MIT 14.76/760: Firms, Markets, Trade and Growth
Sp 2026, Lecture 8: Misallocation (part II)

Dave Donaldson

Misallocation and Productivity

- As we saw last lecture, Hsieh and Klenow (2009) document large dispersion in their measure of wedges (aka TFPR, or what we called $\overline{\mu}_{i,L}$) across firms
  - Large in the US
  - ... even larger in China and India
- But we also raised a concern: why would all firms have CRTS production functions, as HK 2009 assumed?
- Today we'll work on relaxing this assumption

Revisiting Wedges and HK (2009)

- Recall that our general definition of the wedge for firm i, at the actual allocation (with "bars" over it), when only one input ("L"), was

$$\overline{\mu}_{i,L} \equiv \frac{\overline{VMPL_i}}{\overline{w}} \equiv \frac{\overline{p}_i}{\overline{w}} \frac{\partial F_i(\overline{L}_i, A_i)}{\partial L_i}$$

- But if we define the *scale elasticity* as $\gamma_i(L_i) \equiv \frac{L_i}{Q_i} \frac{\partial F_i(L_i, A_i)}{\partial L_i}$ this becomes (using $\overline{R}_i \equiv \overline{p}_i \overline{Q}_i$)

$$\overline{\mu}_{i,L} = \frac{\overline{R}_i}{\overline{w}\overline{L}_i} \gamma_i(\overline{L}_i)$$

- Recall that the key HK (2009) assumption (for measuring wedges) was that all firms have a CRTS technology (i.e. $\gamma_i(\overline{L}_i) = 1$ for all $\overline{L}_i$)

- If instead we don't know $\gamma_i(\overline{L}_i)$ then we have an "identification problem": impossible to back out $\overline{\mu}_{i,L}$ from the available data on $\frac{\overline{R}_i}{\overline{w}\overline{L}_i}$

A More General Setup

- We now follow Carrillo et al (2023)...
- Starting at the "bar" (i.e. $\bar{Q}_i$, $\bar{L}_i$, $\bar{p}_i$, $\bar{w}$) allocation, as long $F_i(\cdot)$ is differentiable, a Taylor expansion implies:

$$\Delta Q_i = \frac{\partial F_i(\bar{L}_i)}{\partial L_i} \Delta L_i + \tilde{\varepsilon}_i$$

where $\tilde{\varepsilon}_i = \frac{\partial F_i(\bar{L}_i, \bar{A}_i)}{\partial A_i} \Delta A_i$ (plus any 2nd-order terms in Taylor expansion)

- Multiplying through by $\bar{p}_i$, we have (defining $\varepsilon_i \equiv \bar{p}_i \tilde{\varepsilon}_i$):

$$\bar{p}_i \Delta Q_i = \bar{\mu}_{i,L} \bar{w} \Delta L_i + \varepsilon_i \qquad (1)$$

Measuring Wedges

- Recall we have:

$$\bar{p}_i \Delta Q_i = \bar{\mu}_{i,L} \bar{w} \Delta L_i + \varepsilon_i$$

- This is like a regression equation (i.e. we have data on $\bar{p}_i \Delta Q_i$ and $\bar{w} \Delta L_i$ for a sample of firms i), but with the twist that the slope parameter (i.e. $\bar{\mu}_{i,L}$) can be different for each i
- There is no hope of measuring each slope $\bar{\mu}_{i,L}$, even if $Cov(\bar{w} \Delta L_i, \varepsilon_i) = 0$
- This is unfortunate, given that our whole interest is in the heterogeneity of the slopes $\bar{\mu}_{i,L}$, since we're interested in wedge dispersion

Measuring Wedges

- One option is to split up firms based on observable characteristics (e.g. bins of size in terms of $\overline{R}$) and do our best to estimate an average $\mathbb{E}\left[\bar{\mu}_i\right]$ separately by bin
- Doing so would require that $Cov(\bar{w}\Delta L_i, \varepsilon_i) = 0$ within each bin. Can we believe that might be true?

Measuring Wedges

- One option is to split up firms based on observable characteristics (e.g. bins of size in terms of $\overline{R}$) and do our best to estimate an average $\mathbb{E}\left[\bar{\mu}_i\right]$ separately by bin
- Doing so would require that $Cov(\bar{w}\Delta L_i, \varepsilon_i) = 0$ within each bin. Can we believe that might be true?
- No. So we need a valid instrumental variable (IV) for the endogenous variable $\bar{w}\Delta L_i$. Any ideas?

Where to find firm-level demand shocks?

Where to find firm-level demand shocks?

- Carrillo et al (2023) draw on features of Ecuador's public procurement system.

- In Ecuador, beginning in 2009, procurement contracts for certain construction services were required to be randomized among qualified applicants
  - Examples: Construction/repairs of public buildings, roads, schools, sewage, wells
  - Amounts of up to 0.0007% of central government's annual budget ($240,000 in 2014) must use this lottery system
  - Applies to both government and state-owned firms

Lottery Descriptive Statistics

	Mean	Median	N
Lotteries			
Contract amount (USD)	46,522	31,597	18,474
	(40,990)		
Contract duration (days)	64.5	60	18,467
	(34)		
# of Participants	10.1	4	18,474
	(16)		
Firms			
# of Lotteries Entered (per year)	3.5	1	9,393
	(7)		
# of Lotteries Won (per year)	0.4	0	9,393
	(1)		

Administrative Data on Firms

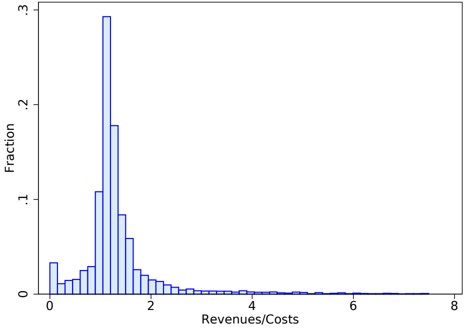
1. Transaction-level data from VAT "annexes" (monthly, 2008-2015)
    - "Sales" = sum of purchases from the firm reported in *other firms'* VAT annexes
    - Sales by type of client (procuring entities, other government, private) to look at crowd-out/in
2. Social security data of wages and employees (monthly, 2007-2015)
3. Corporate income tax filings (annual, 2008-2015)
    - Purchase of materials
    - Check on sales

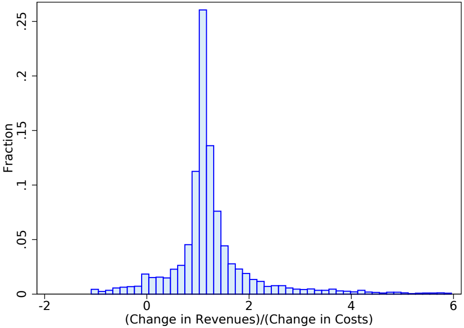
Summary Statistics

	(1) Mean	(2) Median	(3) Standard Deviation	(4) N
Firm age (years)	11.21	10	11.06	9,393
Is incorporated	0.18	0	0.38	9,393
Number of clients (third-party reported)	4.74	2	17.72	9,393
Sales (third-party reported)	122,985	44,046	249,812	9,393
Sales (self-reported)	128,795	48,429	260,573	9,393
Costs (self-reported)	114,623	38,499	246,388	9,393
Profits (self-reported)	14,172	9,613	30,447	9,393
Employees (social security)	4.40	2	10.10	9,393
Wages (social security)	7,399	2,880	25,926	9,393

Notes: This table presents summary statistics for each lottery participant's first year in which they entered a lottery. All monetary variables are in USD.

Comparing Revenues to Costs (in Levels and Annual Changes)





(This dispersion—in levels, on the left—is similar to what HK (2009) find for US manufacturing since their $TFPR_i \approx \frac{R_i}{C_i}$)

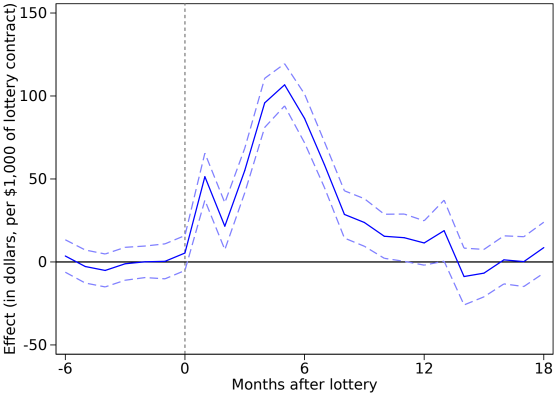
How to Use Lotteries to Generate Demand Shocks?

- Winning the lottery is random and hence exogenous. But entering a lottery is a choice, so probably not exogenous.
- How can we design an IV that only depends on the winning, not the participation?
- We know the lotteries entered by firm i at time t, and we know the dollar amount for winning (and the associated probability of winning) each lottery.
- So we can calculate $\mathbb{E}[W_{it}]$, the expected amount of winnings for firm i at time t
- And then construct an IV based on *deviations* from expected winnings

$$Z_{it} \equiv W_{it} - \mathbb{E}[W_{it}]$$

- This will be (mean-) independent of any variable

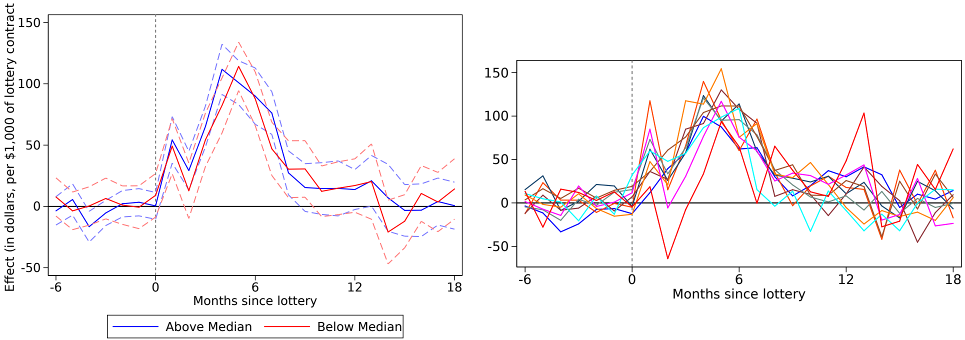
Effect of Lottery Winnings on Revenues (average effects)



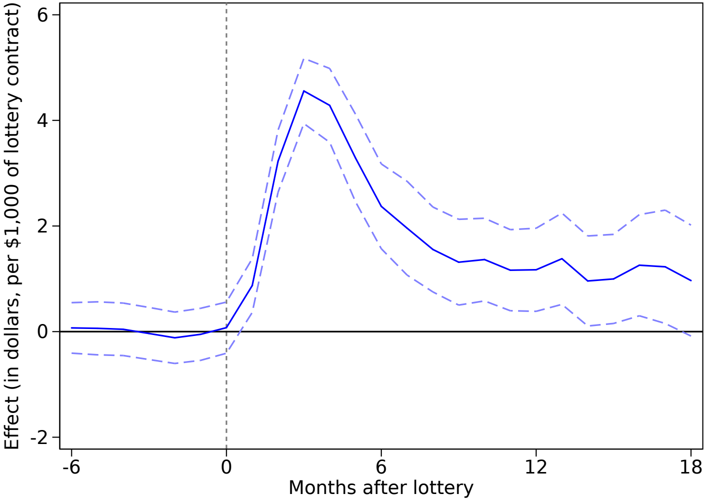
Estimates of a monthly "event-study" specification as follows:

$$R_{it} = \alpha + \sum_{m=-6}^{18} \beta(m) Z_{i,t-m} + \epsilon_{it}$$

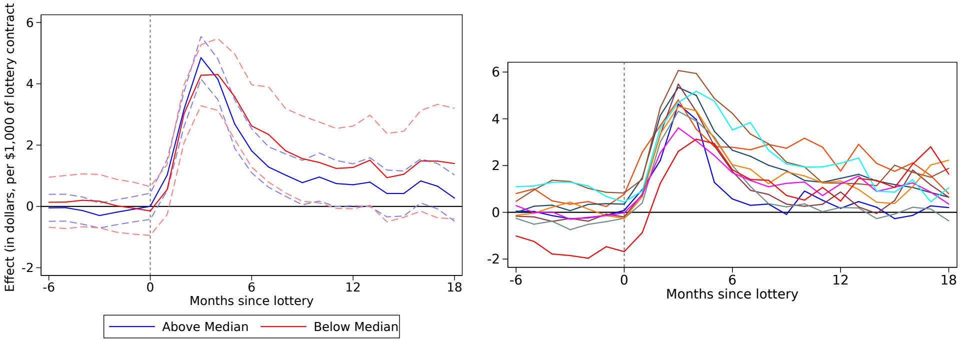
Effect on Total Revenues (heterogeneity by groups based on 2008 revenues, i.e. revenues before the lottery system)



Effect on Labor Payments (average effects)



Effect on Labor Payments (heterogeneity by groups based on 2008 revenues)



Assessing Misallocation

- These figures (and others like them for other categories of costs in Carrillo et al, 2023) suggest that, at least in terms of groups based on pre-period revenues, there is very little cross-firm heterogeneity in the slopes from either the
  - "Reduced form": R_{it} regressed on Z_{it}; or the
  - "First stage": L_{it} regressed on Z_{it}
- Such a scenario is consistent with there being very little heterogeneity in the IV coefficients, ie the slopes $\bar{\mu}_{i,L}$ that we are interested in for studying misallocation
- But can we conduct a version of this test that allows for *any* source of heterogeneity, not just groups based on pre-existing observable differences such as the 2008 revenues used here)?

Assessing Misallocation

- Recall the main equation we derived:

$$\bar{p}_i \Delta Q_i = \bar{\mu}_{i,L} \bar{w} \Delta L_i + \varepsilon_i$$

- Now square this to get:

$$(\bar{p}_i \Delta Q_i)^2 = (\bar{\mu}_i)^2 (\bar{w} \Delta L_i)^2 + 2\bar{\mu}_i \varepsilon_i \bar{w} \Delta L_i + (\varepsilon_i)^2 \tag{2}$$

  which is also “just" a regression equation—with dependent variable $(\bar{p}_i \Delta Q_i)^2$ and regressors $(\bar{w} \Delta L_i)^2$ and $2\bar{w} \Delta L_i$

- So could hope to use this second regression to recover an estimate of $\mathbb{E}\left[\bar{\mu}_i^2\right]$

Assessing Misallocation

- Recall the main equation we derived:

$$\bar{p}_i \Delta Q_i = \bar{\mu}_{i,L} \bar{w} \Delta L_i + \varepsilon_i$$

- Now square this to get:

$$(\bar{p}_i \Delta Q_i)^2 = (\bar{\mu}_i)^2 (\bar{w} \Delta L_i)^2 + 2\bar{\mu}_i \varepsilon_i \bar{w} \Delta L_i + (\varepsilon_i)^2 \quad (2)$$

  which is also "just" a regression equation—with dependent variable $(\bar{p}_i \Delta Q_i)^2$ and regressors $(\bar{w} \Delta L_i)^2$ and $2\bar{w}\Delta L_i$

- So could hope to use this second regression to recover an estimate of $\mathbb{E}\left[\bar{\mu}_i^2\right]$

- And then test:

$$\bar{\mu}_i = \mu \text{ for all } i \iff \mathbb{E}\left[\bar{\mu}_i^2\right] = \left(\mathbb{E}\left[\bar{\mu}_i\right]\right)^2$$

Assessing Misallocation (with endogenous $\overline{w}\Delta L_i$)

- With endogenous regressor(s), need an IV(s)
- But with IV regression, typically can't! recover simple average of slopes (like $\mathbb{E}\left[\bar{\mu}_i\right]$ in (1) or $\mathbb{E}\left[\bar{\mu}_i^2\right]$ in (2))
- However, in the current case Masten and Torgovitsky (REStat, 2017) describe a method that can work (but we won't go over this in detail).

Implications for Amount of Misallocation

- Applying Masten and Torgovitsky (2017) techniques, Carrillo et al (2023) find:
  - Estimate of $(\mathbb{E}[\bar{\mu}])^2 = (1.10)^2 = 1.21$
  - Estimate of $\mathbb{E}[\bar{\mu}^2] = 1.22$
- Fact that $(\mathbb{E}[\bar{\mu}])^2 \approx \mathbb{E}[\bar{\mu}^2] \Rightarrow \bar{\mu}_i \approx \mu$ for all i
- Can also apply same additional assumptions as in HK (2009), about the CES utility function with $\sigma = 3$, to calculate the welfare costs of misallocation (in this sector), i.e. $\frac{TFP(\overline{\mathbf{Q}})}{TFP(\mathbf{Q}*)}$

Bootstrap (200 reps) estimates of $1 - \frac{TFP(\overline{\mathbf{Q}})}{TFP(\mathbf{Q}^*)}$

"Bootstrap" is a nonparametric method for calculating confidence intervals on estimates with complicated distributional properties. (You don't need to know the details of this!)