boldsymbol{\alpha_A}$ from 0.219 to 0.231 (pooled model) and 0.208 to 0.216 (fixed effects model), revealing a methodological discrepancy.

Given that the production function is log-linear with age in levels and output log transformed, $\boldsymbol{\alpha_A}$ represents a semi-elasticity. This implies that a one-year increase in age should raise output by approximately $\boldsymbol{\alpha_A}$×100%. Griliches & Mairesse (1995) expect small or negative age effects on output due to stagnation in older firms. Yet, my estimates, ranging from 0.19 to 0.231 (≈ 19–23.1%), are positive and substantially larger than expected, starkly contrasting with Griliches & Mairesse's findings.

Surviving firms are younger on average than exiting ones, suggesting selection bias inflates $\boldsymbol{\alpha_A}$ in uncorrected models, as surviving older firms are disproportionately productive. The OP correction reduces $\boldsymbol{\alpha_A}$ from 0.213 to 0.19, aligning with expectations by mitigating this bias, while the IMR's increase (e.g., 0.219 to 0.231) is counterintuitive.

The OP reduction matches theory, though $\boldsymbol{\alpha_A}$ remains large, hinting age may proxy for factors like experience. The IMR's increase suggests imperfect selection correction. Treating age and capital symmetrically in OP may cause misspecification. While capital typically drives investment positively, age might have a weaker or even negative effect, older firms may invest less due to maturity or diminished growth opportunities. This misspecification could bias $\boldsymbol{\alpha_A}$ though the direction is uncertain. Still, the OP correction's effect supports its validity, though the large $\boldsymbol{\alpha_A}$ may reflect overestimation or proxy effects.

In summary, the corrected OP estimates of $\boldsymbol{\alpha_A}$ (decreasing from 0.213 to 0.19) better reflects the impact of endogenous exit and aligns with the expectation that selection bias inflates uncorrected estimates. The discrepancy with the IMR approach $\boldsymbol{\alpha_A}$ underscores methodological sensitivity in handling selection. While the positive estimates contrast with Griliches & Mairesse (1995), the OP reduction supports the idea that correcting for exit removes bias, though age's large effect may still indicate it proxies for factors beyond pure aging.

Log Capital ($\boldsymbol{k_{it}}$):

OP estimates of $\boldsymbol{\alpha_K}$ (1.04–1.043) are consistently below previous estimates (1.154–1.156). This reduction supports OP framework's ability in correcting for simultaneity bias for capital and mirrors their findings where corrected capital coefficients drop (Olley and Pakes (1996), pg. 192). Griliches & Mairesse (1995) highlight that OLS tends to inflate capital coefficients, while OP's approach yields lower, potentially more accurate estimates.

Returns to Scale:

In PF(1) $\boldsymbol{\alpha_K}$ and $\boldsymbol{\alpha_L}$ are the elasticities of output with respect to labour and capital, respectively, and their sum ($\boldsymbol{\alpha_K} + \boldsymbol{\alpha_L}$) indicates the returns to scale for these inputs (Griliches & Mairesse, 1995).

OP Framework sums range from 2.322 to 2.307, lower than previous estimates (2.387 to 2.421) suggesting partial correction of upward biases. However, these sums still suggest incredibly high returns to scale, much higher than typically expected. Griliches & Mairesse (1995) note that omitting key inputs from the model like materials or using deflated revenue as a proxy for output can lead to upward bias in $\boldsymbol{\alpha_K}$ and $\boldsymbol{\alpha_L}$ causing returns to scale to be overstated.

Conclusion:

In conclusion, while the OP framework offers improvements over OLS-based methods by addressing simultaneity bias, particularly for capital, it appears less effective in fully correcting bias for labour and may introduce misspecification issues when treating firm age as a state variable. The unexpectedly large and positive age coefficients, along with persistently high returns to scale estimates, point to potential omitted variable bias or structural assumptions that may not hold in practice. These findings underscore the importance of model specification and highlight the challenges of disentangling productivity drivers in firm-level production function estimation.

Table 11: All Parameter Estimations Comparison

Variable	Pooled Unbalanced	FE Unbalanced	Pooled Unbalanced IMR	FE Unbalanced IMR	OP (No Endo. Exit Control + **Bootstrap std. err.) 400 reps**	OP (Endo. Exit Control + **Bootstrap std. err.) 400 reps**
a_{it}	0.219*** (0.002)	0.208*** (0.003)	0.231*** (0.002)	0.216*** (0.004)	0.213*** (0.007)	0.19*** (0.004)
l_{it}	1.264*** (0.006)	1.233*** (0.006)	1.265*** (0.006)	1.265*** (0.006)	1.264*** (0.006)	1.264*** (0.006)
k_{it}	1.155*** (0.004)	1.154*** (0.005)	1.154*** (0.004)	1.156*** (0.005)	1.043*** (0.007)	1.058*** (0.006)
λ_{it} (IMR)			-0.231*** (0.022)	-0.202*** (0.023)		
Constant	4.049*** (0.041)	4.311*** (0.043)	3.984*** (0.043)	4.247*** (0.045)		
N	39569	39569	37389	37389		
R²	0.858	0.861	0.858	0.86		

Legend: * p<0.05; ** p<0.01; *** p<0.001

2 Article Discussion [20 Points] *Discuss the article by Jan De Loecker (Econometrica, 2011): "Product Differentiation, MultiProduct Firms and Estimating the Impact of Trade Liberalization on Productivity." Your discussion should be at most 2 pages with font size 12 and at least 2cm of side margins. In your assessment of the paper, try to be critical: what do you think about it? What is the author actually trying to do? Did he succeed? What are the pros and cons of the article, in your opinion? How does it relate to the existing literature? What do you think is the main contribution of the paper?*

Loecker (2011) aims to provide a more accurate estimation of firm-level productivity by addressing the bias introduced by using deflated revenue as a proxy for physical output. His goal is to isolate true productivity gains from trade liberalisation, separating them from price and demand effects that distort traditional measures.

Relation to Existing Literature

De Loecker (2011) builds on the insights of Klette-Griliches (1996), who addressed price-related biases in productivity estimation at an aggregate level, by extending the analysis to account for firm-specific demand elasticities and using trade policy shocks as exogenous demand shifters. In contrast to approaches like Olley and Pakes (1996) or Levinsohn and Petrin

(2003), which primarily address simultaneity bias arising from input choices, De Loecker directly confronts the bias introduced by unobserved price variation. His work complements studies such as Pavcnik (2002), which document productivity gains from trade liberalisation, by critically reassessing the size of those gains once price effects are properly accounted for.

The Problem: Using deflated revenue as an Output Proxy

Typically, productivity ω_{it} is estimated using a production function like the Cobb-Douglas:

$$\boldsymbol{Q_{it} = L_{it}^{\alpha_l} M_{it}^{\alpha_m} K_{it}^{\alpha_k} \exp\left(\omega_{it} + u_{it}\right)} \qquad (1)$$

$$\boldsymbol{q_{it} = \alpha_L l_{it} + \alpha_M m_{it} + \alpha_K k_{it} + \omega_{it} + u_{it}}$$

However, given physical output (Q_{it}) is rarely observed, revenue ($R_{it} = Q_{it} \times P_{it}$), deflated by an industry price index (P_{st}), is used as a proxy for Q_{it} (in logs: $\tilde{r}_{it} = r_{it} - p_{st}$). Since revenue (in logs: $r_{it} = p_{it} + q_{it}$) reflects both output and firm-specific prices, using deflated revenue ($\tilde{r}_{it} = p_{it} + q_{it} - p_{st}$) as a proxy for output includes both output (q_{it})and firm-specific price variation ($p_{it} - p_{st}$). This means firms charging higher prices can appear more productive regardless of actual efficiency. As a result, standard productivity estimates capture not only efficiency, but also pricing power and demand conditions, leading to biased estimators of input productivity and misleading inferences about firm performance.

De Loecker's Solution:

De Loecker (2011) aims to address this bias by building on Klette-Griliches (1996) and integrating the Cobb-Douglas production function (1) with a CES demand system (2) to separate price effects from physical output.

$$\boldsymbol{Q_{it} = Q_{st} \left(\frac{P_{it}}{P_{st}}\right)^{\eta_s} e^{\xi_{it}}} \qquad (2)$$

In logs: $$\boldsymbol{q_{it} = q_{st} + \eta_s (p_{it} - p_{st}) + \xi_{it}} \qquad (3)$$

Where Q_{st} is segment level demand
P_{it} is the firm specific price
P_{st} is the industry price index
η_s is the elasticity of substitution

This equation links quantity demanded to prices and observable demand shifters (Q_{st}) allowing prices to be expressed as a function of quantities and demand conditions. Rearranging (3) and given that ($\tilde{r}_{it} = p_{it} + q_{it} - p_{st}$), deflated revenue is:

$$\tilde{r}_{it} = r_{it} - p_{st} = \frac{(\eta_s + 1)}{\eta_s} q_{it} - \frac{1}{\eta_s} q_{st} - \frac{1}{\eta_s} \xi_{it}$$

Substituting in the CD function (1) into this gives:

$$\tilde{r}_{it} = \frac{(\eta_s + 1)}{\eta_s} (\alpha_L l_{it} + \alpha_M m_{it} + \alpha_K k_{it} + \omega_{it} + u_{it}) - \frac{1}{\eta_s} q_{st} - \frac{1}{\eta_s} \xi_{it}$$

Which with reduced form coefficients becomes:

$$\boldsymbol{\tilde{r}_{it} = \beta_l l_{it} + \beta_m m_{it} + \beta_k k_{it} + \beta_s q_{st} + \omega^*_{it} + \xi^*_{it} + u_{it}} \qquad (4)$$

And for multiproduct firms becomes:

$$\boldsymbol{\tilde{r}_{it} = \beta_{np} np_{it} + \beta_l l_{it} + \beta_m m_{it} + \beta_k k_{it} + \sum_{s=1}^{S} s_{is} \beta_s q_{st} + \omega^*_{it} + \xi^*_{it} + u_{it}} \qquad (5)$$

Where np_{it} is the log number of products a firm makes
s_{is} represents the share of firm i's output that comes from market segment s at time t
ω^*_{it} is productivity as embedded in revenue data adjusted by demand elasticity effects

The demand system (2) links quantity to price through the elasticity η_s estimated via $\beta_s = -\frac{1}{\eta_s}$ in (4 & 5), which captures how prices vary with demand conditions. When working with revenue rather than physical output data, the revenue function expressed as a function of inputs and demand shifters (4 & 5) inadvertently absorbs price effects into input coefficients and demand terms, leaving productivity shocks (ω_{it}) as a residual that combines both true productivity and price variation. De Loecker (2011) addresses this by explicitly modelling the demand side, which allows him to recover true productivity ($\omega_{it} = \frac{\eta_s}{(\eta_s+1)} \omega^*_{it}$) and input elasticities ($\alpha_h = \frac{\eta_s}{(\eta_s+1)} \beta_h$) by stripping out price distortions ($p_{it} - p_{st}$). For example, if higher firm revenue stems from higher prices rather than greater efficiency, the methodology adjusts for this, ensuring the productivity measure reflects only physical efficiency and not price or demand distortions.

<u>Empirical Findings & Contribution:</u>

Empirically De Loecker (2011) applies his framework to assess the true impact of trade liberalisation (quota reductions) on productivity in the Belgian textile industry (1994-2002). While standard methods estimate an 8% increase in productivity gains from trade liberalisation, his method only estimates a statistically significant 2% productivity gain.

This finding challenges the narrative of trade as a primary driver of firm-level productivity, with the apparent productivity gains from trade liberalisation being more reflective of price effects, specifically how increased import competition affects price, rather than pure efficiency gains. This is a critical insight for policymakers, as it reveals that the welfare gains from trade liberalisation are driven more by increased consumer surplus through lower prices than by improvements in producer productivity.

Strengths and Limitations:

To test the robustness of his findings, De Loecker employs various specifications, including alternative production functions (flexible vs. Cobb-Douglas), output measures (value-added vs. gross), and demand systems (nested CES vs. baseline CES). He consistently found that price effects dominate, with productivity gains remaining around 2%. However, a possible limitation of this approach is the assumption that multiproduct firms allocate inputs proportionately across products which may not fully capture real-world variability and could potentially bias the results.

Overall, De Loecker's methodology represents a significant advance by combining demand and production analysis to correct for price bias, though it requires detailed product-level data (e.g., trade quotas) which may limit its broader applicability. The findings imply that the primary welfare benefits of trade liberalisation arise from increased consumer surplus via lower prices, rather than substantial improvements in firm-level efficiency.

A Closing Note:

De Loecker's work feels particularly relevant given Trump's recent tariff policies. These policies appear to sacrifice the "big win" of trade liberalisation of increased consumer surplus via lower prices for a protectionist strategy that offers uncertain and likely limited productivity benefits. While the intent of the tariffs is to strengthen domestic industries, the result will likely just be higher costs for consumers and little efficiency effect for firms, marking a clear departure from the consumer-focused advantages of open trade that De Loecker emphasises.

References (MLA):

Olley, G. Steven, and Ariel Pakes. “The Dynamics of Productivity in the Telecommunications Equipment Industry.” *Econometrica*, vol. 64, no. 6, 1996, pp. 1263–97. *JSTOR*, https://doi.org/10.2307/2171831. Accessed 6 Apr. 2025.

Levinsohn, James, and Amil Petrin. "Estimating production functions using inputs to control for unobservables." The review of economic studies 70.2 (2003): 317-341.

Pavcnik, Nina. "Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants." The Review of economic studies 69.1 (2002): 245-276.

Klette, Tor Jakob, and Zvi Griliches. "The inconsistency of common scale estimators when output prices are unobserved and endogenous." Journal of applied econometrics 11.4 (1996): 343-361.

Levinsohn, James, and Amil Petrin. "Estimating production functions using inputs to control for unobservables." The review of economic studies 70.2 (2003): 317-341.

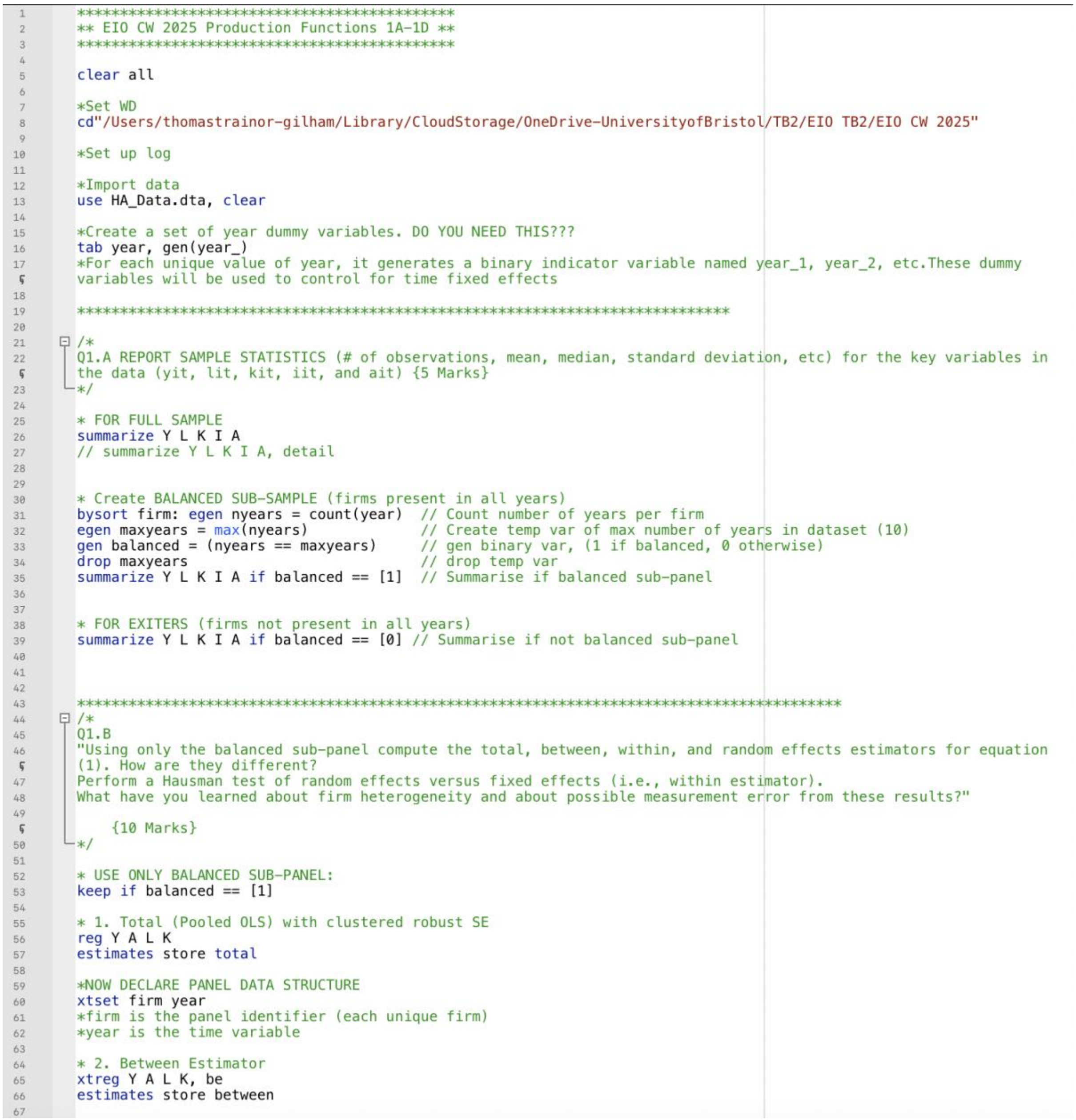
De Loecker, Jan. "Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity." Econometrica 79.5 (2011): 1407-1451.

Griliches, Zvi, and Jacques Mairesse. "Production functions: the search for identification." (1995).

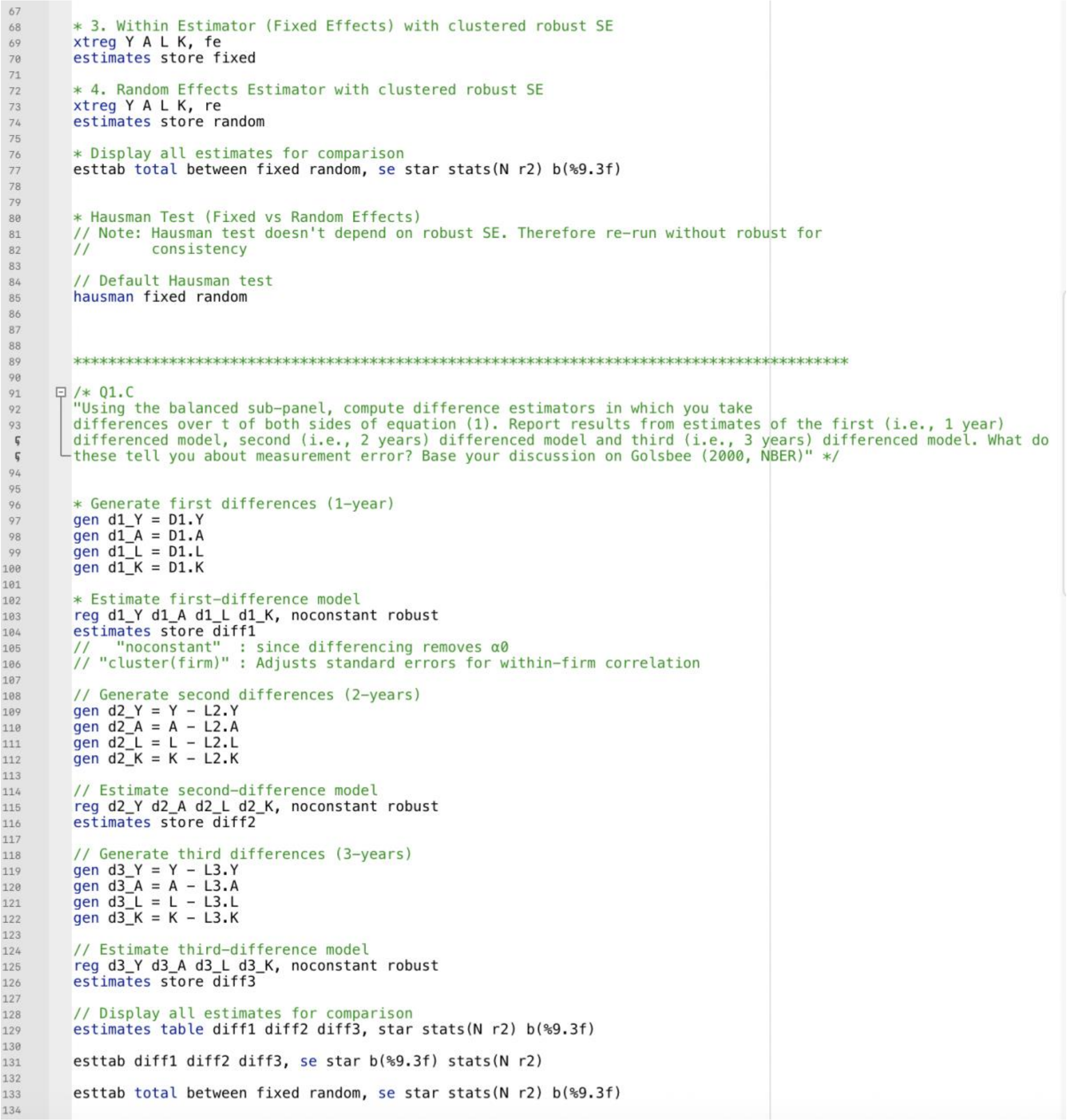
Cameron, A. Colin, and Pravin K. Trivedi. "Stata Version 11 and $ Microeconometrics using Stata." (2009).

Goolsbee, Austan. "What happens when you tax the rich? Evidence from executive compensation." Journal of Political Economy 108.2 (2000): 352-378.

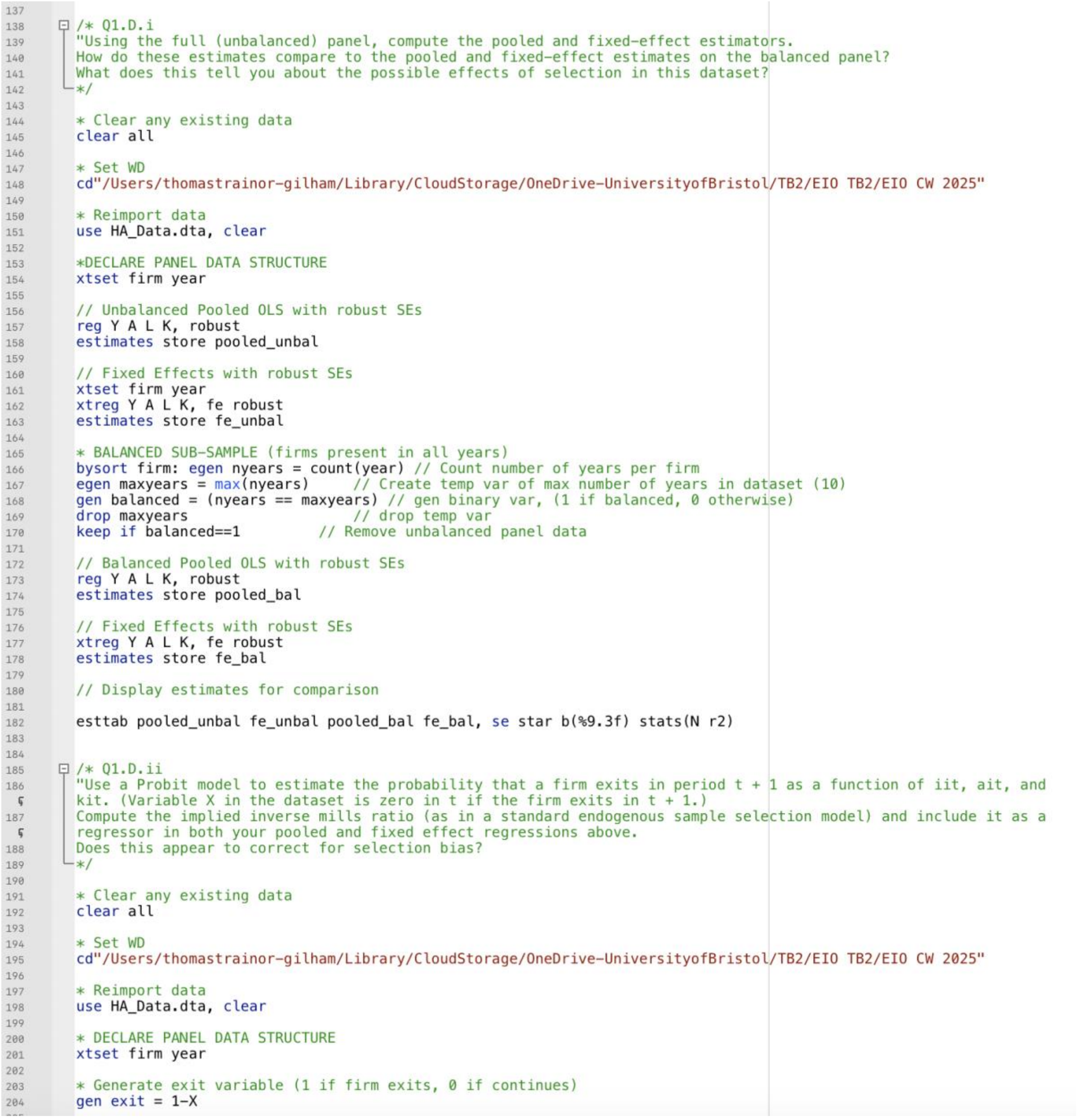
Appendix:



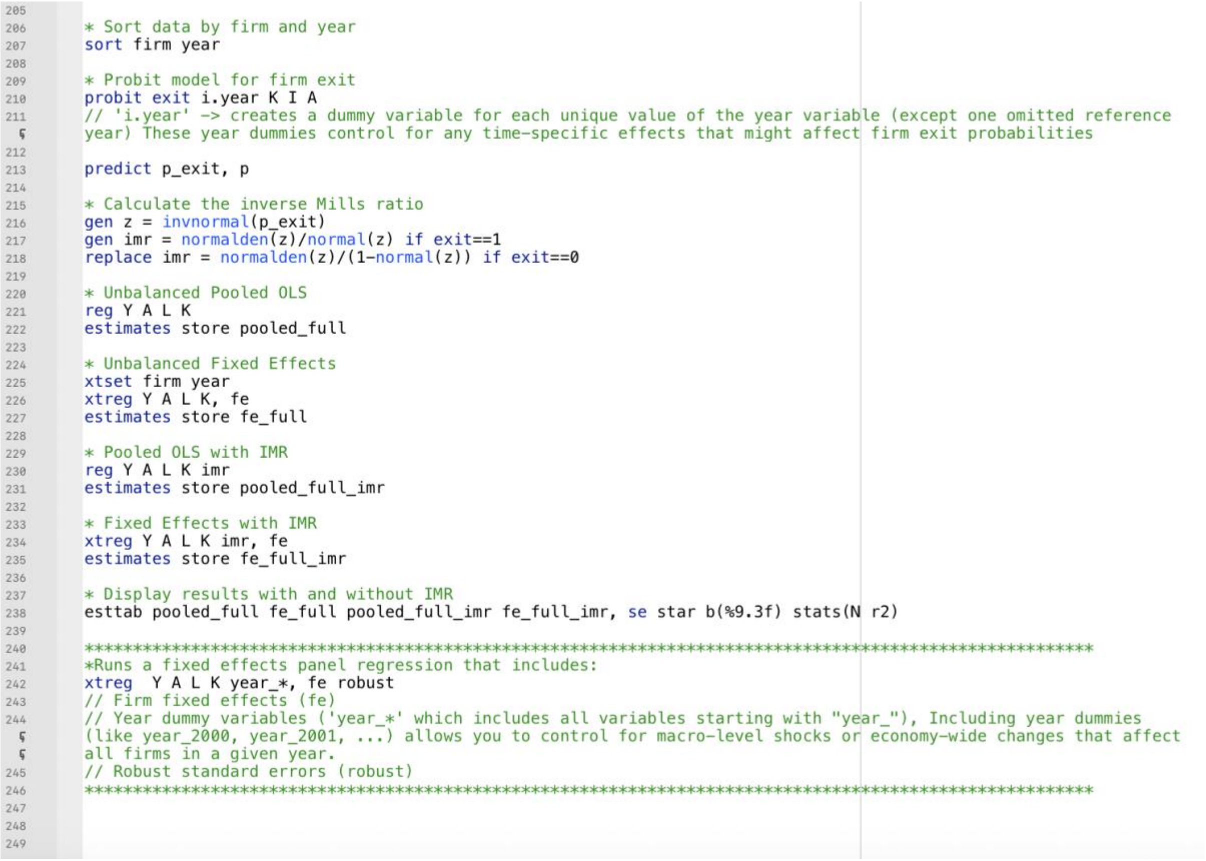
```
********************************************
** EIO CW 2025 Production Functions 1A-1D **
********************************************

clear all

*Set WD
cd"/Users/thomastrainor-gilham/Library/CloudStorage/OneDrive-UniversityofBristol/TB2/EIO TB2/EIO CW 2025"

*Set up log

*Import data
use HA_Data.dta, clear

*Create a set of year dummy variables. DO YOU NEED THIS???
tab year, gen(year_)
*For each unique value of year, it generates a binary indicator variable named year_1, year_2, etc.These dummy
variables will be used to control for time fixed effects

****************************************************************************

/*
Q1.A REPORT SAMPLE STATISTICS (# of observations, mean, median, standard deviation, etc) for the key variables in
the data (yit, lit, kit, iit, and ait) {5 Marks}
*/

* FOR FULL SAMPLE
summarize Y L K I A
// summarize Y L K I A, detail


* Create BALANCED SUB-SAMPLE (firms present in all years)
bysort firm: egen nyears = count(year)  // Count number of years per firm
egen maxyears = max(nyears)             // Create temp var of max number of years in dataset (10)
gen balanced = (nyears == maxyears)     // gen binary var, (1 if balanced, 0 otherwise)
drop maxyears                           // drop temp var
summarize Y L K I A if balanced == [1]  // Summarise if balanced sub-panel


* FOR EXITERS (firms not present in all years)
summarize Y L K I A if balanced == [0] // Summarise if not balanced sub-panel



*****************************************************************************************
/*
Q1.B
"Using only the balanced sub-panel compute the total, between, within, and random effects estimators for equation
(1). How are they different?
Perform a Hausman test of random effects versus fixed effects (i.e., within estimator).
What have you learned about firm heterogeneity and about possible measurement error from these results?"

    {10 Marks}
*/

* USE ONLY BALANCED SUB-PANEL:
keep if balanced == [1]

* 1. Total (Pooled OLS) with clustered robust SE
reg Y A L K
estimates store total

*NOW DECLARE PANEL DATA STRUCTURE
xtset firm year
*firm is the panel identifier (each unique firm)
*year is the time variable

* 2. Between Estimator
xtreg Y A L K, be
estimates store between

```



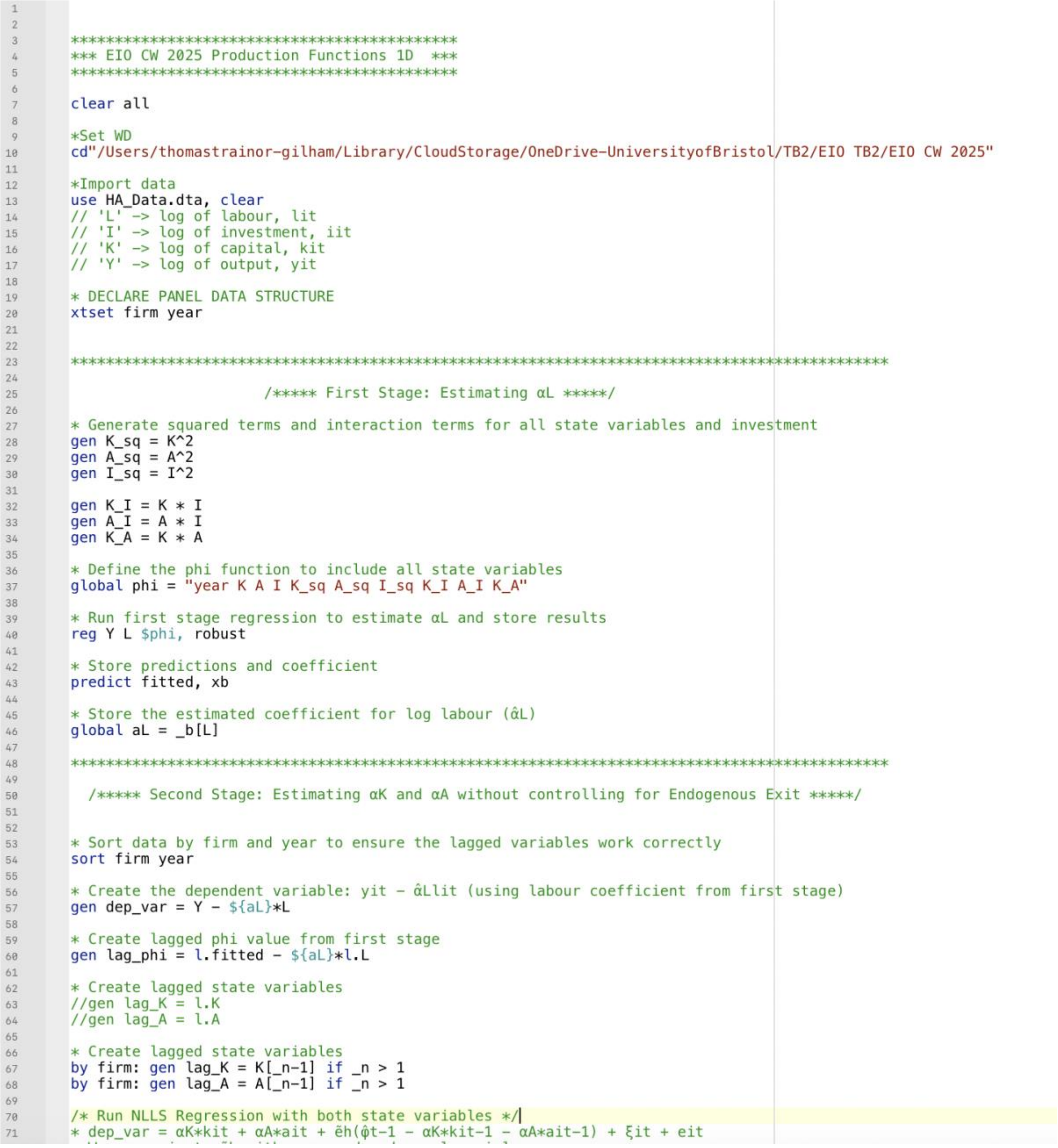
```
* 3. Within Estimator (Fixed Effects) with clustered robust SE
xtreg Y A L K, fe
estimates store fixed

* 4. Random Effects Estimator with clustered robust SE
xtreg Y A L K, re
estimates store random

* Display all estimates for comparison
esttab total between fixed random, se star stats(N r2) b(%9.3f)


* Hausman Test (Fixed vs Random Effects)
// Note: Hausman test doesn't depend on robust SE. Therefore re-run without robust for
//       consistency

// Default Hausman test
hausman fixed random



****************************************************************************************

/* Q1.C
"Using the balanced sub-panel, compute difference estimators in which you take
differences over t of both sides of equation (1). Report results from estimates of the first (i.e., 1 year)
differenced model, second (i.e., 2 years) differenced model and third (i.e., 3 years) differenced model. What do
these tell you about measurement error? Base your discussion on Golsbee (2000, NBER)" */


* Generate first differences (1-year)
gen d1_Y = D1.Y
gen d1_A = D1.A
gen d1_L = D1.L
gen d1_K = D1.K

* Estimate first-difference model
reg d1_Y d1_A d1_L d1_K, noconstant robust
estimates store diff1
//   "noconstant"  : since differencing removes α0
// "cluster(firm)" : Adjusts standard errors for within-firm correlation

// Generate second differences (2-years)
gen d2_Y = Y - L2.Y
gen d2_A = A - L2.A
gen d2_L = L - L2.L
gen d2_K = K - L2.K

// Estimate second-difference model
reg d2_Y d2_A d2_L d2_K, noconstant robust
estimates store diff2

// Generate third differences (3-years)
gen d3_Y = Y - L3.Y
gen d3_A = A - L3.A
gen d3_L = L - L3.L
gen d3_K = K - L3.K

// Estimate third-difference model
reg d3_Y d3_A d3_L d3_K, noconstant robust
estimates store diff3

// Display all estimates for comparison
estimates table diff1 diff2 diff3, star stats(N r2) b(%9.3f)

esttab diff1 diff2 diff3, se star b(%9.3f) stats(N r2)

esttab total between fixed random, se star stats(N r2) b(%9.3f)
```



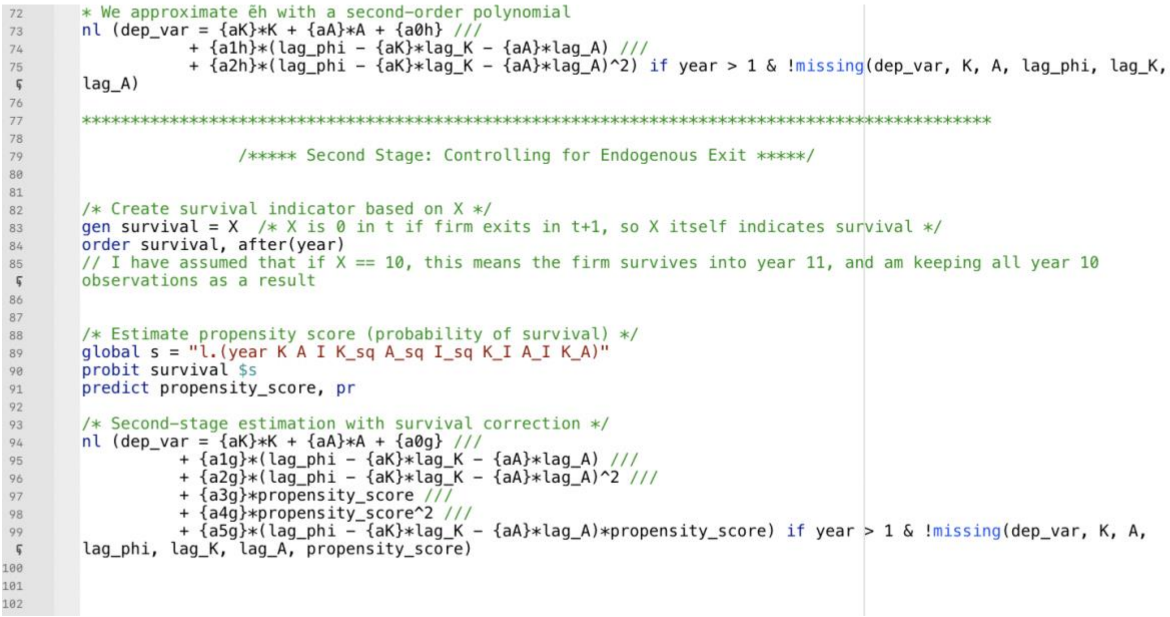
```
/* Q1.D.i
"Using the full (unbalanced) panel, compute the pooled and fixed-effect estimators.
How do these estimates compare to the pooled and fixed-effect estimates on the balanced panel?
What does this tell you about the possible effects of selection in this dataset?
*/

* Clear any existing data
clear all

* Set WD
cd"/Users/thomastrainor-gilham/Library/CloudStorage/OneDrive-UniversityofBristol/TB2/EIO TB2/EIO CW 2025"

* Reimport data
use HA_Data.dta, clear

*DECLARE PANEL DATA STRUCTURE
xtset firm year

// Unbalanced Pooled OLS with robust SEs
reg Y A L K, robust
estimates store pooled_unbal

// Fixed Effects with robust SEs
xtset firm year
xtreg Y A L K, fe robust
estimates store fe_unbal

* BALANCED SUB-SAMPLE (firms present in all years)
bysort firm: egen nyears = count(year) // Count number of years per firm
egen maxyears = max(nyears)     // Create temp var of max number of years in dataset (10)
gen balanced = (nyears == maxyears) // gen binary var, (1 if balanced, 0 otherwise)
drop maxyears                   // drop temp var
keep if balanced==1         // Remove unbalanced panel data

// Balanced Pooled OLS with robust SEs
reg Y A L K, robust
estimates store pooled_bal

// Fixed Effects with robust SEs
xtreg Y A L K, fe robust
estimates store fe_bal

// Display estimates for comparison

esttab pooled_unbal fe_unbal pooled_bal fe_bal, se star b(%9.3f) stats(N r2)


/* Q1.D.ii
"Use a Probit model to estimate the probability that a firm exits in period t + 1 as a function of iit, ait, and
kit. (Variable X in the dataset is zero in t if the firm exits in t + 1.)
Compute the implied inverse mills ratio (as in a standard endogenous sample selection model) and include it as a
regressor in both your pooled and fixed effect regressions above.
Does this appear to correct for selection bias?
*/

* Clear any existing data
clear all

* Set WD
cd"/Users/thomastrainor-gilham/Library/CloudStorage/OneDrive-UniversityofBristol/TB2/EIO TB2/EIO CW 2025"

* Reimport data
use HA_Data.dta, clear

* DECLARE PANEL DATA STRUCTURE
xtset firm year

* Generate exit variable (1 if firm exits, 0 if continues)
gen exit = 1-X
```



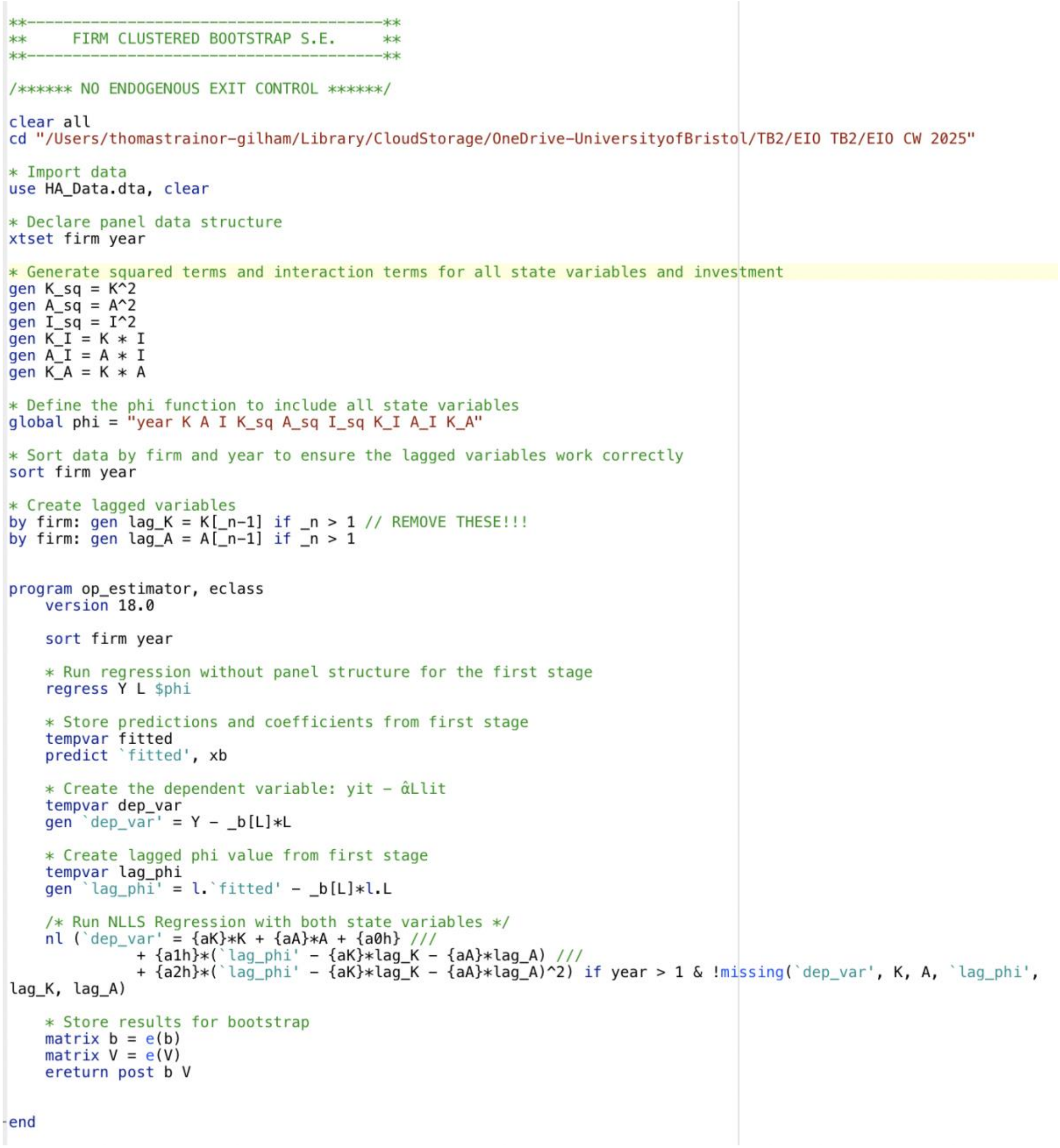
```
* Sort data by firm and year
sort firm year

* Probit model for firm exit
probit exit i.year K I A
// 'i.year' -> creates a dummy variable for each unique value of the year variable (except one omitted reference
year) These year dummies control for any time-specific effects that might affect firm exit probabilities

predict p_exit, p

* Calculate the inverse Mills ratio
gen z = invnormal(p_exit)
gen imr = normalden(z)/normal(z) if exit==1
replace imr = normalden(z)/(1-normal(z)) if exit==0

* Unbalanced Pooled OLS
reg Y A L K
estimates store pooled_full

* Unbalanced Fixed Effects
xtset firm year
xtreg Y A L K, fe
estimates store fe_full

* Pooled OLS with IMR
reg Y A L K imr
estimates store pooled_full_imr

* Fixed Effects with IMR
xtreg Y A L K imr, fe
estimates store fe_full_imr

* Display results with and without IMR
esttab pooled_full fe_full pooled_full_imr fe_full_imr, se star b(%9.3f) stats(N r2)

*********************************************************************************************************
*Runs a fixed effects panel regression that includes:
xtreg  Y A L K year_*, fe robust
// Firm fixed effects (fe)
// Year dummy variables ('year_*' which includes all variables starting with "year_"), Including year dummies
(like year_2000, year_2001, ...) allows you to control for macro-level shocks or economy-wide changes that affect
all firms in a given year.
// Robust standard errors (robust)
*********************************************************************************************************
```



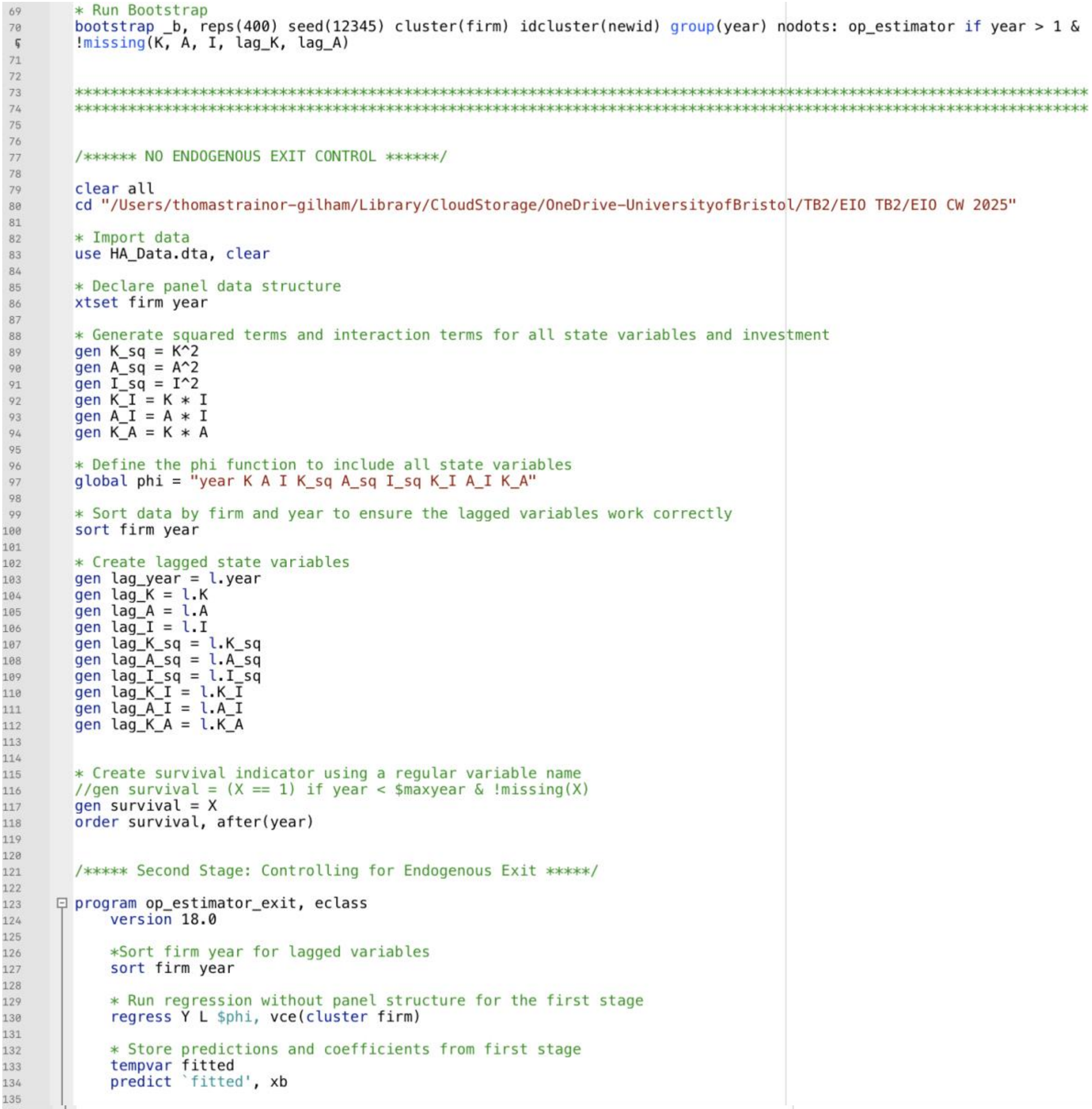
```


********************************************
*** EIO CW 2025 Production Functions 1D  ***
********************************************

clear all

*Set WD
cd"/Users/thomastrainor-gilham/Library/CloudStorage/OneDrive-UniversityofBristol/TB2/EIO TB2/EIO CW 2025"

*Import data
use HA_Data.dta, clear
// 'L' -> log of labour, lit
// 'I' -> log of investment, iit
// 'K' -> log of capital, kit
// 'Y' -> log of output, yit

* DECLARE PANEL DATA STRUCTURE
xtset firm year


*********************************************************************************************

                      /***** First Stage: Estimating αL *****/

* Generate squared terms and interaction terms for all state variables and investment
gen K_sq = K^2
gen A_sq = A^2
gen I_sq = I^2

gen K_I = K * I
gen A_I = A * I
gen K_A = K * A

* Define the phi function to include all state variables
global phi = "year K A I K_sq A_sq I_sq K_I A_I K_A"

* Run first stage regression to estimate αL and store results
reg Y L $phi, robust

* Store predictions and coefficient
predict fitted, xb

* Store the estimated coefficient for log labour (α̂L)
global aL = _b[L]

*********************************************************************************************

  /***** Second Stage: Estimating αK and αA without controlling for Endogenous Exit *****/


* Sort data by firm and year to ensure the lagged variables work correctly
sort firm year

* Create the dependent variable: yit - α̂Llit (using labour coefficient from first stage)
gen dep_var = Y - ${aL}*L

* Create lagged phi value from first stage
gen lag_phi = l.fitted - ${aL}*l.L

* Create lagged state variables
//gen lag_K = l.K
//gen lag_A = l.A

* Create lagged state variables
by firm: gen lag_K = K[_n-1] if _n > 1
by firm: gen lag_A = A[_n-1] if _n > 1

/* Run NLLS Regression with both state variables */
* dep_var = αK*kit + αA*ait + ẽh(ϕ̂t-1 - αK*kit-1 - αA*ait-1) + ξit + eit
```



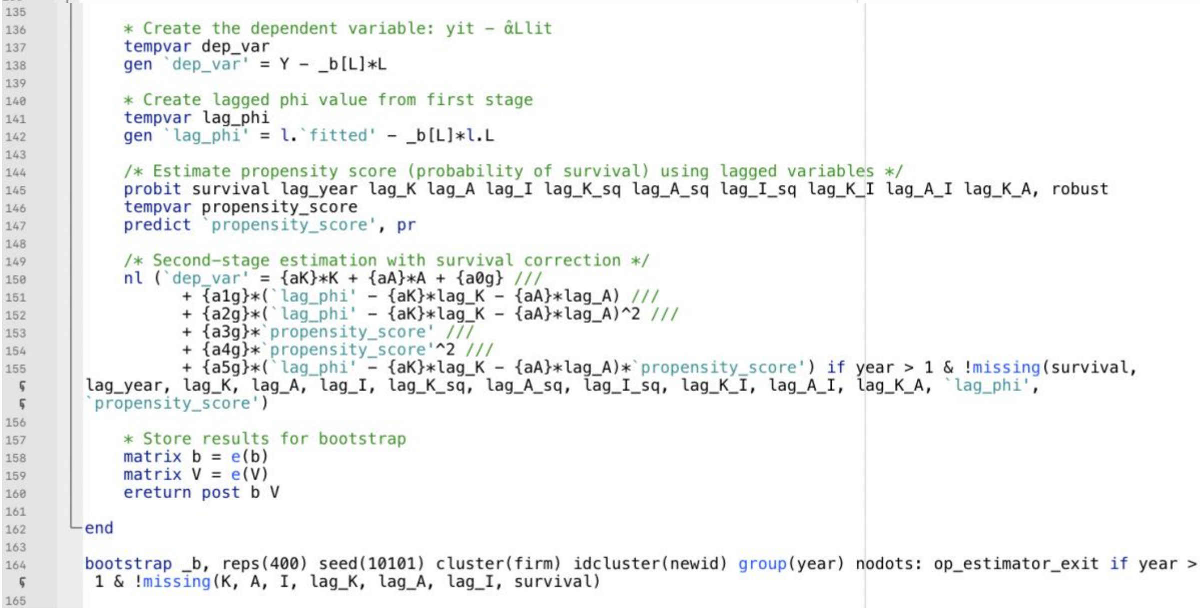
```
* We approximate ēh with a second-order polynomial
nl (dep_var = {aK}*K + {aA}*A + {a0h} ///
           + {a1h}*(lag_phi - {aK}*lag_K - {aA}*lag_A) ///
           + {a2h}*(lag_phi - {aK}*lag_K - {aA}*lag_A)^2) if year > 1 & !missing(dep_var, K, A, lag_phi, lag_K,
lag_A)

*********************************************************************************************

                /***** Second Stage: Controlling for Endogenous Exit *****/


/* Create survival indicator based on X */
gen survival = X  /* X is 0 in t if firm exits in t+1, so X itself indicates survival */
order survival, after(year)
// I have assumed that if X == 10, this means the firm survives into year 11, and am keeping all year 10
observations as a result


/* Estimate propensity score (probability of survival) */
global s = "l.(year K A I K_sq A_sq I_sq K_I A_I K_A)"
probit survival $s
predict propensity_score, pr

/* Second-stage estimation with survival correction */
nl (dep_var = {aK}*K + {aA}*A + {a0g} ///
          + {a1g}*(lag_phi - {aK}*lag_K - {aA}*lag_A) ///
          + {a2g}*(lag_phi - {aK}*lag_K - {aA}*lag_A)^2 ///
          + {a3g}*propensity_score ///
          + {a4g}*propensity_score^2 ///
          + {a5g}*(lag_phi - {aK}*lag_K - {aA}*lag_A)*propensity_score) if year > 1 & !missing(dep_var, K, A,
lag_phi, lag_K, lag_A, propensity_score)
```



```
**-----------------------------------------**
**     FIRM CLUSTERED BOOTSTRAP S.E.       **
**-----------------------------------------**

/****** NO ENDOGENOUS EXIT CONTROL ******/

clear all
cd "/Users/thomastrainor-gilham/Library/CloudStorage/OneDrive-UniversityofBristol/TB2/EIO TB2/EIO CW 2025"

* Import data
use HA_Data.dta, clear

* Declare panel data structure
xtset firm year

* Generate squared terms and interaction terms for all state variables and investment
gen K_sq = K^2
gen A_sq = A^2
gen I_sq = I^2
gen K_I = K * I
gen A_I = A * I
gen K_A = K * A

* Define the phi function to include all state variables
global phi = "year K A I K_sq A_sq I_sq K_I A_I K_A"

* Sort data by firm and year to ensure the lagged variables work correctly
sort firm year

* Create lagged variables
by firm: gen lag_K = K[_n-1] if _n > 1 // REMOVE THESE!!!
by firm: gen lag_A = A[_n-1] if _n > 1


program op_estimator, eclass
    version 18.0

    sort firm year

    * Run regression without panel structure for the first stage
    regress Y L $phi

    * Store predictions and coefficients from first stage
    tempvar fitted
    predict `fitted', xb

    * Create the dependent variable: yit − α̂Llit
    tempvar dep_var
    gen `dep_var' = Y − _b[L]*L

    * Create lagged phi value from first stage
    tempvar lag_phi
    gen `lag_phi' = l.`fitted' − _b[L]*l.L

    /* Run NLLS Regression with both state variables */
    nl (`dep_var' = {aK}*K + {aA}*A + {a0h} ///
              + {a1h}*(`lag_phi' − {aK}*lag_K − {aA}*lag_A) ///
              + {a2h}*(`lag_phi' − {aK}*lag_K − {aA}*lag_A)^2) if year > 1 & !missing(`dep_var', K, A, `lag_phi',
lag_K, lag_A)

    * Store results for bootstrap
    matrix b = e(b)
    matrix V = e(V)
    ereturn post b V


end
```



```
* Run Bootstrap
bootstrap _b, reps(400) seed(12345) cluster(firm) idcluster(newid) group(year) nodots: op_estimator if year > 1 &
!missing(K, A, I, lag_K, lag_A)


******************************************************************************************************************
******************************************************************************************************************


/****** NO ENDOGENOUS EXIT CONTROL ******/

clear all
cd "/Users/thomastrainor-gilham/Library/CloudStorage/OneDrive-UniversityofBristol/TB2/EIO TB2/EIO CW 2025"

* Import data
use HA_Data.dta, clear

* Declare panel data structure
xtset firm year

* Generate squared terms and interaction terms for all state variables and investment
gen K_sq = K^2
gen A_sq = A^2
gen I_sq = I^2
gen K_I = K * I
gen A_I = A * I
gen K_A = K * A

* Define the phi function to include all state variables
global phi = "year K A I K_sq A_sq I_sq K_I A_I K_A"

* Sort data by firm and year to ensure the lagged variables work correctly
sort firm year

* Create lagged state variables
gen lag_year = l.year
gen lag_K = l.K
gen lag_A = l.A
gen lag_I = l.I
gen lag_K_sq = l.K_sq
gen lag_A_sq = l.A_sq
gen lag_I_sq = l.I_sq
gen lag_K_I = l.K_I
gen lag_A_I = l.A_I
gen lag_K_A = l.K_A


* Create survival indicator using a regular variable name
//gen survival = (X == 1) if year < $maxyear & !missing(X)
gen survival = X
order survival, after(year)


/***** Second Stage: Controlling for Endogenous Exit *****/

program op_estimator_exit, eclass
    version 18.0

    *Sort firm year for lagged variables
    sort firm year

    * Run regression without panel structure for the first stage
    regress Y L $phi, vce(cluster firm)

    * Store predictions and coefficients from first stage
    tempvar fitted
    predict `fitted', xb

```



```

    * Create the dependent variable: yit − α̂Llit
    tempvar dep_var
    gen `dep_var' = Y - _b[L]*L

    * Create lagged phi value from first stage
    tempvar lag_phi
    gen `lag_phi' = l.`fitted' - _b[L]*l.L

    /* Estimate propensity score (probability of survival) using lagged variables */
    probit survival lag_year lag_K lag_A lag_I lag_K_sq lag_A_sq lag_I_sq lag_K_I lag_A_I lag_K_A, robust
    tempvar propensity_score
    predict `propensity_score', pr

    /* Second-stage estimation with survival correction */
    nl (`dep_var' = {aK}*K + {aA}*A + {a0g} ///
          + {a1g}*(`lag_phi' - {aK}*lag_K - {aA}*lag_A) ///
          + {a2g}*(`lag_phi' - {aK}*lag_K - {aA}*lag_A)^2 ///
          + {a3g}*`propensity_score' ///
          + {a4g}*`propensity_score'^2 ///
          + {a5g}*(`lag_phi' - {aK}*lag_K - {aA}*lag_A)*`propensity_score') if year > 1 & !missing(survival,
lag_year, lag_K, lag_A, lag_I, lag_K_sq, lag_A_sq, lag_I_sq, lag_K_I, lag_A_I, lag_K_A, `lag_phi',
`propensity_score')

    * Store results for bootstrap
    matrix b = e(b)
    matrix V = e(V)
    ereturn post b V

end

bootstrap _b, reps(400) seed(10101) cluster(firm) idcluster(newid) group(year) nodots: op_estimator_exit if year >
 1 & !missing(K, A, I, lag_K, lag_A, lag_I, survival)

```