Markups and Public Procurement: Structural Estimation and Causal Evidence from Czech Construction Tenders

Master of Science Thesis by Marek Chadim,

Department of Economics, Stockholm School of Economics *

November 6, 2024

I document the evolution of market power using firm-level data from the Czech construction sector since 2006. Contrary to the global trend of rising markups, I find that aggregate markups have decreased, declining from 40% above marginal cost in 2006 to 30% in 2021, driven primarily by firms in the upper tail of the markup distribution. By linking this data with government tenders, I examine the relationship between markups and public procurement. I find that markups are significantly higher when controlling for unobserved productivity; government contractors have price-to-marginal-cost ratios that are 0.3 higher than those of private-sector firms; and firm-level markups increase by 12% upon a firm's entry into public procurement.

Keywords Firm Behavior: Empirical Analysis; Production, Cost, Capital, Total Factor Productivity; National Government Expenditures and Related Policies: Procurement; Pricing and Market Structure; Industry Studies: Construction

JEL D22, D24, H57, L11, L74

*Email: 42624@student.hhs.se. Replication files: https://github.com/marek-chadim/Markups-and-Public-Procurement. I thank Jaakko Meriläinen and Matěj Bajgar for helpful comments, and Jiří Skuhrovec of Datlab, s.r.o. for providing administrative data on public procurement. This thesis builds upon my undergraduate thesis (doi: http://hdl.handle.net/20.500.11956/184831) at the Institute of Economic Studies, Faculty of Social Sciences, Charles University Prague, where I accessed the Czech firm-level financial data.

1 Introduction

In this paper, I investigate the relationship between markups and public procurement in the Czech construction sector over the period 2006–2021. By linking firm-level financial data with public procurement records, I explore how firms' entry into public procurement markets affects their pricing power, measured through markups. The primary focus is on understanding whether firms that win government contracts exhibit higher markups compared to those operating solely in the private sector, and how these markups evolve over time.

My research shows that, contrary to the global trend of rising markups, the average markup in the Czech construction sector has declined over the period studied. However, firms engaged in public procurement tend to maintain significantly higher markups than their private-sector counterparts. Specifically, I find that markups increase by approximately 12% after firms enter the public procurement market. This suggests that public procurement may provide opportunities for firms to exert greater pricing power, potentially due to reduced competition or other factors such as discretion and favoritism.

These findings contribute to ongoing discussions on the efficiency of public procurement systems and the extent to which they foster or inhibit competition. They also highlight the broader implications of public procurement on market power, particularly in contexts where government contracts constitute a significant portion of economic activity. This paper builds upon existing literature by providing empirical evidence of markup dynamics in public procurement, filling a gap in the understanding of how firms' market power evolves when they engage with the public sector.

1.1 Public Procurement

Public procurement accounts for approximately 12% of GDP across OECD countries, playing a critical role in government expenditure and economic activity (OECD, 2021). Ensuring efficiency in procurement is essential to prevent waste and maximize taxpayer value. Despite its significance, persistent inefficiencies arise due to issues like discretion, political favoritism, and limited competition—topics widely explored in the literature.

One major challenge in public procurement is balancing discretion with the risk of rent-seeking. Discretion allows procurement officials to tailor decisions to specific needs but also creates opportunities for corruption. In the Czech Republic, Palguta and Pertold (2017) document that officials often manipulate procurement thresholds, adjusting contract values to avoid competitive bidding. This practice bypasses open tenders and reduces transparency. Szucs (2024) finds more discretion to public agencies in Hungary results in higher prices, less productive contractors and more politically connected winners. Decarolis et al. (2020) find that discretionary procedures in Italian government contracts, especially those with fewer bidders, increase the likelihood of corruption, frequently benefiting politically connected firms. Both studies suggest that while discretion can enhance efficiency, it requires robust oversight to avoid exploitation for personal or political gain. Political favoritism further contributes to inefficiency. Firms with political connections often secure contracts at inflated prices without delivering better outcomes. Baránek and Titl (2024) show that, in the Czech Republic, politically connected firms win contracts priced 6% higher than competitively awarded ones, with no corresponding improvement in quality. This misallocation of resources echoes findings in Italy, where Bandiera et al. (2009) report that public bodies with weaker governance consistently overpay for comparable goods, exacerbating procurement inefficiencies. A lack of competition, particularly single-bid contracts, also undermines procurement effectiveness. In the Czech Republic, Titl (2023) shows that about 23% of public contracts are awarded through single-bid procedures, which drive up prices. A 2012 reform aimed at reducing single-bidding resulted in a 10% drop in procurement costs, underscoring the advantages of fostering competition. In the U.S., Kang and Miller (2022) find that single-bid contracts are prevalent in federal procurement, and that unchecked discretion can stifle competition, leading to higher prices.

Overall, these findings highlight that while flexibility and discretion are important for procurement, they must be paired with strong oversight and governance to prevent corruption and ensure cost-effective, competitive outcomes.

1.2 Markups as a Measure of Market Power

The study of markups, defined as the ratio of price to marginal cost

$$\mu \equiv \frac{P}{c},$$

is crucial for understanding market power in both theoretical and empirical economics. Recent research has focused on how markups reflect competitive dynamics within industries, influencing welfare, pricing strategies, and policy interventions. Syverson (2024) emphasizes the dual role of markups: they signal the presence of market power and quantify its extent. Markups go beyond price-setting; they encapsulate key features of imperfect competition, allowing researchers to measure how far a firm deviates from competitive pricing. Syverson also highlights the deadweight loss associated with markups; firms with market power, unable or unwilling to engage in perfect price discrimination, often forgo socially beneficial transactions to protect inframarginal profits.

De Loecker et al. (2020) show a significant increase in U.S. markups since 1980, rising from 21% above marginal cost to 61% by 2016. This growth is concentrated among the largest firms, which have expanded their share of economic activity, indicating a growing concentration of market power. Hall (2018) corroborates this trend, further strengthening evidence of rising markups. Additionally, Autor et al. (2020) explore how increased market power has affected labor markets, particularly with the emergence of "superstar firms"—large, highly productive companies dominating their industries. These firms capture a disproportionate share of profits while employing fewer workers, contributing to the decline in labor's share of income. Autor argues that this concentration exacerbates income inequality as capital captures a larger portion of economic gains. The causes of rising markups are debated. Some researchers link this trend to weakened antitrust enforcement, which allows firms to consolidate power.

In contrast, Miller (2024) attributes the rise to technological advancements. His review suggests that productivity gains, reductions in marginal costs, and product quality improvements have enabled firms to increase markups without necessarily harming consumer welfare. This aligns with the view that technological progress boosts firm efficiency, allowing them to command higher prices through improved cost structures. However, the welfare implications of rising markups remain complex. Berry et al. (2019) caution that while higher markups may reflect efficiency gains, they could also indicate reduced competition, especially in industries with high entry barriers, where dominant firms extract excessive rents. This highlights the need to consider industry-specific

factors when interpreting markup trends and assessing policy implications. Alternatively, Shapiro and Yurukoglu (2024) argue that rising markups do not necessarily signal weakened competition. In many sectors, they reflect competitive dynamics, where the most efficient firms grow larger and capture market share by offering superior products at lower costs. While some industries may experience a decline in competition, others might become more competitive due to technological progress and efficiency gains.

Estimating Markups Two primary methods are commonly used to estimate markups: the demand-based and production-based approaches. The demand-based approach estimates markups by analyzing a firm's residual demand curve, using demand elasticities to infer market power. This method is prevalent in industrial organization (IO), where firms' pricing decisions reflect their ability to raise prices above marginal cost. In perfectly competitive markets, firms have little influence over prices, but as competition diminishes, firms can increase markups. While this approach is central to studies of imperfect competition and monopoly pricing, it requires extensive data and relies on assumptions about consumer preferences and market conditions. It is especially sensitive to the specification of demand systems and the availability of price and quantity data.

A significant drawback of the demand-based method is the need for detailed instruments to identify shifts in demand. In markets with differentiated products, the heterogeneity of consumer preferences complicates estimation. De Loecker and Syverson (2021) discuss how these assumptions can affect the accuracy of demand-based estimates. While useful, this method may be impractical in industries where demand data are scarce, necessitating alternative approaches.

De Loecker and Warzynski (2012) introduced a flexible production-based method for estimating firm-level markups directly from production data. By combining input-output elasticities with input revenue shares, this approach captures market power in both product and factor markets. It overcomes many limitations of demand-based models by relying on more readily available production data. However, the De Loecker and Warzynski (DLW) method has its own limitations. One critique is its assumption of Hicks-neutral productivity, where productivity shifts equally across all inputs. Raval (2023) shows that this assumption, combined with labor market frictions like hiring costs or monopsony power, can introduce bias by treating inputs such as labor and materials similarly. Additionally, Bond et al. (2021) highlight that the DLW approach suffers from "omitted price bias" when using revenue data rather than actual output quantities, as revenue-based estimates fail to capture firm-specific price variations. Despite these criticisms, De Ridder et al. (2024) demonstrate that revenue-based markups still reveal important trends over time and between firms,

making the DLW approach valuable for studying market power patterns, even if precise markup levels require careful interpretation.

The production-based approach has been widely applied both within and outside IO. It has played a crucial role in documenting the rise of markups and contributed to discussions on market power and industry concentration in the U.S. and globally (De Loecker et al., 2020). Although this method faces challenges, particularly in measuring output elasticities, it has renewed interest in how productivity data can inform debates on market power. Beyond IO, this approach has connected productivity data to global discussions on declining labor shares, the effects of globalization, and labor market issues such as monopsony power (Autor et al., 2020), making it a valuable tool in competition policy discussions.

The DLW method relies on a first-order condition for the cost minimization of a variable input in production. This method requires the output elasticity of the variable input and its revenue share. A key assumption is that, in each period, producers minimize costs by choosing inputs optimally, free from frictions. The following markup formula arises from production and cost data:

$$\mu_{it} = \theta_{it}^{V} \frac{P_{it} Q_{it}}{P_{it}^{V} X_{it}^{V}}, \tag{1}$$

where θ_{it}^{V} represents the output elasticity of input X^{V}, which is generally specific to the producer and time period.

The flexibility of this approach stems from its independence from assumptions about market conduct or a particular demand system. Multiple first-order conditions—one for each variable input—allow for the estimation of markups with different inputs. Regardless of the input used, two key components are needed: the revenue share and the output elasticity. De Loecker and Warzynski (2012) assume that firms are price-takers in input markets, which does not preclude input providers from charging markups, potentially leading to double marginalization. The derived formula highlights that marginal cost is inferred from a single variable input, without assuming substitution elasticities among inputs or returns to scale.

An essential component of this method is the output elasticity θ_{it}^{V}, estimated using the control function approach pioneered by Olley and Pakes (1996), modified by Levinsohn and Petrin (2003), and consolidated by Ackerberg et al. (2015). This method links the production function to the economic model describing firm behavior and the competitive environment in which firms operate.

1.3 Markups and Public Procurement

I use the empirical framework of De Loecker and Warzynski (2012) to analyze how markups differ between government contractors and private-sector firms. Additionally, I examine the impact of public procurement entry on firm markups. To explore this, I correlate markups with a firm's public procurement status and evaluate whether they change upon entering public procurement, controlling for input usage. The empirical model is detailed in Appendix A. It is important to note that I do not take a specific stance on any particular economic model when interpreting the estimated markup parameters. However, I reference various mechanisms to interpret the findings.

Much of the existing literature on public procurement has focused on comparisons between tenders, with little attention paid to comparisons between public procurement and private markets. Economic models in industrial organization, which consider heterogeneous producers and firm-specific markups, suggest that more productive firms set higher markups, as these firms can afford to pay the costs associated with public procurement entry. Therefore, it is expected that government contractors will have higher markups. This relationship, which stems from supply-side heterogeneity (productivity), is predicted by many models. Another strand of trade literature investigates the role of quality differences between firms. If government contractors produce higher-quality goods using higher-quality inputs, they can charge higher markups. These mechanisms suggest that, in the cross section, government contractors should have higher markups. However, the dynamics of markups over time as firms enter public procurement markets—compared to firms already engaged in public procurement or those in the private sector—remain unclear. This paper examines both the validity of economic models that link public procurement to markups and provides new evidence on the dynamics between markups and public procurement status.

Given the above, I expect higher markups for government contractors. However, these differences are influenced by both supply- and demand-side factors that affect costs and prices. The procedure of De Loecker and Warzynski (2012) provides estimates of both markups and productivity, enabling further decomposition of the markup differences between private-sector firms and government contractors. By controlling for differences in marginal costs (i.e., productivity), I assess whether government contractors still maintain higher markups. This allows me to isolate the productivity component and explore other factors affecting prices, such as corruption, favoritism, discretion, inefficiencies, and single-bidding practices, as highlighted in the public procurement literature.

2 Data

Public procurement contracts in the Czech Republic are awarded under nationwide regulations, which require procurers to publish contract details in an online system. The dataset covers contracts awarded by central, regional, and municipal governments, as well as government-owned enterprises, spanning the period from 2005 to 2021. Since the majority of these contracts pertain to construction projects, I utilize a dataset that includes all firms active in the Czech construction sector between 2006 and 2021. This dataset, obtained through Charles University's licensed access to the MagnusWeb firm-level database (Dun & Bradstreet Czech Republic, a.s.), provides full company accounts for an unbalanced panel of 1,297 firms. The sample is restricted to firms with financial data for at least two consecutive periods, and the top and bottom one percentile of firms are trimmed based on their sales-to-cost of goods sold ratio to improve robustness in markup estimation. By linking firm-level data with public procurement records, I can track each firm's involvement in public procurement. At any given time, I can determine whether a firm is active exclusively in the private sector, has entered public procurement, has exited it, or remains an ongoing government contractor.

Table 1: **Firms and Public Procurement in Czech Construction**

Year	No. firms	No. Firms Active in Public Procurement
2006	227	86
2007	290	97
2008	348	112
2009	412	131
2010	497	138
2011	506	130
2012	457	152
2013	235	97
2014	243	96
2015	245	118
2016	338	174
2017	660	376
2018	708	388
2019	764	408
2020	769	426
2021	562	298

3 Results from Structural Estimation

In this section, I apply the De Loecker and Warzynski (2012) framework to estimate markups for Czech manufacturing firms, testing whether government contractors, on average, exhibit different markups compared to private-sector firms. Additionally, I leverage the substantial entry into public procurement markets within my dataset to analyze how markups evolve as firms enter and exit these markets. To the best of my knowledge, this is the first study to provide robust econometric evidence of this relationship.

To estimate markups in the context of public procurement, I incorporate a firm's procurement status, as well as any other factor influencing optimal input demand, into the control function. Specifically, I include a firm's public procurement status in all input demand equations (as an element of z_{it}) and allow it to directly influence the law of motion for productivity. After estimating the output elasticity of the variable inputs, I compute the implied markups using the first-order conditions (FOCs) derived in Equation 1.

The analysis begins by documenting the main patterns of markups in the Czech construction sector. I focus on these estimates to highlight several key findings. First, I examine the relationship between markups and public procurement status both cross-sectionally and over time. Second, I explore how markups relate to other economic variables. Finally, I discuss the broader implications of my results on aggregate markup trends.

3.1 The Evolution of Markups in Czech Construction

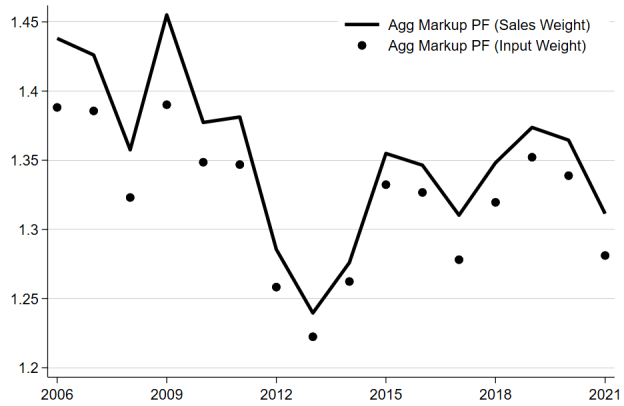
Aggregate Markup The markup measure in Equation 1 is the product of the output elasticity, θ, and the inverse of the variable input's revenue share, $\frac{PQ}{P^V V}$. The revenue share is directly observed in firms' income statements, while I estimate the output elasticities. These elasticities are both firm- and time-specific, reflecting technological differences across firms and changes over time. The average markup is calculated as:

$$\mu_t = \sum_i m_{it}\mu_{it},$$

where m_{it} represents the weight of each firm. I use the share of sales as the weight and also compare it with total costs as the input weight. The gap between sales-weighted and input-weighted aggregate markups is notable. Figure 1 illustrates the evolution of average markups in

the construction sector over time. Early in the sample period, markups remained stable around 1.45, declined to 1.25 during 2012–2014, and then increased slightly above 1.35 by the 2020s. In 2006, the average markup stood at 44% above marginal cost, compared to 31% in 2021.

Figure 1: Average Markups. Estimated output elasticities $\hat{\theta}_{it}^{V}$ are time- and firm-specific.

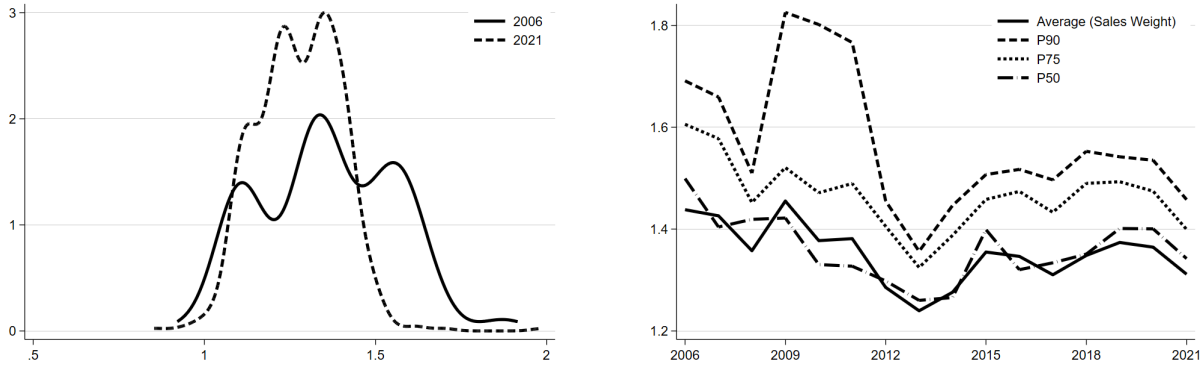


The Distribution of Markups Averages alone do not fully capture the changes in the distribution of markups. The strength of the De Loecker and Warzynski (2012) method lies in its ability to estimate firm-specific markups, allowing for an analysis of the entire distribution. One key finding is that the overall decline in markups is driven by a small number of firms, with the majority experiencing only a modest decrease. To illustrate the shifts in the distribution, I plot the kernel density of unweighted markups for 2006 and 2021 in Figure 2a. The variance of markups has narrowed over time, with a notable thinning and shortening of the upper tail. This reduction in the upper tail largely accounts for the decline in the average markup. Appendix B provides the data underlying Figure 1.

Since the kernel density does not account for firm weights, I also plot the moments of the sales-weighted markup distribution over time in Figure 2b. Firms are ranked by markup, with percentiles weighted by each firm's market share, making the percentiles comparable to the share-weighted average. The ranking is updated annually, so the top firms may vary from year to year.

The decline in average markups is driven by firms in the upper half of the markup distribution. The median (P50) and lower percentiles remain mostly stable. The sharpest drop occurs at the 90th percentile, particularly before 2012, where markups fall from 1.8 to 1.4. This indicates that the overall reduction in average markup is primarily due to a few firms experiencing significantly lower markups than in earlier periods. This finding aligns with Titl (2023), who shows that the 2012

Figure 2: The Distribution of Markups $\hat{\mu}_{it}$



(a) Kernel Density (unweighted)

(b) Percentiles (revenue-weighted)

Czech public procurement reform, which eliminated single-bid contracts, led to a 10% reduction in prices relative to estimated costs for these contracts.

3.2 Relating Markups and Public Procurement

Do Government Contractors Have Different Markups? Given the availability of firm-specific markups, I can directly relate a firm's markup to its public procurement status using a regression framework. I estimate the percentage difference in markups between government contractors and private-sector producers. A key advantage of using log markups is that the results remain robust even if the variable inputs used to compute markups are subject to adjustment costs. As long as government contractors are not disproportionately affected by these costs, the results hold. Additionally, I rely on logged markups because the variation in firm-level markups is substantial, and ordinary least squares (OLS) helps minimize proportional deviations rather than absolute deviations. After estimating the regression, I convert these percentage differences into absolute markup values.The specification used is:

$$\ln \mu_{it} = \delta_0 + \delta_1 pp_{it} + \mathbf{b}'_{it}\sigma + \nu_{it}, \tag{2}$$

where pp_{it} is a public procurement dummy, and δ_1 measures the percentage markup premium for government contractors.[1] The vector $\mathbf{b}_{it}$ contains all control variables, with σ as the corresponding coefficients. It is important to note that I do not interpret δ_1 as a causal parameter; instead, this

[1] I control for variable input and capital use to capture differences in size and factor intensity, as well as year-subindustry interactions to account for aggregate markup trends.

specification tests whether government contractors, on average, have different markups. To my knowledge, this relationship has not been previously documented, making these results an important contribution.

While I am not primarily interested in the coefficients of the control variables, I will revisit the correlation between markups and other economic factors later. This regression is estimated at the manufacturing level, with a full interaction of year and sub-industry dummies.[2] Once δ_1 is estimated, I compute the absolute markup difference by applying the percentage difference to the constant term, which represents the domestic average markup. I denote this difference by μ_{PP}, calculated as $\hat{\mu}_{PP} = \hat{\delta}_1 \exp(\hat{\delta}_0)$, following the parameter estimates. Table 2 presents the results.

Table 2: Markups and Public Procurement I: Cross-Section

Markup (level)	Coefficient	Std. Err.	z	[95% Conf. Interval]
Public Procurement Premium	0.327	0.018	18.08	[0.2918, 0.3628]

Note: Estimates are obtained after running equation 2. The standard error is derived from a nonlinear combination of the relevant parameter estimates, using the delta method, and is cluster robust. All regressions include variable input, capital, and full 47 year and NACE 2-digit division dummies as controls. N = 7,261. Adjusted R^2 = 0.758.

The percentage premium parameter estimate $\hat{\delta}_1$ is 0.149, with a highly precise standard error of 0.003. These results align with economic models where government-contracting firms, on average, charge higher markups due to greater productivity, allowing them to undercut rivals in the tender process. This prediction is supported by comparing the average markup of government contractors to private-sector firms in the cross-section. However, some models suggest that firms with identical productivity levels will charge the same markup, implying that productivity differences are the primary driver of markup variations. The estimation procedure provides both markups $\mu_{it} = \frac{P_{it}}{C_{it}}$ and productivity ω_{it}, allowing for a more nuanced analysis. When both a firm's public procurement status and productivity are included, the coefficient on public procurement (δ_1) decreases and changes sign, from 0.15 to 0.03, as expected. Controlling for productivity accounts for differences in marginal costs, meaning the coefficient on public procurement status now reflects the variation in average prices between government contractors and non-contractors. To clarify, I estimate the following:

$$(\ln P_{it} - \ln C_{it}) = \delta_0 + \delta_1 pp_{it} + \delta_2 \omega_{it} + b'_{it}\sigma + \nu_{it},$$

Here, δ_1 measures the average price difference (in percentages) if ω_{it} fully captures $\ln C_{it}$. Importantly, the public procurement effect remains significant even after controlling for productivity.

[2] I also estimated the model by year and industry; the magnitude varies as expected—see Appendix C.

In fact, the public procurement dummy explains approximately 20% of the markup difference, indicating that factors beyond productivity—such as differences in average prices—play a crucial role in explaining the markup disparity between government contractors and private-sector firms. This finding is consistent with the public procurement literature, which highlights varying levels of competitiveness between public procurement and private markets. Additionally, simple differences in demand elasticities and income across markets could also explain price variations, though data constraints prevent further discrimination between these mechanisms.

These results carry potentially significant policy implications. Models with heterogeneous firms emphasize the reallocation of market share from less efficient producers to more efficient ones, with government contractors expected to be more productive, enabling them to cover the fixed costs of entering public procurement markets. However, the findings call for a more cautious interpretation of the public procurement productivity premium and its role in aggregate productivity growth. Given that measured productivity is a residual of a sales-generating production function, it captures the unexplained portion of sales from the factors used in production and may include unobserved differences in input and output quality, as well as market power effects. These findings underscore the need to study the public procurement-productivity relationship alongside market power within an integrated framework. In the next section, I further explore the markup trajectory as a function of public procurement status.

Public Procurement Entry and Markup dynamics So far, I have estimated differences in average markups between government contractors and private-sector firms. My dataset also allows me to investigate whether markups vary significantly within the group of government contractors. In particular, I explore whether a specific pattern of markups emerges for firms entering public procurement markets, both before and after they become government contractors. These results can help test theories of self-selection into public procurement markets. I now focus on three categories of government contractors identifiable in my sample: *starters* (firms that enter public procurement), *quitters* (firms that exit public procurement), and *continuous contractors* (firms that contract with the government throughout the sample period) (see Appendix D). I estimate the following regression, comparing markups before and after public procurement entry and exit, while also estimating the markup differential for firms that consistently receive government contracts:[3]

$$\ln \mu_{it} = \gamma_0 + \gamma_1 \text{Entry}_{it} + \gamma_2 \text{Exit}_{it} + \gamma_3 \text{Always}_i + \mathbf{b}_{it}'\sigma + \nu_{it}, \tag{3}$$

[3]I exclude 152 firms that enter or exit public procurement markets more than once during the sample period.

where $\text{Entry}_{it} = 1$ if a firm becomes a government contractor and 0 otherwise, and $\text{Exit}_{it} = 1$ if a firm ceases contracting with the government. The constant term captures the average log markup for private-sector firms, including those that enter or exit public procurement markets. The primary interest lies in the coefficient γ_1, which measures the percentage markup difference for starters between the pre- and post-public procurement periods. Similarly, γ_2 measures the effect of public procurement exit on markups, while γ_3 captures the markup difference for continuous contractors, which I expect to be positive. Given the static nature of most models, there is little theoretical guidance for γ_1; hence, these results provide new empirical evidence on markup dynamics and public procurement status. I compute the implied markup-level effect of public procurement entry as:

$$\hat{\mu}_{PP}^{entry} = \hat{\gamma}_1 \exp(\hat{\gamma}_0),$$

and report the results in Table 3. I find that public procurement entry is associated with substantially higher markups, approximately 12%, after controlling for aggregate markup changes. The other coefficients align with expectations (see Appendix E). Notably, including productivity (as done earlier) reveals a significant effect for public procurement entry (t-statistic = 4.5). This indicates that price changes are linked to public procurement entry, likely influenced by factors such as competition, corruption, discretion, demand conditions (e.g., elasticities), and quality differences, as discussed previously. Table 3 presents both percentage and level estimates. My results indicate that public procurement entry is associated with a significant markup increase of approximately 12%, corresponding to a level increase between 0.23 and 0.35 (95% confidence interval, calculated using the delta method). When allowing the markup effect to vary with public procurement intensity—by interacting the public procurement dummies with the share of export sales in total sales—the coefficient on public procurement entry is larger, indicating a 14.5% increase. This allows me to trace the markup trajectory over time based on the share of public procurement sales in total sales.

Interpreting My Results I report two major findings: (i) in the cross-section, government contractors have higher markups than their private-sector counterparts within the same industry, and (ii) in the time series, markups increase when firms enter public procurement markets, even after controlling for aggregate demand and supply effects via year dummies. How can these results be explained? Several hypotheses can account for these findings. First, more efficient producers are likely to face more efficient competitors, charge lower prices, sell more in the private market, and outcompete rivals in public procurement tenders. Their cost advantage enables them to set

Table 3: Markups and Public Procurement II: Entry Effect

Entry Effect on Markups	Percentage ($\hat{\gamma}_1$)	Level ($\hat{\mu}_{PP}^{entry}$)
Public Procurement Premium	0.120	0.275
	(0.006)	(0.022)

Note: Estimates are obtained after running equation 3. The $\hat{\mu}_{PP}^{entry}$ standard error is derived from a nonlinear combination of the relevant parameter estimates, using the delta method, and both standard errors are cluster robust. All regressions include variable input, capital and full 47 year and NACE 2-digit division dummies as controls. Firms that enter or exit public procurement markets are not included in the regression: N=5,744. Adjusted R^2 = 0.736.

higher markups under certain conditions regarding the relative efficiency of firms in both private and public procurement markets. These firms also tend to exhibit higher measured productivity. An alternative explanation is that demand elasticities differ in public procurement markets, or that government valuations of goods differ from those in the private sector. The exact mechanism underlying these results cannot be tested with the available data, as I lack firm-specific price information that could help distinguish the markup difference between cost and price effects. However, I demonstrated that even after controlling for cost differences, Czech government contractors still exhibit higher average markups, suggesting that other factors influencing prices are at play. This aligns with recent work by Titl (2023), Baránek and Titl (2024), and Szucs (2024), which highlight the role of competition, discretion, and favoritism in public procurement markets.

In conclusion, my evidence shows that markups differ for government contractors and increase significantly—both economically and statistically—when firms enter public procurement markets.

Markups and Productivity Although not the primary focus of my analysis, I extend the investigation by relating firm-level markups to productivity. The De Loecker and Warzynski (2012) framework provides estimates for both markups and productivity. Specifically, after estimating the production function coefficients, I derive productivity as follows:

$$\hat{\omega}_{it} = \hat{\phi}_{it} - f(v_{it}, k_{it}; \hat{\beta}),$$

where $f(v_{it}, k_{it}; \hat{\beta})$ represents the predicted output based on variable inputs and capital, using the estimated coefficients $\hat{\beta}$. A broad class of industrial organization models predicts that firms with lower marginal costs (i.e., higher productivity) tend to charge higher markups, all else being equal. For instance, in Cournot competition models, more productive firms capture larger market shares and, consequently, set higher markups. I estimate equation 2 again, this time replacing export status with productivity, and find a highly significant and positive coefficient of 0.89 on

productivity. These results align with a wide range of theoretical models, confirming that more productive firms charge higher markups. However, I refrain from further analysis, as productivity estimates may include price or demand variation, making them imperfect measures of marginal costs. The De Loecker and Warzynski (2012) framework could be used to explore the distinct roles of productivity and markups in public procurement entry and exit behavior—an important avenue for future research, though beyond the scope of this paper. It is also important to note that the productivity estimates capture all unexplained variations in total revenue from input factors, potentially including market power effects, as emphasized in public procurement literature. [4]

Heterogeneity and Robustness I discuss several robustness checks below. First, I explore differences in markups for government contractors in the civil engineering sub-industry relative to other construction projects, as well as differences in markups for government contractors classified as sole proprietors compared to companies and cooperatives. Next, I address the role of the decreasing public procurement markup premium in driving the aggregate markup trajectory over time. Lastly, I evaluate the robustness of using deflated sales to proxy for output.

I have shown that government contractors generally have higher markups and that these increase after entering public procurement. However, government contractors operate in different markets, and my markup estimates encompass a mix of market-specific markups. To investigate whether markups differ across construction sub-industries, I leverage firm-specific public procurement project data. Specifically, I revisit the effect of public procurement entry on markups, accounting for contracting intensity, to examine the distinct effects on private-sector and public procurement markups.

In the Czech Republic, public procurement in construction includes building construction, civil engineering, and specialized activities such as electrical and plumbing installations. To determine if markups are higher for contractors involved in complex projects like road and railway construction, I introduce interaction terms for the NACE 2-digit division 42 (civil engineering) in equation 2. I estimate a point coefficient of 0.037 (standard error = 0.01), reflecting approximately a 4% higher markup for government contracts in civil engineering compared to average contracts for building construction and specialized activities. This result aligns with the ”superstar firm” hypothesis (Autor et al., 2020), as civil engineering firms tend to have higher factor intensity (capital and cost of goods sold). Due to data limitations, I cannot test other hypotheses, such as differen-

[4]Both markups and productivity enter significantly in a public procurement entry and exit logit regression, controlling for aggregate effects. This suggests that both play distinct roles in shaping public procurement entry behavior.

tial quality leading to greater markups, which remains a topic for future research. I also examine whether markups differ for government contracts supplied by sole proprietors compared to those supplied by companies or cooperatives. My dataset identifies 12 sole proprietors, and I find a highly significant (t-statistic = 8.09) interaction term of 0.082, suggesting a notable markup premium for sole proprietors. These findings may indicate possible discretion, as documented by Baránek and Titl (2024).

In addition, I interact the public procurement dummy with yearly indicators for 2006–2021, using 2006 as the reference year. The results show significantly lower markup premia over time, with differences in markup premia relative to 2006 growing from around -1% to more than -4% by the end of the period. These findings mirror those in Appendix C, obtained by estimating equation 2 separately for each year. This suggests that the declining public procurement markup premium has played a substantial role in the overall decline in aggregate markups.

Decomposing the Public Procurement Entry Markup Effect Thus far, I have demonstrated that markups increase when firms enter public procurement markets. However, my markup estimates reflect a combination of private-sector and public procurement market markups. While the De Loecker and Warzynski (2012) framework can estimate market-specific markups, my dataset does not include information on hours worked or employees by destination market, which limits my ability to disentangle private-sector and public procurement markups. I instead focus on the share of public procurement sales in total sales, interacting this with the public procurement entry dummy to determine if private-sector markups change upon public procurement entry.

For firms with a small share of sales from public procurement (e.g., less than 1%), I examine whether private-sector markups change. I find a significant coefficient of 0.117 for γ_1, corresponding to a level effect of 0.27, consistent with previous estimates. However, when adjusting for the public procurement sales share, the markup entry effect is minimal for firms selling a small proportion in public procurement markets. For firms with less than 1% of sales from public procurement, markups increase by only 0.12%, suggesting no significant change in private-sector markups.

This approach has limitations. As public procurement sales shares may increase over time, separating private and public procurement markups becomes challenging. Moreover, the assumption that inputs are used in proportion to final sales may not hold, and a more optimal weight should be considered in future research. To estimate market-specific markups by firm, one could introduce a demand system for each market alongside an assumption about the cost function. While the De Loecker and Warzynski (2012) approach avoids the need to specify these assumptions, I can

still compare markups across producers and examine how markups change with public procurement entry over time, without decomposing market-specific markups within a firm.

Unobserved Prices and Revenue Data In estimating output elasticities, I implicitly treat deflated sales as a proxy for physical quantity, potentially introducing omitted variable bias, as discussed by Klette and Griliches (1996). However, in my context, I am less concerned with obtaining precise productivity estimates. As noted by De Loecker and Warzynski (2012), unobserved prices primarily affect productivity estimates. In my case, unobserved prices likely bias output elasticities downward, as increases in input usage generally drive prices down under common demand and cost specifications. This suggests that I may underestimate markups, but this bias predominantly affects level estimates rather than the relationship between markups and public procurement status. To address this issue, I correct markups using a proxy for productivity, $h(\cdot)$, controlling for price variation correlated with productivity. Although demand shocks unrelated to the non-parametric output, including the proxy $\phi_t(\cdot)$[5], may still bias the input coefficient estimates, this does not impact my primary results, as I consistently relate logarithmic markups to public procurement status. I estimate the following regression:

$$\left(\ln \theta_{it}^V - \ln \alpha_{it}^V\right) = \theta_0 + \theta_1 pp_{it} + \nu_{it},$$

where price bias is expected to affect output elasticities. When employing a more flexible production technology, such as the translog, I encounter a trade-off: allowing output elasticities to vary introduces potential bias from unobserved prices. Nevertheless, my average percentage difference in markups remains consistent, provided the difference $\ln \hat{\theta}_{it}^V - \ln \theta_{it}^V$ is not correlated with public procurement status pp_{it}. To mitigate this issue, I control for inputs in all markup regressions. Using the translog production function, I estimate:

$$\hat{\theta}_{it}^V = \theta_{it}^V + \rho(v_{it}, k_{it}),$$

where $\rho(\cdot)$ captures the potential bias stemming from unobserved firm-level price deviations. This strategy is applied consistently throughout my analysis.

[5]Refer to Appendix A for further details on this notation.

3.3 Conclusion

Using firm-level data on publicly traded firms in the Czech construction sector, I analyze the evolution of market power by estimating firm-specific markups and documenting their distributional characteristics. Since 2006, markups have dropped from 40% to nearly 30% in 2021, representing a 10 percentage point reduction. This decline is predominantly driven by firms with the highest initial markups. Over time, the distribution of markups has become less skewed, characterized by a thinner upper tail.

I also establish the link between markups and public procurement using a method that accommodates a wide range of price-setting models while recovering firm-specific markups. This approach eliminates the need for assumptions regarding constant returns to scale or explicit measures of the user cost of capital. Using data on Czech construction firms, I test whether (i) government contractors charge higher markups on average, and (ii) markups change when firms enter or exit public procurement markets. The Czech Republic provides a particularly interesting context, having transitioned from a centrally planned to a market economy with relatively fast GDP growth. However, institutional weaknesses, particularly corruption and inefficiencies in public procurement, have been identified as key impediments to economic progress. The results show significant and robust differences in markups between government contractors and private-sector firms. Government contractors consistently exhibit higher markups, which aligns with the observed productivity premium among these firms. However, these findings raise important questions about the nature of these revenue-based productivity differences. I also provide new econometric evidence that markups increase significantly when firms enter public procurement markets. These findings suggest that government contractors exhibit greater market power, which shifts significantly when firms enter public procurement. Furthermore, examining the heterogeneity in public procurement markup premia, I find that civil engineering projects and sole proprietors benefit from significantly higher pricing power compared to other public procurement construction projects and entities. By linking the fall in aggregate markups to the dynamics of public procurement markups, I provide empirical evidence of the factors driving the overall decline in markups. These results represent a preliminary step toward understanding the relationship between market power and public procurement. They corroborate existing evidence and offer potential explanations for the substantial gains in market power observed when firms become government contractors.

4 Causal Analysis Under Unconfoundedness

In this section, I examine several methods for estimating causal effects of engagement in public procurement under unconfoundedness, closely following Imbens and Xu (2024). The analysis focuses on a binary treatment, with my primary interest in the efficiency of government tenders in the Czech construction sector vis-à-vis private markets during 2006-2021. I begin by briefly introducing the potential outcomes framework and two key assumptions—unconfoundedness and overlap—followed by a discussion of estimation strategies applicable under these assumptions and supplementary analyses, primarily placebo tests, used to validate these assumptions and enhance research credibility.

4.1 Potential Outcomes Framework

I adopt the potential outcome model (Rubin, 1974). For each firm i, for $i = 1, \ldots, N$, two potential outcomes exist: $Y_i(0)$ represents the outcome (log markups) had firm i not participated in public procurement, and $Y_i(1)$ represents the outcome for the same firm had it participated in public procurement. The difference between those two potential outcomes, $\tau_i \equiv Y_i(1) - Y_i(0)$ is the causal effect of the public procurement for that firm. The binary treatment for firm i, participation in public procurement, is denoted by $W_i \in \{0, 1\}$. The realized outcome is $Y_i \equiv Y_i(W_i) = (1 - W_i)Y_i(0) + W_i Y_i(1)$. In addition, firm i's pretreatment characteristics are denoted by X_i. In my case, the basic vector of covariates includes sales, costs, wages, assets, employment category, and indicators for civil engineering and specialized activities construction divisions. I also consider setting where the covariate vector is augmented to include lagged markup variables.

My primary estimand is the average treatment effect for the treated (ATT),

$$ATT \equiv \frac{1}{N_{tr}} \sum_{i:W_i=1} \{Y_i(1) - Y_i(0)\} = \overline{Y}(1) - \overline{Y}(0),$$

where N_{tr} is the number of treated units. In other settings researchers may also be interested in the average treatment effect $ATE \equiv \frac{1}{N} \sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\}$. I allow for treatment effect heterogeneity and focus on the ATT, as it makes less sense to me to contemplate the effect of public procurement for firms in the control group with long-term high markups.

I cannot directly estimate the ATT because I do not observe the control outcomes for the treated units. To make progress, I define my estimand as the covariate adjusted difference between treated

and controls,

$$\mathbb{E}\left[\mathbb{E}[Y_i|W_i=1,X_i]-\mathbb{E}[Y_i|W_i=0,X_i]\mid W_i=1\right].$$

This is an object I can estimate consistently given a random sample. However, it is only under two critical assumptions: unconfoundedness and overlap that this estimand is equal to the *causal estimand*, the ATT (see, e.g. Abadie and Cattaneo (2018) for formal treatment). A separate question concerns the plausibility of the assumptions, which I will address with placebo and sensitivity analyses.

Assumption 1 (Unconfoundedness)**.** *The unconfoundedness assumption, states that, conditional on the covariates, the treatment assignment is independent of the pair of potential outcomes:*

$$W_i \perp\!\!\!\perp \{(Y_i(0), Y_i(1)\} \mid X_i.$$

Identifying the ATT in fact only requires $W_i \perp\!\!\!\perp Y_i(0) \mid X_i$, a weaker version of unconfoundedness. This assumption stands in contrast to traditional econometric assumptions of exogeneity that were articulated in terms of residuals, themselves defined in terms of functional forms. Unconfoundedness separates the functional form part of the assumptions from their essence. Essentially, it is sufficient that researchers understand (a crucial aspect of) the "design," or the treatment assignment mechanism, without full knowledge of the data-generating process of the potential outcomes. A key result in Rosenbaum and Rubin (1983) shows that Assumption 1 implies

$$W_i \perp\!\!\!\perp \{(Y_i(0), Y_i(1)\} \mid e(X_i),$$

in which $e(X_i) \equiv \Pr(W_i = 1 \mid X_i)$ is the propensity score for unit i. This result is important because it reduces the dimension of the conditioning set from the dimension of X_i to one, the dimension of the propensity score.

When the parametric outcome model (the markup equation) is correctly specified, unconfoundedness implies a zero conditional mean for the error term.

Unconfoundedness is a very strong assumption. For a general discussion on this topic, see Imbens (2004). While I acknowledge concerns about its validity in the absence of a clear understanding of the treatment assignment mechanism, I believe that supplementary such as placebo tests and sensitivity analyses, can help assess the plausibility of this assumption and, by doing so, improve the credibility of analyses based on it. I will illustrate these methods in the next section.

In the context of the my data, it is evident that appropriate pretreatment variables should be controlled for. In other cases, whether one should adjust for differences between treated and control units based on specific covariates is less clear (see Cinelli et al. (2022) for a discussion).

Assumption 2 (Overlap)**.** *Etimataing the average effect at every value for the covariates requires overlap, or that the propensity score is between zero and one:*

$$0 < \Pr(W_i = 1 \mid X_i) < 1.$$

If the ATT is of interest, in fact only a weaker overlap assumption, $\Pr(W_i = 1 \mid X_i) < 1$, is required. Overlap is crucial in identifying the ATT when I am unwilling to make functional form assumptions about the conditional means of the potential outcomes. When X_i includes fewer than a handful of covariates, inspecting pairs of the covariates' marginal or joint distributions by treatment status may be sufficient for assessing overlap. However, this approach becomes impractical in high-dimensional settings. In such cases, a more attractive method is to inspect the distribution of the propensity scores, estimated by a flexible method, by treatment status. The lack of overlap in covariate distributions implies, and is implied by, a lack of overlap in the propensity score distributions. Approaches assuming correct functional forms, which allow for interpolation or extrapolation of treatment effects across all covariate levels and their combinations, thereby formally eliminate the need for overlap.

Ensuring overlap or improving balance typically involves dropping some units from the full sample. Although in principle this leads to some loss of information, the improvement in robustness and reduction of bias may outweigh the loss in precision. In practice, concerns about bias suggest that aggressive trimming may lead to more robust and credible estimates. Crump et al. (2009) characterize subsamples optimized for precise average treatment effect estimation, with a rule of thumb suggesting trimming data with estimated propensity scores outside $[0.1, 0.9]$. Another approach, particularly well suited to settings where the focus is on the ATT, is to created a matched sample in which all treated units are matched to a distinct control unit in terms of the estimated propensity score. Beyond ensuring overlap, this method creates a sample that is much better balanced in the covariate distributions. This is particularly true in cases with poor overlap in the raw data, such as my sample, where trimming to ensure overlap is likely to more important than the choice of specific estimation strategies.

4.2 Estimation Given Unconfoundedness and Overlap

Outcome modeling Denote the conditional means of the two potential outcomes as $\mu_w(x) \equiv [Y_i(w)|X_i = x]$, for $w \in \{0, 1\}$. The most commonly used method is a simple linear regression using the treatment indicator and covariates (the level terms) as regressors. The regression method models the conditional means of potential outcomes parametrically, that is, $\mu_0(x) = x^\top \beta$ is a linear function of x, and the treatment effect is constant: $\mu_1(x) = \mu_0(x) + \tau$. Relaxing the functional form assumptions slightly, one can use two separate linear regressions to model $\mu_0(x)$ and $\mu_1(x)$. This estimator is sometimes referred to as the Oaxaca-Blinder estimators (Kline, 2011). In general, one can model the outcome semi- or non-parametrically (eg. Heckman et al. (1998).

Adjusting covariate imbalance The methods focusing directly adjusting covariate imbalance between the treatment and control groups include blocking on covariates, covariate matching, and weighting methods to achieve covariate balance (Zubizarreta et al., 2023). Adjusting for differences in the propensity scores can be implemented through methods such as blocking/matching and reweighting. For example, one of the IPW estimators for the ATT is the Hájek estimator $\hat{\tau}_{IPW} = \sum_{i:W_i=1} Y_i/N_{tr} - \sum_{i:W_i=0} \hat{\omega}_i Y_i$, with weights $\hat{\omega}_i = \hat{e}(X_i)/(1 - \hat{e}(X_i))/\sum_{j:W_j=0} \hat{e}(X_j)/(1 - \hat{e}(X_j))$. Here, $\hat{e}(x)$ is the estimated propensity score.

Doubly robust methods The augmented inverse propensity weighting (AIPW) estimator that combines weighting and regression can be written as:

$$\hat{\tau}_{AIPW} = \frac{1}{N_{tr}} \sum_{i:W_i=1} (Y_i - \hat{\mu}_0(X_i)) - \frac{1}{N_{tr}} \sum_{i:W_i=0} \hat{\omega}_i (Y_i - \hat{\mu}_0(X_i)) \ ,$$

in which $\hat{\omega}_i$ is the IPW (or balancing) weights. They can be viewed as combining an outcome model with an adjustment term, which consists of an IPW estimator applied to the residuals from the outcome model. In recent years, machine learning methods have were introduced for estimating causal effects due to methodological improvements, eg. Athey et al. (2019). Many of these estimators adopt the form of an AIPW estimator and ensures the stability of the moment conditions used to identify the causal parameter against small perturbations in nuisance functions, including the conditional mean $\mu_w(x)$ and the propensity score $e(x)$. Chernozhukov et al. (2018) show estimators with slower convergence rates for the estimators of these nuisance functions.

4.3 Alternative Estimands and Heterogeneous Treatment Effects

Understanding effect heterogeneity is crucial for discerning the mechanisms and impacts of a policy, for a more precise evaluation of policy effectiveness, and for guiding personalized policy assignments. Econometrically, researchers can study heterogeneous treatment effects by estimating the CATT, i.e., $\tau(x) \equiv \frac{1}{N_x} \sum_{i:X_i=x,W_i=1} \tau_i$, in which N_x is the number of treated units whose covariate values equal to x. Another important group of estimands are the quantile treatment effects. They are defined as the difference between the quantiles of the treated and untreated potential outcome distributions for the population or the treated group. Because Assumptions 1 and 2 allow for the identification of the full marginal distribution of $Y_i(0)$ and $Y_i(1)$, quantile treatment effects are identified under those assumptions. One potentially underexplored area is that estimates of CATT or quantile treatment effects can inform the plausibility of the unconfoundedness assumption, given that researchers often possess insights into the range of these effects.

4.4 Validation through Placebo and Sensitivity Analyses

To evaluate the credibility of treatment effect estimates, I rely on placebo and sensitivity analyses.

Placebo analyses I indirectly assess unconfoundedness by formally testing a conditional independence restriction that is testable. This testable assumption differs from unconfoundedness in two aspects. First, it conditions on a subset of the full set of covariates that appear in the unconfoundedness assumption. Second, it uses one of the remaining covariates as a pseudo-outcome that serves as a proxy for the target outcome. I use a placebo test which estimates the treatment effect on a pretreatment variable, known to be unaffected by the treatment. A lagged outcome is appealing as it is typically a good proxy for the target outcome. Imbens (2015) discusses testing additional implications of the conditional independence relationship.

Sensitivity analyses By assumming unconfoundedness only holds conditional on observed covariates X_i and an unobserved confounder U_i, a causal relationship is considered insensitive to unobserved confounding if the estimated effect remains robust against strong dependence with U_i. I follow Imbens (2003), who benchmarks the association between the unobserved U_i and the potential outcomes and treatment with those estimated on observed covariates and introduces contour plots for interpretation.

4.5 The Czech Construction Public Procurement Data

To demonstrate the methodology in practice, I now reanalyze data from the Czech construction sector, building on original structural inference of the underlying markup distribution to evaluate the effect of public procurement engagement between 2010 and 2021 on firms' pricing power. The primary outcome of interest is post-contract markups. This empirical analysis presents a challenge, as I lack detailed information on the treatment assignment mechanism. However, the availability of lagged outcomes allows for validation of the unconfoundedness assumption.

The control group comprises 161 firm-year observations from firms without any government tender revenues between 2010 and 2021, while the treatment group consists of 151 firm-years from firms that received payments from government construction tenders during 2010–2019. These treated firms may include both repeat government contractors and one-time providers. The total contract value amounts to 25.9 billion Czech Crowns in 2021 (1 CZK $\approx$ 21.5 USD in 2021).

Contract Value (CZK)	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**Total**
Sum (Billions)	2.20	0.27	0.60	2.82	1.11	0.84	0.59	1.99	5.05	3.84	2.98	3.60	25.90
No. Contractors	7	8	9	13	11	14	12	17	17	15	15	13	151

I am unable to rely on random assignment to ensure comparability between the treatment and control groups at the time of public procurement entry. However, I will present results based on the assumption that unconfoundedness holds once I condition on a set of observable covariates, including the year in which the contract was won. Notably, I also have data on past markups spanning from 4 years (2006–2009) to 15 years (2006–2020) before the firm won a contract. These historical outcomes can be used either as conditioning variables or as placebo outcomes.

In the subsequent analysis, I will use the natural logarithm of markups from three post-contract periods as the outcome variables. These are denoted as $Y_{i,0}, \ldots, Y_{i,2}$, where $t = 0$ represents the year the contract was awarded. I will treat the log markups from the three years immediately preceding the contract win, i.e., $Y_{i,-3}, Y_{i,-2}, Y_{i,-1}$, as well as their average, as placebo outcomes. Additionally, the log markups from the three years prior to these, i.e., $Y_{i,-6}, Y_{i,-5}, Y_{i,-4}$, will be used as covariates for adjustment. This adjustment will also include a set of time-invariant and pre-contract variables such as financial statement data—specifically sales, cost of goods sold, assets, and number of employees—as well as the year the contract revenue was received and the NACE 2-digit division. I will also include lagged values of the public procurement indicator to compare firms with similar pre-treatment histories of engagement in public procurement.

First, I present some summary statistics.

Table 4: Pretreatment Summary Statistics

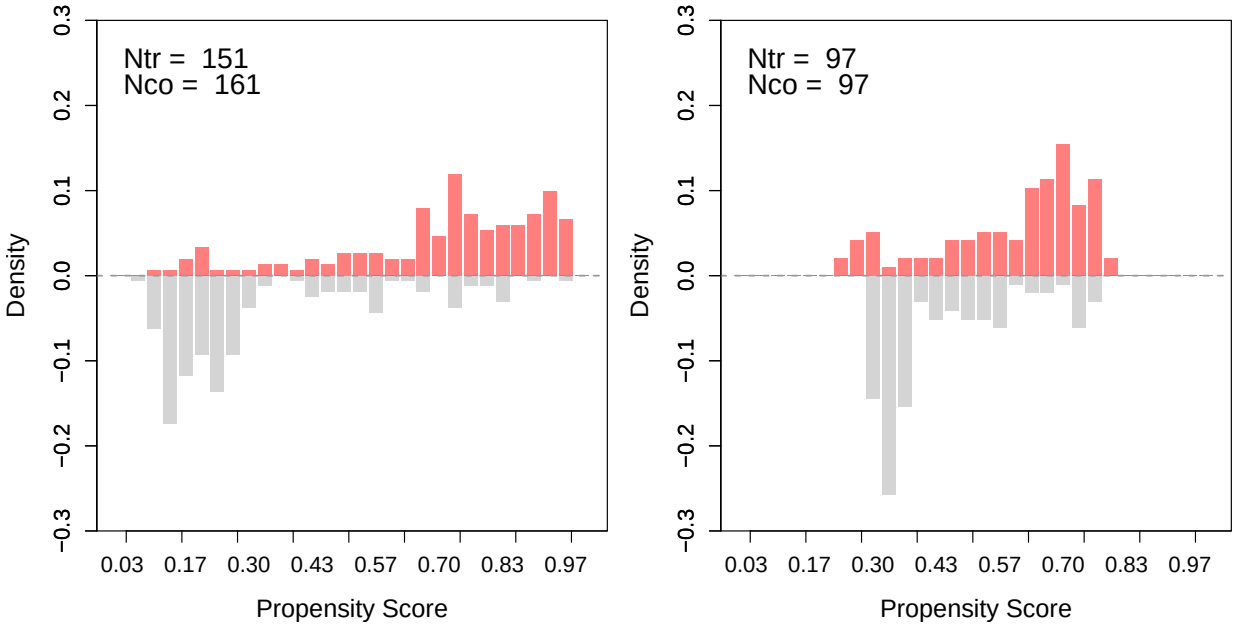
Observations Stratified by Engagement in Public Procurement	$\{i,t;W_{it}=0\}$ N=161 **mean (s. d.)**	$\{i,t;W_{it}=1\}$ N=151 **mean (s. d.)**	**diff / sd**
No. Employees	52.57 (42.03)	104.96 (61.67)	0.993
$\ln(\text{Sales})_{t-4}$	18.52 (1.01)	19.27 (0.85)	0.799
$\ln(\text{Costs})_{t-4}$	17.89 (0.95)	18.52 (0.82)	0.713
$\ln(\text{Markup})_{t-4}$	0.24 (0.13)	0.34 (0.17)	0.640
$\ln(\text{Assets})_{t-4}$	16.05 (1.68)	16.97 (1.37)	0.603
NACE 42	0.04 (0.20)	0.11 (0.32)	0.259
NACE 43	0.40 (0.49)	0.44 (0.50)	0.081

Note: I report observations from firms in a strongly balanced panel, excluding those already active in public procurement in 2006 (N = 26). The period from 2006 to 2009 is unavailable due to matching and placebo requirements of at least four years. This results in $N \times T = 312$ observations during 2010-2021.

$$\text{diff/sd} = \left|\bar{X}_0 - \bar{X}_1\right| \left(\sqrt{(s_0^2 + s_1^2)/2}\right)^{-1}$$

In Table 4, I present the averages and standard deviations for the observations used in the estimation sample. The sample consists of firms observed throughout the entire 16-year period, excluding those already engaged in public procurement as of 2006. I further condition the sample on four lagged values of the pre-treatment variable. The final dataset includes 26 firms, tracked across 10 periods of instantaneous treatment effects between 2010 and 2021, yielding 161 observations from firms exclusively active in the private sector ($i,t;W_{it}=0$) and 151 firm-year observations from firms receiving government contracts ($i,t;W_{it}=1$). I report the average covariate values by treatment status, normalized by their standard deviations. Table 4 shows that the baseline difference in average markups between treated firms (future contractors) and control firms (those only active in the private sector during 2010–2021) amounts to 0.66 standard deviations. I do not report the t-statistic for this difference, as the t-statistic partially reflects sample size. A larger normalized difference, however, signals a more severe overlap issue. Differences exceeding 0.25 standard deviations are in geneal considered substantial (Imbens and Wooldridge, 2009). Estimating the treatment effects essentially involves using the controls to estimate the conditional mean $\mu_0(x) \equiv [Y_i(0)|X_i = x]$, and leveraging this estimated regression function to predict the missing control outcomes for the treated units. Given the substantial disparity in covariate distributions between the two groups—up to 0.99 standard deviations from zero—linear regression is likely to rely heavily on extrapolation, making the results highly sensitive to the functional form of the model.

Figure 3: Assessing Overlap



(a) Full sample (b) Trimmed sample

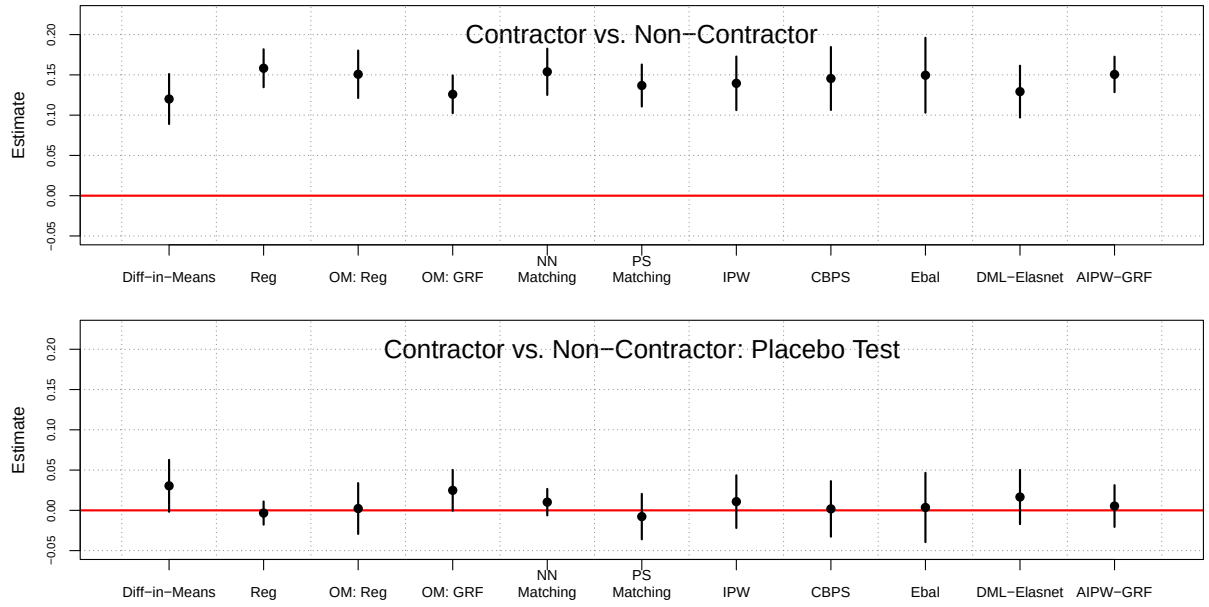
Note: Propensity scores estimated through Generalized Random Forest and re-estimated after trimming in (b).

Figure 3 evaluates the overlap between the treatment and control groups using the covariates described earlier. The figure shows that, although the propensity score distribution for individuals in the treatment groups differs from that of the control group, the propensity scores of the treatment groups still lie within the support of the control group. To improve overlap, I exclude observations with propensity scores exceeding 0.8 (removing 8 controls and 53 contractors) and further refine the control group by applying 1:1 propensity score matching without replacement.

Figure 5 displays the ATT estimates from various estimators on the real outcome (log markup in Year 0) and the placebo outcome (the average log markups from Year -3 to Year -1). The figure shows that different covariate adjustment methods yield consistent results, corroborating the findings from the previous section: participation in public procurement leads to a significant increase in markups, averaging up to 15% across methods. When doubly robust estimators, such as AIPW-GRF, are applied, the estimates for the placebo outcomes are nearly zero, providing further evidence in support of the unconfoundedness assumption.

I then estimate the ATT and CATT for log markups from Year -3 to Year 3 separately, using both difference-in-means and AIPW-GRF methods. The results are presented in Figure 5. This figure resembles an event study plot commonly used in panel data analysis, although my primary identification assumption here is unconfoundedness. The estimates from AIPW-GRF and

Figure 4: ATT Given Unconfoundedness and Placebo Estimates

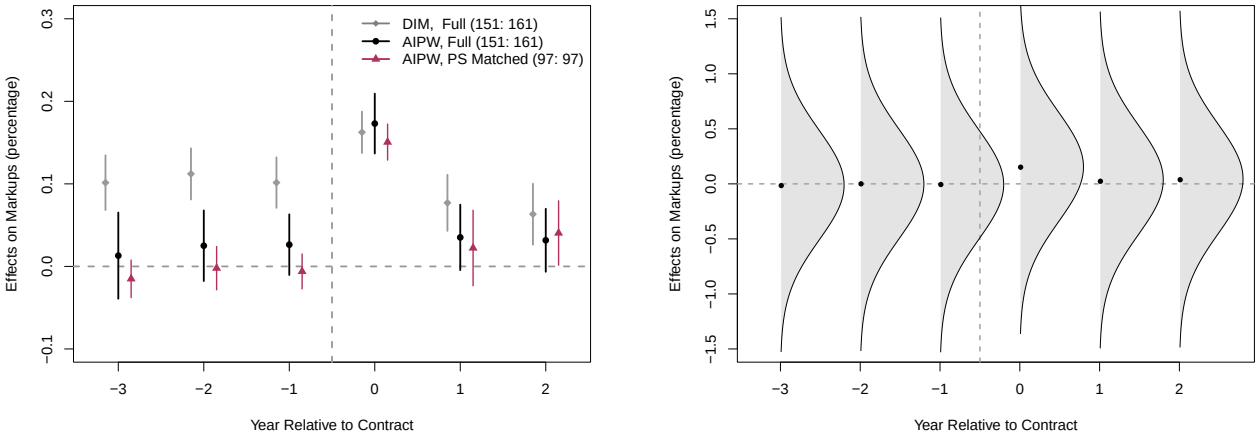


Note: Top figure: ATT (percentage) estimates of government contractors versus non-contractors on the average annual log markups in Year 0 along with their 95% confidence intervals. Bottom figure shows the ATT estimates on the placebo outcome, the average markups from Year –3 to Year –1, along with their 95% confidence intervals. Eleven estimators are employed, including difference-in-means, linear regression, Oaxaca Blinder, GRF as an outcome model, 1:5 nearest neighbor matching with bias correction, propensity score matching, IPW with propensity scores estimated by GRF, CBPS, entropy balancing, double/debiased machine learning with elastic net (implemented using `DoubleML`), and AIPW with GRF (implemented using `grf`).

Effect on Markups	**Contract**	**Pre-Contract Average**
Difference-in-Means	0.12 (0.02)	0.03 (0.02)
Regression	0.16 (0.01)	-0.00 (0.01)
Oaxaca Blinder	0.15 (0.01)	0.00 (0.02)
GRF	0.13 (0.01)	0.03 (0.01)
NN Matching	0.15 (0.01)	0.01 (0.01)
PS Matching	0.13 (0.01)	-0.00 (0.01)
IPW	0.14 (0.02)	0.01 (0.02)
CBPS	0.15 (0.02)	0.00 (0.02)
Entropy Balancing	0.15 (0.03)	-0.00 (0.02)
DML-ElasticNet	0.16 (0.01)	-0.01 (0.01)
AIPW-GRF	0.15 (0.01)	0.00 (0.01)

Note: ATT estimates for the real outcome, markups in Year 0, and placebo outcome, average markups from Year –3 to Year –1, using the trimmed data. Robust standard errors are in parentheses. The outcome is measured in natural logarithms (percentage effects). These estimates are visualized in Figure 4.

Figure 5: ATT and CATT Estimates



(a) ATT (b) CATT

Note: Figures show the ATT and CATT estimates. The outcome variables include markups from 3 years before winning to 6 years after winning. The estimates for pre-contract outcomes serve as placebo tests. **Subfigure A**: the difference-in-means estimator (gray diamonds) and the AIPW-GRF estimator (black solid circles for the original data and red triangles for the trimmed data). **Subfigure B**: the distributions of CATT estimates using AIPW-GRF based on the trimmed sample (with the black dots representing the corresponding ATT estimates).

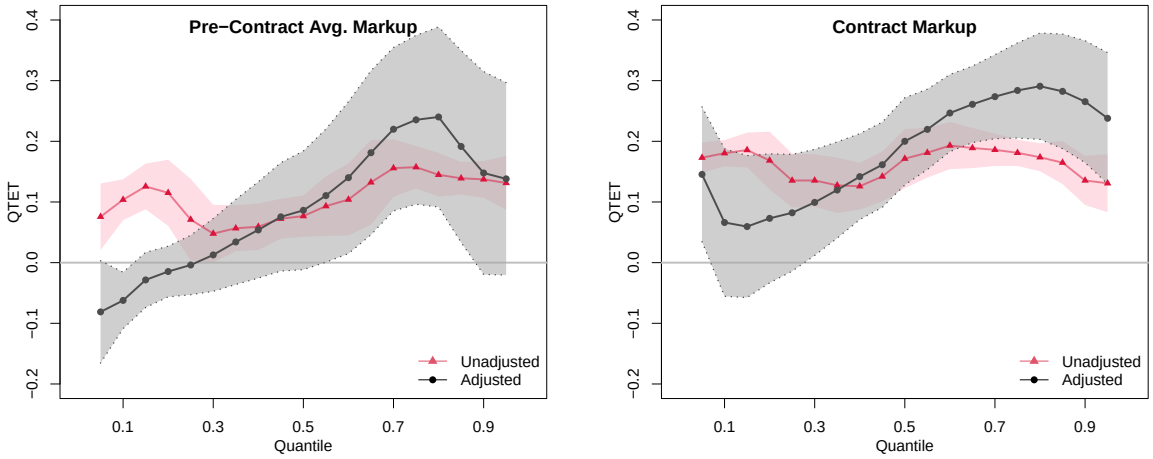
difference-in-means diverge significantly, with the former producing much more credible results. The difference-in-means approach performs poorly in the placebo tests, while AIPW-GRF yields placebo estimates that are nearly zero. These findings not only provide strong corroborative evidence for the placebo tests (i.e., the CATT estimates align closely with zero in the pre-contract years) but also suggest a lack of treatment effect heterogeneity among government contractors. Notably, the distribution of the CATT appears unimodal.

I also estimate the quantile treatment effects on the treated (QTET) using the IPW approach proposed by Firpo (2007). Figure 6 plots the QTET estimates based on the trimmed data.

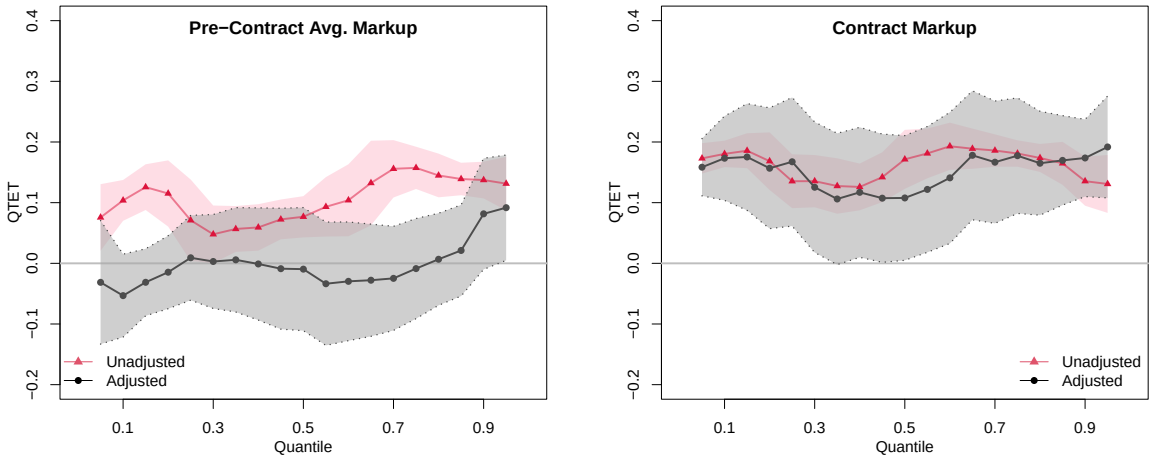
Appendix Figure F presents results from a sensitivity analysis. The estimated causal effect of public procurement remains robust to potential confounders. For instance, the estimate remains positive and substantial (0.157) even when a confounder's correlations with treatment and outcome are tripled relative to those of Y_{-1}.

Overall, I find placebo tests provide strong evidence for the unconfoundedness assumption, enhancing the credibility of the causal estimates. Despite challenges in justifying unconfoundedness due to limited insight into the treatment assignment process, lagged outcomes proved especially valuable, likely capturing both selection and outcome variables, and serving as reliable placebo outcomes. I estimated the propensity score with flexible methods, and assessed overlap through

Figure 6: Quantile Treatment Effects



(a) Contractors vs Controls (Full Sample)



(b) Contractors vs Controls (Trimmed Sample)

Note: Figures show the quantile treatment effects on the treated (QTET) with or without adjusting for the covariates (in grey and pink, respectively). Each dot corresponds to a QTET estimate at a particular quantile, while gray/pink areas represent bootstrapped 95% confidence intervals. Unadjusted models do not incorporate covariates, while adjustment models use the full set of covariates to estimate the propensity scores with a logit. **Subfigure A**: based on the **full sample**. **Subfigure B**: based on the **trimmed sample**.

propensity score distributions. I trimmed the data accordingly and used modern techniques, such as doubly robust estimators, to estimate average causal effects. To explore treatment effect heterogeneity, I considered alternative estimands like the conditional average treatment effect (CATT) and quantile treatment effects (QTET). Placebo tests with pretreatment outcomes validated unconfoundedness, and sensitivity analyses confirmed the robustness of the findings. I followed Imbens and Xu (2024) and benefited from their detailed online tutorial with R code to assist researchers in implementing these methods at `https://yiqingxu.org/tutorials/lalonde`.

5 Causal Panel Data Analysis

In this section I follow Arkhangelsky and Imbens (2024), who survey the recent causal panel data literature. The literature has extended earlier work on difference-in-differences or two-way-fixed-effect estimators. It has more generally incorporated factor models or interactive fixed effects. It has also developed novel methods using synthetic control approaches

The Econometrics Panel Data Literature The earlier econometric panel data literature paid close attention to the dynamics in the outcome process. It distinguished between state dependence and unobserved heterogeneity (Chamberlain, 1984) and various dynamic forms of exogeneity (Engle et al., 1983). The earlier literature also studied models that combined the presence of unit-fixed effects with lagged dependent variables, leading to concerns about biases of least squares estimators in short panels (Nickell, 1981) and the use of instrumental variable approaches (Arellano and Bond, 1991; Richard Blundell and Stephen Bond, 1998). In contrast, an important theme in the current literature that was not discussed as much in the earlier literature concerns the presence of general heterogeneity in causal effects, both over time and across units, associated with observed as well as unobserved characteristics. The recognition of the importance of heterogeneity has led to findings that previously popular estimators are sensitive to the presence of such heterogeneity and to the development of more robust alternatives.

5.1 Data Setup, General Assumptions, Estimands, and Estimators

Consider observations on N units, indexed by $i = 1, \ldots, N$, over T periods, indexed by $t = 1, \ldots, T$. The outcome of interest is denoted by Y_{it}, and the treatment is denoted by W_{it}, both doubly indexed by the unit and time indices. Collect the outcomes and treatment assignments into two $N \times T$ matrices:

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & \ldots & Y_{1T} \\ Y_{21} & Y_{22} & Y_{23} & \ldots & Y_{2T} \\ Y_{31} & Y_{32} & Y_{33} & \ldots & Y_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{N2} & Y_{N3} & \ldots & Y_{NT} \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} W_{11} & W_{12} & W_{13} & \ldots & W_{1T} \\ W_{21} & W_{22} & W_{23} & \ldots & W_{2T} \\ W_{31} & W_{32} & W_{33} & \ldots & W_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_{N1} & W_{N2} & W_{N3} & \ldots & W_{NT} \end{pmatrix},$$

with the rows corresponding to units and the columns corresponding to time periods. One may also observe other exogenous variables, denoted by X_{it} or X_i, depending on whether they vary over time or only by unit. Optimally, the focus is on a balanced panel where for all units $i = 1, \ldots, N$ one observes outcomes for all $t = 1, \ldots, T$ periods. In practice, concerns can arise from the panel being unbalanced either because we observe units for different lengths of time or because data is missing for some of them. In the most general case the treatment may vary both across units and over time, with units switching in and out of the treatment group:

$$\underset{(\text{general})}{\mathbf{W}^{\text{gen}} =} \begin{pmatrix} 1 & 1 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & 1 & \ldots & 0 \\ 1 & 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & 0 & \ldots & 0 \end{pmatrix}$$

With this type of data, we can use variation of the treatment within units and variation of the treatment within time periods to identify causal effects. Especially in settings without dynamic effects, the presence of both types of variation may improve the credible dynamic treatment effects are present, yet, assuming their absence without justification leads to difficult-to-interpret results.

The recent DID/TWFE literature has focused on the staggered adoption case where units remain in the treatment group once they adopt the treatment, but they vary in the time at which they adopt the treatment. Some may adopt early, while others adopt later:

$$\underset{(\text{staggered adoption})}{\mathbf{W}^{\text{stag}} =} \begin{pmatrix} 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & 0 & 0 & \ldots & 1 & 1 \\ 0 & 0 & 0 & 1 & \ldots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & 1 & 1 & \ldots & 1 & 1 \end{pmatrix}$$

This case is also referred to as the absorbing treatment setting. Clearly, this setting leads to much richer information about the possible presence of dynamic effects, with the ability, under some assumptions, to separate dynamic effects from heterogeneity across calendar time.

Potential Outcomes, General Assumptions, and Estimands I use the use the potential outcome notation for N units and T periods. Let $\underline{\mathbf{w}}$ denote the full T-component column vector of treatment assignments,

$$\underline{\mathbf{w}} \equiv (w_1, \ldots, w_T)^\top,$$

and $\underline{\mathbf{W}}_i$ the vector of treatment values for unit i. Let $\underline{\mathbf{w}}^t$ the t-component column vector of treatment assignments up to time t:

$$\underline{\mathbf{w}}^t \equiv (w_1, \ldots, w_t)^\top,$$

so that $\underline{\mathbf{w}}^T = \underline{\mathbf{w}}$, and similar for $\underline{\mathbf{W}}_i^t$. In general we can index the potential outcomes for unit i in period t by the full T-component vector of assignments $\underline{\mathbf{w}}$:

$$Y_{it}(\underline{\mathbf{w}}).$$

Even this notation already makes a key assumption, the Stable Unit Treatment Value Assumption, or SUTVA (see Rubin, 1978; Imbens and Rubin, 2015). SUTVA requires that there is no interference or spillovers between units. This is a strong assumption, and in many applications, it may be violated. The recent causal panel data literature does not emphasize allowing for such interference.

Without further restrictions, this setup describes for each unit and each time period 2^T potential outcomes, as a function of multi-valued treatment $\underline{\mathbf{w}}$. As a result one can define for every period t unit-level treatment effects for every pair of assignment vectors $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$:

$$\tau_{it}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'} \equiv Y_{it}(\underline{\mathbf{w}}') - Y_{it}(\underline{\mathbf{w}}),$$

with the corresponding population average effect defined as

$$\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'} \equiv \mathbb{E}\left[Y_{it}(\underline{\mathbf{w}}') - Y_{it}(\underline{\mathbf{w}})\right].$$

These unit-level and average causal effects are the basic building blocks of many of the estimands considered in the literature. A key challenge is that there are many, $2^{T-1} \times (2^T - 1)$ to be precise, distinct average effects of the form $\tau_{t}^{\underline{\mathbf{w}},\underline{\mathbf{w}}'}$. Even with $T = 2$ there are already six different average causal effects, and with T larger, this number quickly increases. This means that in practice there is a need to limit or focus on summary measures of all these causal effects, *e.g.,* averages over effects at different times.

Assumption 3. (NO ANTICIPATION) *The potential outcomes satisfy*

$$Y_{it}(\underline{\mathbf{w}}) = Y_{it}(\underline{\mathbf{w}}'),$$

for all i, and for all combinations of t, $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$ such that $\underline{\mathbf{w}}^t = \underline{\mathbf{w}}'^t$.

With experimental data, one can compare outcomes in period t for units with the same treatment path up to and including t, but whose treatment paths diverge in the future. The average difference between such average outcomes should be zero in expectation under the no-anticipation assumption.

In observational studies the assumption need not hold. One remedy for this problem is to allow for limited anticipation, assuming the treatment can be anticipated for a fixed number of periods. Algorithmically, this amounts to redefining $\underline{\mathbf{w}}$ by shifting it by the fixed number of periods. The no anticipation assumption reduces the total number of potential treatment effects from $2^{T-1} \times (2^T - 1)$ to $(\sum_{t=1}^{T} 2^{t-1})(\sum_{t=1}^{T} 2^t - 1)$. The basic building blocks, unit-period specific treatment effects, are now of the type

$$\tau_{it}^{\underline{\mathbf{w}}^t, \underline{\mathbf{w}}^{t'}} \equiv Y_{it}(\underline{\mathbf{w}}^{t'}) - Y_{it}(\underline{\mathbf{w}}^t),$$

with the potential outcomes for period t indexed by treatments up to period t only. This current structure still allows us to distinguish between static treatment effects, *i.e.,* $\tau_{it}^{(\underline{\mathbf{w}}^{t-1},0),(\underline{\mathbf{w}}^{t-1},1)}$, which measures the response of current outcome to the current treatment, holding the past ones fixed, and dynamic ones, i.e., $\tau_{it}^{(\underline{\mathbf{w}}^{t-1},w^t),(\underline{\mathbf{w}}'^{t-1},w^t)}$, which does the opposite. A large literature in biostatistics on dynamic models is also relevant for these problems, e.g. Robins et al. (2000).

A stronger assumption is that the potential outcomes only depend on the contemporaneous assignment, ruling out dynamic effects of any type.

Assumption 4. (NO DYNAMIC / CARRY-OVER EFFECTS) *The potential outcomes satisfy*

$$Y_{it}(\underline{\mathbf{w}}) = Y_{it}(\underline{\mathbf{w}}'),$$

for all i and for all combinations of t, $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}'$ such that $w_{it} = w'_{it}$.

It restricts the treatment effects and, thus, the potential outcomes for the post-treatment periods. If one is willing to make the no-dynamic effects assumption, the potential outcomes can be written as $Y_{it}(0)$ and $Y_{it}(1)$ with a scalar argument. In this case, the total number of treatment effects for

each unit is greatly reduced to T (one per period), and we can simplify them to

$$\tau_{it} \equiv Y_{it}(1) - Y_{it}(0),$$

where τ_{it} has no superscripts because there are only two possible arguments of the potential outcomes, $w \in \{0, 1\}$.

A conceptually different assumption is that of absorbing treatments, that is where the assignment mechanism corresponds to staggered adoption.

Assumption 5. (STAGGERED ADOPTION)

$$W_{it} \geq W_{it-1} \quad \forall t = 2, \ldots, T.$$

Defining the adoption date A_i as the date of the first treatment, $A_i \equiv T+1-\sum_{t=1}^{T} W_{it}$ for units that are treated in the sample, and $A_i \equiv \infty$ for never-treated ones. In the staggered adoption case, we can write the potential outcomes, again with some abuse of notation, in terms of the adoption date, $Y_{it}(a)$, for $a = 1, \ldots, T, \infty$, and the realized outcome as $Y_{it} = Y_{it}(A_i)$.

Given staggered adoption but absent the no-anticipation and no-dynamics assumptions, write the building blocks as

$$\tau_{it}^{a,a'} \equiv Y_{it}(a') - Y_{it}(a),$$

with the corresponding population average

$$\tau_{t}^{a,a'} \equiv \mathbb{E}\left[Y_{it}(a') - Y_{it}(a)\right].$$

Define the average for subpopulations by the adoption date:

$$\tau_{t|a''}^{a,a'} \equiv \mathbb{E}\left[Y_{it}(a') - Y_{it}(a) | A_i = a''\right].$$

This estimand is conceptually similar to the average effect on the treated in cross-sectional settings, with the important difference that selection now operates over two dimensions: units and periods.

The Two-Way-Fixed-Effect Characterization I start with the Two-Way-Fixed-Effect (TWFE) specification in a panel setting with no anticipation and no dynamics and parallel trends or constant treatment effects.

Assumption 6. (THE TWO-WAY-FIXED-EFFECT MODEL) *The control outcome $Y_{it}(0)$ satisfies a two-way-fixed-effect structure:*

$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}.$$

The unobserved component ε_{it} is (mean-)independent of the treatment assignment W_{it}.

Assumption 7. (PARALLEL TRENDS ASSUMPTION) *The potential outcomes satisfy*

$$Y_{it}(1) = Y_{it}(0) + \tau \quad \forall (i,t).$$

The combination of these two assumptions leads to a model for the realized outcome, defined as $Y_{it} \equiv W_{it}Y_{it}(1) + (1 - W_{it})Y_{it}(0)$,

$$Y_{it} = \alpha_i + \beta_t + \tau W_{it} + \varepsilon_{it}.$$

One can estimate the parameters of this model by least squares:

$$(\hat{\tau}^{\text{TWFE}}, \hat{\alpha}, \hat{\beta}) = \arg\min_{\tau,\alpha,\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \alpha_i - \beta_t - \tau W_{it}\right)^2.$$

while imposing one restriction on the α_i or β_t (*e.g.,* fixing one of the α_i or one of the β_t equal to zero) to avoid perfect collinearity of the regressors, but this normalization does not affect the value for the estimator of the parameter of interest, τ. The fundamental justification for the TWFE estimator, in one form or another, is based on a parallel trend assumption. This states that, in one form or another, the units who are treated would have followed, in the absence of the treatment, a path that is parallel to the path followed by the control units, in an average sense.

The Staggered Adoption Case Let $A_i \equiv T + 1 - \sum_{t=1}^{T} W_{it}$ be the adoption date (the first time unit i is treated if a unit is ever treated), with the convention that $A_i \equiv \infty$ for units who never adopt the treatment, and N_a is the number of units with adoption date $A_i = a$. Define also the average treatment effect by time and adoption date,

$$\tau_{t|a} \equiv \mathbb{E}\left[Y_{it}(1) - Y_{it}(0) \middle| A_i = a\right].$$

The key is that these average treatment effects can vary both by time and by adoption date. Goodman-Bacon (2021) decomposes the TWFE estimator as follows. Define for all time-periods t and all adoption dates a the average outcome in period t for units with adoption date a:

$$\overline{Y}_{t|a} \equiv \frac{1}{N_a} \sum_{i:A_i=a} Y_{i,t}.$$

Then, for all pairs of time periods $t > t'$ and pairs of adoption dates a, a' such that $t' < a \leq t$ (units with adoption date a change treatment between t and t') and either $a' \leq t'$ or $t < a'$ (units with adoption date a' do not change treatment status between t and t', they are either already treated before period t', or only adopt the treatment after period t), define the following double difference that is the building block for the TWFE estimator:

$$\hat{\tau}_{t,t'}^{a,a'} \equiv \left(\overline{Y}_{t|a} - \overline{Y}_{t'|a}\right) - \left(\overline{Y}_{t|a'} - \overline{Y}_{t'|a'}\right)$$

The interpretation of this double difference plays a key role in the interpretation of the TWFE estimator $\hat{\tau}$. The group with adoption date a changes treatment status between periods t' and t, so the difference $\overline{Y}_{t|a} - \overline{Y}_{t'|a}$ reflects a change in treatment but this treatment effect is contaminated by the time trend in the control outcome under the TWFE structure. For the group with an adoption date a', the difference $\overline{Y}_{t|a'} - \overline{Y}_{t'|a'}$ does not capture a change in treatment status. If $t < a'$, it is a difference in average control outcomes, and $\hat{\tau}_{t,t'}^{a,a'}$ is a standard DID estimand, which under the TWFE model for the control outcomes has an interpretation as an average treatment effect. However, if $a' < t'$, the difference $\overline{Y}_{t|a'} - \overline{Y}_{t'|a'}$ is a difference in average outcomes under the treatment. In the presence of treatment effect heterogeneity, and in the absence of a TWFE model for the outcomes under treatment, it is a weighted average of treatment effects, with the weights adding up to one but with some of the weights negative. The TWFE estimator $\hat{\tau}^{\text{TWFE}}$ can be characterized as a linear combination of the building blocks $\hat{\tau}_{t,t'}^{a,a'}$, including those where the non-changing group has an early adoption date $a' < t'$. The coefficients in that linear combination depend on various aspects of the data, including the number of units N_a in each of the corresponding adoption groups.

Alternative DID-type Estimators for the Staggered Adoption Setting To deal with the negative weights, researchers have recently proposed a number of different modifications to the TWFE estimator. They maintain the TWFE assumption for the control outcomes but avoid the additional assumption on treatment effect heterogeneity.

Callaway and Sant'Anna (2020) propose two ways of dealing with the negative weights. Their first approach takes a group with adoption date a, and compares average outcomes in any post-adoption period $t \geq a$ ($\overline{Y}_{t|a}$ for $t \geq a$) to average outcomes for the same group (the group with adoption date a) immediately prior to the adoption ($\overline{Y}_{a-1|a}$). It then subtracts the difference in outcomes for the same two time periods for the single group that never adopts the treatment. Callaway and Sant'Anna (2020) further suggest using as an alternative control group the average of the groups that do adopt the treatment, but restricting this to those who adopt after period t.

Sun and Abraham (2020) suggest reporting the unweighted average of the same building blocks as Callaway and Sant'Anna (2020) over the periods t after the first period, with the weights within a period equal to the fraction of units with an adoption date prior to that, excluding first period adopters. An additional issue emphasized by Sun and Abraham (2020) is related to the validation of the two-way model. Sun and Abraham (2020) show that common implementation of testing for parallel trends using pre-treatment data using two-way specifications with leads of treatments also include comparisons with negative weights. As a result, they caution against such procedures.

de Chaisemartin and d'Haultfœuille (2020) deal with the negative weights by focusing on one-period ahead double differences, with control groups that adopt later ($a > t$). They aggregate these by averaging over all groups that adopt later and then they average over the time periods, weighted by the fraction of adopters in each period. One challenge with the de Chaisemartin and d'Haultfœuille (2020) approach is that by limiting the comparisons to those that are separated by a single period, the standard errors may be large relative to those for estimators based on more comparisons. Although the additivity assumption may be more likely to hold over such short horizons, there is also increased sensitivity to the presence of dynamic effects.

Borusyak et al. (2021) focus on a model for the baseline outcomes that is richer than the TWFE model:

$$Y_{it}(0) = A_{it}^{\top}\lambda_i + X_{it}^{\top}\delta + \epsilon_{it}$$

where A_{it} and X_{it} are observed covariates, leading to a factor-type structure. This setup reduced to the TWFE for $A_{it} \equiv 1$ and $X_{it} \equiv (\mathbf{1}_{t=1}, \ldots, \mathbf{1}_{t=T})$. They propose estimating λ_i and δ by least squares using only observations for control units only, and later construct unit-time specific imputations for the unobserved control outcomes for the treated units, leading to unit/period-specific treatment effect estimates:

$$\hat{\tau}_{it} = Y_{it} - A_{it}^{\top}\hat{\lambda}_i + X_{it}^{\top}\hat{\delta}.$$

5.2 Robust Confidence Set and Sensitivity Analysis

Rambachan and Roth (2023) evaluate the robustness of the significant results under different extents of the parallel trends assumption (7). Suppose the dynamic treatment effects β in pretreatment placebo and posttreatment periods can be expressed as

$$\beta = \underbrace{\begin{pmatrix} 0 \\ \tau_{\text{post}} \end{pmatrix}}_{=:\tau \text{ treatment effects}} + \underbrace{\begin{pmatrix} \delta_{\text{placebo}} \\ \delta_{\text{post}} \end{pmatrix}}_{=:\delta \text{ difference in trends}}$$

where τ_{post} represents the treatment effects of interest and δ represents the difference in trends between the treated and comparison groups that would have occurred absent treatment. Consider the restriction that the posttreatment violation of parallel trends between consecutive periods is no more than $\bar{M}$ times the maximum difference in trends between consecutive placebo periods, i.e,

$$\delta \in \Delta^{RM}(\bar{M}) = \left\{ \delta : \forall t \geq 0, |\delta_{t+1} - \delta_t| \leq \bar{M} \cdot \max_{s \in \{-2,-1\}} |\delta_{s+1} - \delta_s| \right\}.$$

The ATT of interest can be expressed as $\theta = l'_{att}\delta_{post}$, where $l_{att} = [\frac{n_1}{\sum_{t=1}^{T_{post}} n_t}, \cdots, \frac{n_{T_{post}}}{\sum_{t=1}^{T_{post}} n_t}]'$ and n_t represents the number of observations at the posttreatment period t. Define the set of values for θ that are consistent with a given value of β under the restriction $\delta \in \Delta^{RM}(\bar{M})$ as $S(\beta, \Delta^{RM}) = [\theta^{lb}(\beta, \Delta^{RM}), \theta^{ub}(\beta, \Delta^{RM})]$ where

$$\theta^{lb}(\beta, \Delta^{RM}(\bar{M})) := l'_{att}\beta_{\text{post}} - \left(\max_{\delta} l'_{att}\delta_{\text{post}}, \text{ s.t. } \delta \in \Delta^{RM}(\bar{M}), \delta_{\text{placebo}} = \beta_{\text{placebo}} \right)$$
$$\theta^{ub}(\beta, \Delta^{RM}(\bar{M})) := l'_{att}\beta_{\text{post}} - \left(\min_{\delta} l'_{att}\delta_{\text{post}}, \text{ s.t. } \delta \in \Delta^{RM}(\bar{M}), \delta_{\text{placebo}} = \beta_{\text{placebo}} \right)$$

Under a finite sample normal approximation for the estimated dynamic treatment effects $\hat{\beta}$, i.e., $\hat{\beta} \approx N(\tau + \delta, \Sigma_n)$, Rambachan and Roth (2023) define the confidence sets of θ, $\mathcal{C}_n\left(\hat{\beta}_n, \Sigma_n\right)$ by

$$\inf_{\delta \in \Delta^{RM}(\bar{M}), \tau} \inf_{\theta \in \mathcal{S}(\delta+\tau, \Delta^{RM}(\bar{M}))} \mathbb{P}_{\hat{\beta}_n \sim \mathcal{N}(\delta+\tau, \Sigma_n)} \left(\theta \in \mathcal{C}_n\left(\hat{\beta}_n, \Sigma_n\right) \right) \geq 1 - \alpha.$$

To assess the robustness of the significant estimated ATT, determine the threshold value of $\bar{M}$ where the corresponding confidence interval $\mathcal{C}_n(\hat{\beta}_n, \hat{\Sigma}_n)$ just includes 0.

5.3 Relaxing the Two-Way Fixed Effect Structure

A key strand of the recent causal panel literature starts with the introduction of the Synthetic Control (SC) method (Abadie and Gardeazabal, 2003). The SC literature focused on imputing missing potential outcomes by creating synthetic versions of the treated units constructed as convex combinations of control units. This more algorithmic, as opposed to model-based, approach has inspired much new research, ranging from factor-model approaches that motivate synthetic-control type algorithms to hybrid approaches that link synthetic control methods to the earlier TWFE methods and highlight their connections.

Matrix Completion Methods and Factor Models A set of methods that relaxes the TWFE assumptions focuses directly on factor models, where the outcome is assumed to have the form

$$Y_{it}(0) = \sum_{r=1}^{R} \alpha_{ir}\beta_{tr} + \varepsilon_{it}.$$

First, note that this generalizes the TWFE specification: fix the rank at $R = 2$, and set $\alpha_{i2} = 1$ for all i and $\beta_{t1} = 1$ for all t, this is identical to the TWFE specification.

Athey et al. (2021) take an approach that models the entire matrix of potential control outcomes as

$$Y_{it}(0) = L_{it} + \alpha_i + \beta_t + \varepsilon_{it},$$

where the ε_{it} is random noise, uncorrelated with the other components. The matrix $\mathbf{L}$ with typical element L_{it} is a low-rank matrix. As mentioned above the unit and time components α_i and β_t could be subsumed in the low-rank component as they on their own form a rank-two matrix, but in practice it improves the performance of the estimator substantially to keep these fixed effect components in the specification separately from the low-rank component $\mathbf{L}$. The reason is that we regularize the low rank component $\mathbf{L}$, but not the individual and time components.

Athey et al. (2021) propose the Nuclear-Norma-Matrix-Completion (NNMC) estimator based on minimizing

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1 - W_{it})\left(Y_{it} - L_{it} - \alpha_i - \beta_t\right)^2 + \lambda\|\mathbf{L}\|_*,$$

over $\mathbf{L}$, α, and β. The missing $Y_{it}(0$ values are then imputed using the estimated parameters. Here the nuclear norm $\|\mathbf{L}\|_*$ is the sum of the singular values $\sigma_l(\mathbf{L})$ of the matrix $\mathbf{L}$, based on the

singular value decomposition $\mathbf{L} = \mathbf{S}\Sigma\mathbf{R}$, where $\mathbf{S}$ is $N \times N$, Σ is the $N \times T$ diagonal matrix with the singular values and $\mathbf{R}$ is $T \times T$. The penalty parameter λ is chosen through out-of-sample crossvalidation. The nuclear norm regularization shrinks towards a low rank estimator for $\mathbf{L}$, similar to the way LASSO shrinks towards a sparse solution in linear regression.

Xu (2017) studied direct estimation of factor models as an alternative to synthetic control methods, where the number of factors is estimated or pre-specified and the model is directly estimated after some normalization (building on the factor model literature in econometrics, (e.g., Bai (2009)). Based on this model on can impute the missing potential outcomes for the treated unit/time-period pairs and use that to estimate the average effect for the treated.

5.4 Hybrid Methods

Two recent methods combine some of the benefits from the synthetic control approach with either TWFE ideas or with unconfoundedness methods. These methods are particularly attractive because the nest TWFE, while being able to accomodate more flexible outcome models. There are in essence two approaches. One can directly generalize the outcome model, or one can use a local version of the TWFE model. This is somewhat similar to the way one can generalize a linear regression model by making the regression function more flexible through the inclusion of additional function of the regressors, or by estimating it locally through kernel methods.

Synthetic Difference In Differences Consider the SC estimator for the treatment effect, characterized as a weighted least squares regression,

$$\min_{\beta,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\omega}_i \left(Y_{it} - \beta_t - \tau W_{it}\right)^2.$$

Contrast this with the TWFE estimator,based on a slightly different least squares regression:

$$\min_{\beta,\alpha,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \alpha_i - \beta_t - \tau W_{it}\right)^2.$$

The two differences are that the SC regression uses weights $\hat{\omega}_i$, and that the TWFE regression in has unit-specific fixed effects α_i. Arkhangelsky et al. (2021) exploit the omission of the unit fixed effects from the synthetic control regression by proposing the Synthetic Difference In Difference (SDID) estimator that includes both the unit fixed effects α_i and the SC weights $\hat{\omega}_i$, as well as

analogous time weights $\hat{\lambda}_t$, leading to

$$\min_{\beta,\alpha,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\omega}_i \hat{\lambda}_t \left(Y_{it} - \alpha_i - \beta_t - \tau W_{it}\right)^2 .$$

The time weights $\hat{\lambda}_t$ are calculated in a way similar to the unit weights,

$$\min_{\lambda} \sum_{i=1}^{N-1} \left(Y_{iT} - \sum_{s=1}^{T-1} \lambda_s Y_{is} \right)^2 ,$$

subjet to the restriction that $\lambda_s \geq 0$, and $\sum_{s=1}^{T-1} \lambda_s = 1$. The weights for treated units and periods are equal to 1.

Augmented Synthetic Control Ben-Michael et al. (2021) augment the SC estimator by regressing the outcomes in the treatment period on the lagged outcomes using data for the control units. Suppose that, one uses ridge regression for this first step, again in the setting with unit N and period T the only treated unit/time-period pair:

$$\hat{\eta} = \arg\min_{\eta} \sum_{i=1}^{N-1} \left(Y_{iT} - \eta_0 - \sum_{s=1}^{T-1} \eta_s Y_{is} \right)^2 + \lambda \sum_{s=1}^{T-1} \eta_s^2,$$

with ridge parameter λ chosen through cross-validation. A standard unconfoundedness approach would predict the potential control outcome for the treated unit/time period pair as

$$\hat{Y}_{NT} = \hat{\eta}_0 + \sum_{s=1}^{T-1} \hat{\eta}_s Y_{Ns}.$$

The augmented SC estimator modifies this by combining it with SC weights in a way that can be seen either as a bias-adjustment to the unconfoundedness estimator, or a bias-adjustment to the SC estimator:

$$\begin{aligned} \hat{Y}_{NT} &= \hat{\eta}_0 + \sum_{s=1}^{T-1} \hat{\eta}_s Y_{Ns} + \sum_{i=1}^{N-1} \omega_i \left(Y_{iT} - \hat{\eta}_0 - \sum_{s=1}^{T-1} \hat{\eta}_s Y_{is} \right) \\ &= \sum_{i=1}^{N-1} \omega_i Y_{iT} + \sum_{s=1}^{T-1} \hat{\eta}_s \left(Y_{Ns} - \sum_{j=1}^{N-1} \omega_j Y_{js} \right). \end{aligned}$$

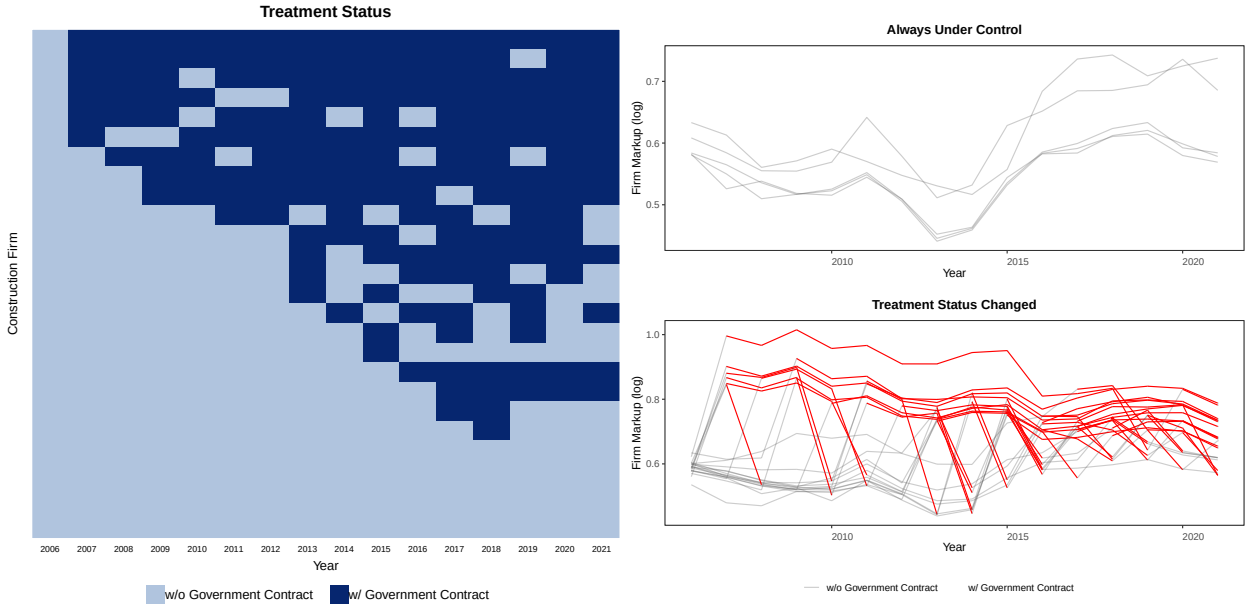
Ben-Michael et al. (2022) extend this approach to the case with staggered adoption.

5.5 Application to Markups and Public Procurement

My goal is to study whether switching from private sector contracts to public sector contracts increases markups for construction firms in the Czech Republic using a generalized DID design. The outcome variable is the natural logarithm of the markup of firm i during year t. The strongly balanced panel dataset consists of 26 Czech construction firms over 16 years, from 2006 to 2021.

My analysis encompasses five parts: (1) fundamental summary and visualization, (2) point estimates, (3) dynamic treatment effects, (4) diagnostic tests, and (5) sensitivity analyses.

Figure 7: Visualizing Treatment and Outcome in the Balanced Panel



Note: I visualize the treatment status using the package `panelView` (Mou et al., 2023). The treated observations are visually represented by deep blue, whereas observations under control status are indicated by a lighter shade of blue. Using the same package, I also depict the trajectory of the outcome variable within the study's time window for each individual unit in the balanced panel. Control units in gray, treated units are represented by red.

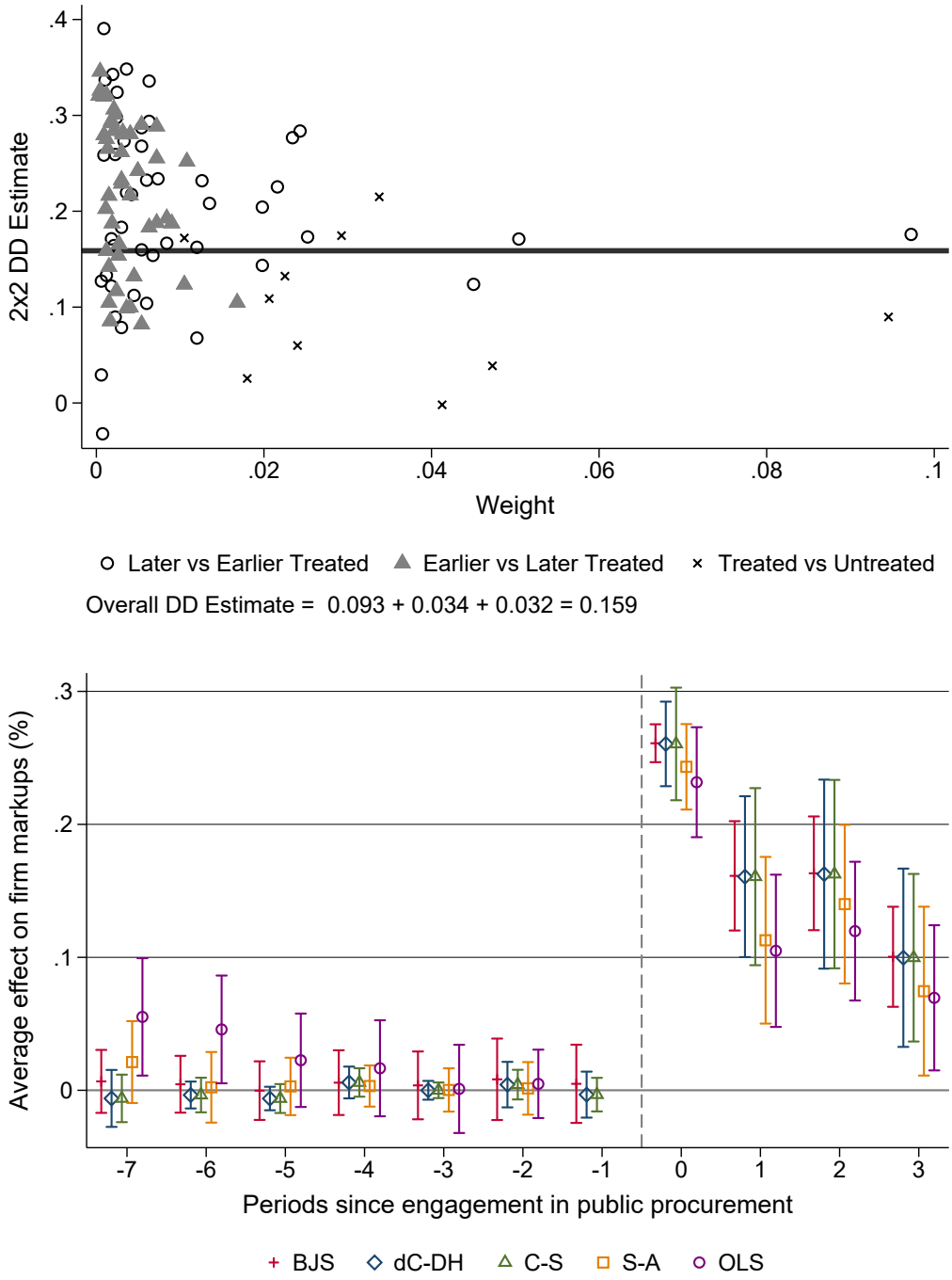
To begin, I visualize the patterns of treatment and outcome variables using plots, which are shown in Figure 7. In this application, treatment reversals clearly take place. I focus on the balanced panel without always treated observations (26 firms over 16 periods).

First, I redefine the treatment as a dummy variable indicating whether the firm ever received sales from government tenders, i. e. following a staggered treatment pattern, to obtain the Goodman-Bacon (2021) decomposition and to compare the dynamic OLS estimates with alternative DID-type estimators robust to heterogeneity in the effect of public procurement on markups across time and cohorts under the staggered adoption setting. The results are shown in Figure 8.

The staggered adoption setup allows me to implement several estimators to obtain the treatment effect estimates. I first estimate a two-way fixed-effects (TWFE) model. The estimated coefficient using TWFE is 0.159, with a standard error of 0.023. Goodman-Bacon (2021) demonstrates that the two-way fixed-effects (TWFE) estimator in a staggered adoption setting can be represented as a weighted average of all possible 2x2 difference-in-differences (DID) estimates between different cohorts. However, when treatment effects change over time heterogeneously across cohorts, the "forbidden" comparisons that use post-treatment data from early adopters as controls for late adopters may introduce bias in the TWFE estimator. The decomposition shows that the estimates from the DIDs comparing comparing ever-treated cohorts switching into treatment and other ever-treated cohorts that are still in their pretreatment periods (the triangles labeled "Earlier vs Later Treated") contribute the least to the TWFE estimate (weight 0.17). The DIDs comparing ever-treated cohorts switching into treatment with the never-treated (crosses labeled "Treated vs Untreated") rank second in terms of its contribution (weight 0.34). The "forbidden" DIDs comparing ever-treated cohorts switching into treatment and other ever-treated cohorts that are already treated (the circles labeled "Later vs Earlier Treated") receive the most weight, 0.49. Both of the timing treated groups have average estimates around 0.19, while the average DID estimate comparing treated to untreated is around 0.09.

I first use the dynamic TWFE regression, which includes a series of interaction terms between a dummy that indicates whether a unit is a treated unit and each lead (lag) indicator relative to the treatment in a TWFE regression. This specification allows for effects to vary over time and usually uses the period immediately preceding a switch into the treatment as the reference period. If the regression is saturated and there is no heterogeneous treatment effects across cohorts, this dynamic TWFE can consistently estimate the dynamic treatment effect. Additionally, to further address the issue of heterogeneous treatment effects across cohorts, I use methods specifically designed to handle heterogeneous treatment effects, such as the methods outlined in Borusyak et al. (2021), de Chaisemartin and d'Haultfœuille (2020), Callaway and Sant'Anna (2020), and Sun and Abraham (2020). In short, results from the heterogeneity robust estimator are substantively similar to those from conventional TWFE models. However, the issue emphasized by Sun and Abraham (2020) is apparent in the visual validation of the two-way model. The common implementation of testing for parallel trends using pre-treatment data using two-way specifications with leads of treatments also include comparisons with negative weights. In my application, OLS "pre-trends" to become significant at further leads from treatment, as opposed to the alternative estimators.

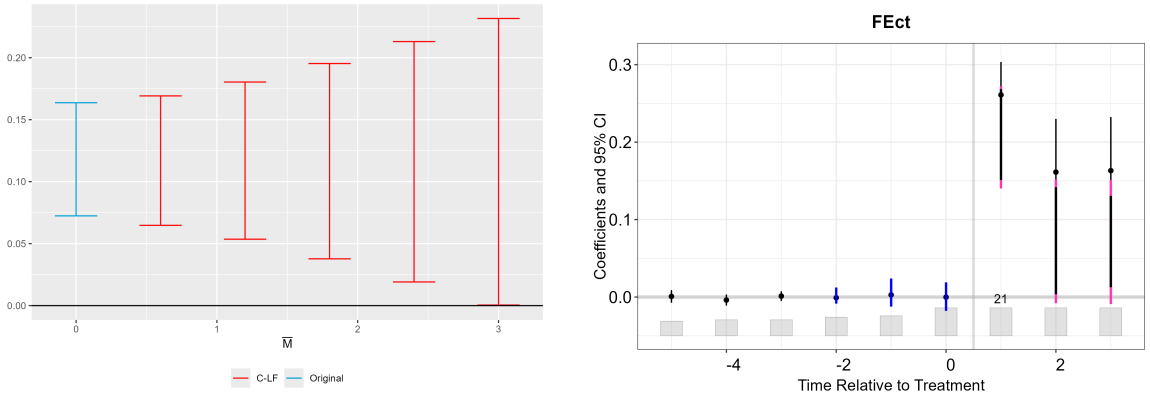
Figure 8: TWFE decomposition and alternative DID-type estimators



Note: Top figure: The Goodman-Bacon (2021) decomposition breakdown of the TWFE estimate into a weighted average of all possible 2×2 DID estimates across different cohorts. Bottom figure shows the ATT estimates on the average log markups from Year –5 to Year 3, along with their cluster robust 95% confidence intervals. Four estimators are employed: Borusyak et al. (2021), de Chaisemartin and d'Haultfœuille (2020), Callaway and Sant'Anna (2020), and Sun and Abraham (2020). Balanced panel: 26 firms and 16 time periods.

A concern highlighted by Roth (2024) is that some HTE-robust estimators produce pre-treatment and post-treatment coefficients asymmetrically, necessitating cautious interpretation. This asymmetry motivates obtaining robust confidence intervals in the following way. I use the `fect` (Liu et al., 2022) package, using the same method as independently proposed by Borusyak et al. (2021), to obtain the average treatment effect (ATT) estimate of 0.142, with a 200-round cluster bootstrap standard error of 0.018. Next, I remove observations in three placebo periods immediately preceding treatment onset within pre-treatment for model fitting, impute the counterfactual for these and the post-treatment periods, and calculate the event-study coefficients. This method ensures symmetric generation of coefficients for both post-treatment and placebo periods. Then I compute robust confidence sets (CSs) proposed by Rambachan and Roth (2023) that allow for PTA violations under relative magnitude (RM) restrictions, which allow for post-treatment violations of the PTA that are at most $\bar{M}$ times the size of the maximum violation in pretreatment *placebo* periods. Robust CSs are uniformly valid for the partially identified treatment effects under the RM restriction. Rambachan and Roth (2023, p. 2653) suggest $\bar{M} = 1$ as a "natural benchmark" with similar numbers of pretreatment and post-treatment periods, corresponding to the assumption that any PTA violations are no worse than the observed pretrend. Figure 9 marks the robust CS for the ATT in three periods after treatment estimated by `FEct` in pink. In this example, the robust CS excludes zero, and the breakdown value from the sensitivity analysis is $\tilde{M} = 3.0$, suggesting that the estimated public procurement effect is robust to potential violations of the PTA under the RM restriction. I also obtain separate confidence intervals for three post-treatment period similarly. In the figure below, I use $\bar{M} = 1$ to derive robust confidence intervals using coefficients from the placebo periods. I depict the placebo periods in blue and the robust confidence intervals in pink.

Figure 9: Absorbing Treatment Effect of Public Procurement on Markups: Sensitivity Analysis

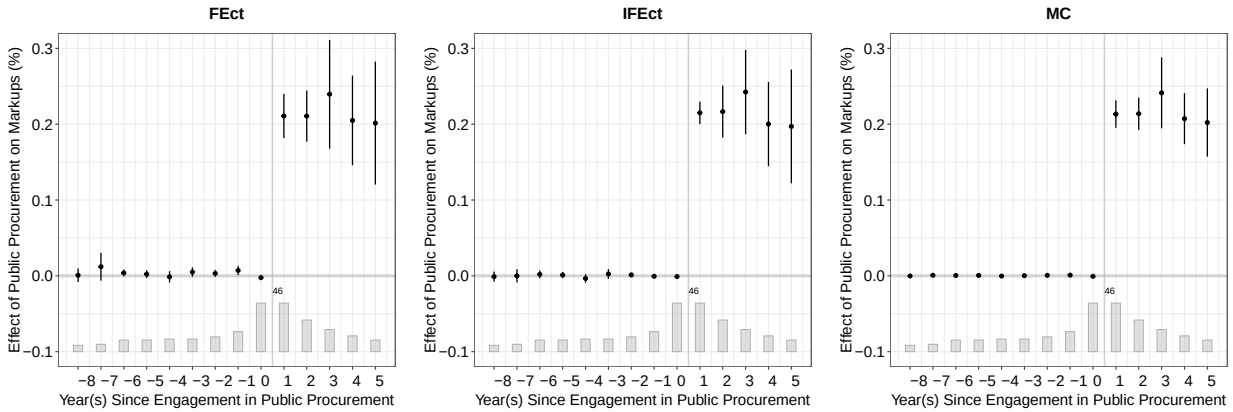


I now consider the general structure of the data, allowing the treatment to switch on and off, by implementing the imputation TWFE (`FEct`), interacted factor (`IFEct`) and matrix completion (`MC`) estimators. I use three periods before treatment for placebo tests, and obtain uncertainty estimates using clustered bootstrap at the unit level 1,000 times. Diagnostic tests are based on estimations obtained from the `fect` package. The equivalence test evaluates whether the 90% confidence intervals (corresponding to a 5% significance level) for the residuals in the pretreatment periods surpass a predetermined range. The null hypothesis posits that the residual exceeds this specified range for each pretreatment period. The criterion uses the default range, set at $0.36\sigma_{\varepsilon}$, where $0.36\sigma_{\varepsilon}$ represents the standard deviation of the outcome variable partialling out the two-way fixed effects. Next, by excluding observations in the last pretreatment periods during model fitting, I test whether the estimated ATT in these "placebo periods" significantly deviates from zero. Last, by excluding observations in the periods after exiting the treated status during model fitting, I test the null hypothesis that the average pseudo-treatment effect within this range is equal to zero.

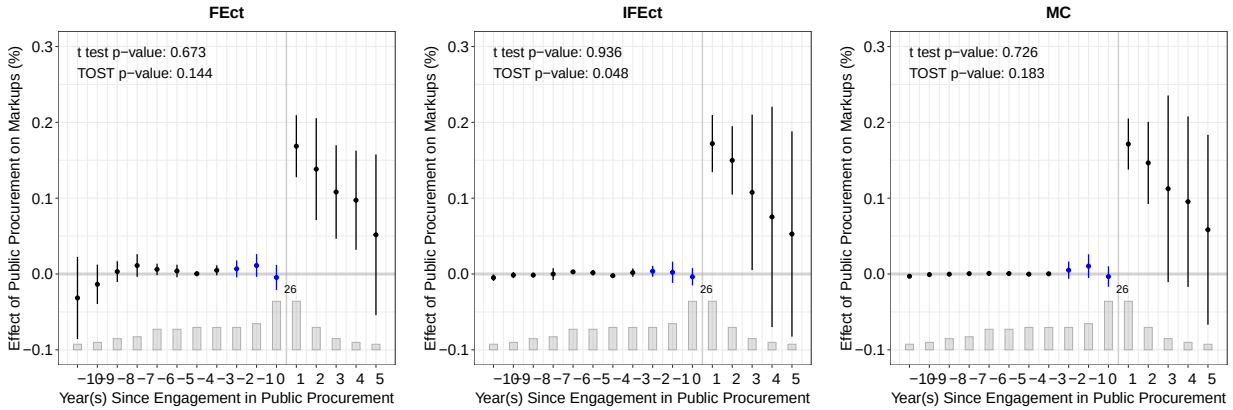
I plot the estimated dynamic treatment effects in Figure 10(a). I find that, first, the residual averages in the pretreatment periods are almost flat and around zero and the effect is stable ar around 0.2 in five periods after the treatment begins. Second, with the placebo test, we cannot reject the null of zero placebo effect for at the 5% level in all models, but we can reject the null whose magnitude is bigger than the default equivalence threshold ($p = 0.0048$) only for IFEct. Figure 10(b) shows the results from the placebo test. Finally, I report the results from the test for carryover effects in Figure 10(c), in which I test the carryover effects up to two years after engagement in public procurement ends. The test suggests that there are slightly negative carryover effects at least two years after engagement in public procurement ends.

Overall, the results shown that IFEct does best in all diagnostic tests and is likely to be the most suitable model among the three. The number of factors in the IFEct model is pre-specified as 2 and the estimated ATT is 0.206 with a standard error of 0.015. For matrix completion, the cross-validation optimal lambda is lower than 13 of the 16 singular values of the matrix L, which has rank equal to the number of periods (minimum of $N = 26$ and $T = 16$), i.e. the matrix completion approach uses a rank 13 factor model for the outcome partialled out by the two-way fixed effects. Violation of the no carryover effect assumption does not necessarily invalidate the research design, but suggests that a more flexible estimation strategy is required. While attractive due to respecting the true data, the general case analysis is likely to benefit from richer data to leverage the model approach. In the last section I conclude with a flexible method, albeit under absorbing treatment.

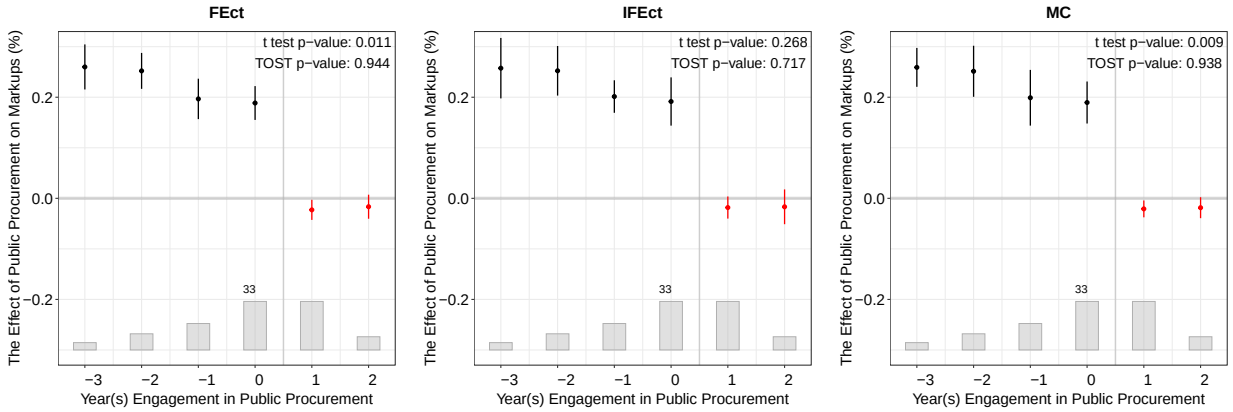
Figure 10: Non-absorbing Treatment Effect of Public Procurement on Markups: Counterfactual Estimators



(a) Dynamic Treatment Effects



(b) Placebo Test



(c) Test for Carryover Effects

Note:. The bar plot at the bottom of each panel illustrates the number of treated units at the given time period relative to the onset of the treatment. Three pretreatment periods serving as the placebo are rendered in blue in panel (b). Two periods after the treatment rendered in red in panel (c) are used to test for the presence of carryover effects. The p-values for the t test of the effects and for the TOST are shown at the top corners of panels (b) and (c).

Implementing Synthetic Difference in Differences I present the approach from Ciccia (2024) of the estimation procedure for event-study Synthetic Difference in Differences (SDID) estimators. In a setting with N units observed over T periods, $N_{tr} < N$ units receive treatment D starting from period a, where $1 < a \leq T$. The treatment D is binary, i.e. $D \in \{0, 1\}$, and it affects some outcome of interest Y. The outcome and the treatment are observed for all (i, t) cells, meaning that the data has a balanced panel structure. The cohort-specific SDID estimator from can be rearranged as follows:

$$\hat{\tau}_a^{sdid} = \frac{1}{T_{tr}^a} \sum_{t=a}^{T} \left(\frac{1}{N_{tr}^a} \sum_{i \in I^a} Y_{i,t} - \sum_{i=1}^{N_{co}} \omega_i Y_{i,t} \right) - \sum_{t=1}^{a-1} \left(\frac{1}{N_{tr}^a} \sum_{i \in I^a} \lambda_t Y_{i,t} - \sum_{i=1}^{N_{co}} \omega_i \lambda_t Y_{i,t} \right)$$

where λ_t and ω_i are the optimal weights chosen to best approximate the pre-treatment outcome evolution of treated and (synthetic) control units. The SDID weights ω_i are closely related to the weights used in Abadie et al. (2010), with two minor differences. First, Arkhangelsky et al. (2021) allow for an intercept term ω_0, meaning that the weights no longer need to make the unexposed pre-trends perfectly match the exposed ones; rather, it is sufficient that the weights make the trends parallel. The reason for this extra flexibility in the choice of weights is that the use of fixed effects α_i will absorb any constant differences between different units. Second, following Doudchenko and Imbens (2016), a regularization penalty is added to increase the dispersion, and ensure the uniqueness, of the weights. The main difference between the unit weights and the time weights is the of regularization for the former but not the latter. This choice reflects allowing for correlated observations within time periods for the same unit, but not across units within time period, beyond what is captured by the systematic component of outcomes as represented by a latent factor model.

As is well known (Ashenfelter and Card, 1985), DID relies on the assumption that log markups in private sector only firms would have evolved in a parallel way absent the engagement in public procurement. Figure 11 illustrates how SDID operates relative to DID and SC to produce $\hat{\tau}_{2018}^{sdid}$. Here, pre-intervention trends volatile, so the DID estimate should be considered suspect. In contrast, synthethic control re-weights the unexposed firms so that the weighted outcomes for these firms match government contractos pre-intervention as close as possible, and then attributes any post-intervention divergence of government contractors from this weighted average to the intervention. What SDID does here is to re-weight the unexposed control units to make their time trend parallel (but not necessarily identical) to government contractors pre-intervention, and then

applies a DID analysis to this re-weighted panel. Moreover, because of the time weights, it only focuses on a subset of the pre-intervention time periods when carrying out this last step, specifically $\lambda_{2017} = 0.838$ and $\lambda_{2015} = 0.162$. These time periods are selected so that the weighted average of historical outcomes predict average treatment period outcomes for control units, up to a constant. Appendix Figure G presents the SDID trends and time weights for all cohorts.

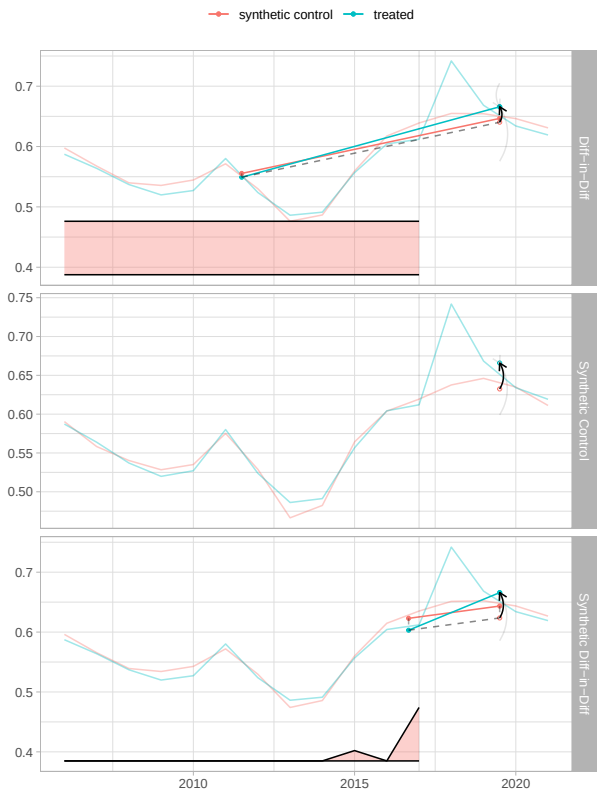


Figure 11: Comparison between DID, synthetic control, and SDID estimates

It is useful to contrast the data-driven SDID approach to selecting the time weights to both DID, where all pre-treatment periods are given equal weight, and to event studies where typically the last pre-treatment period is used as a comparison and so implicitly gets all the weight (*e.g.,* Borusyak and Jaravel (2016); Freyaldenhoven et al. (2019)).The use of weights in the SDID estimator effectively makes the two-way fixed effect regression "local," in that it emphasizes (puts more weight on) units that on average are similar in terms of their past to the target (treated) units, and it emphasizes periods that are on average similar to the target (treated) periods. This localization can bring two benefits relative to the standard DID estimator. Intuitively, using only similar units and similar periods makes the estimator more robust. Perhaps less intuitively, the use of the

weights can also improve the estimator's precision by implicitly removing systematic (predictable) parts of the outcome. Together, these weights make the DID strategy more plausible. This idea is not far from the current empirical practice. Raw data rarely exhibits parallel time trends for treated and control units, and researchers use different techniques, such as adjusting for covariates or selecting appropriate time periods to address this problem (*e.g.*, Abadie (2005); Callaway and Sant'Anna (2020)). Graphical evidence that is used to support the parallel trends assumption is then based on the adjusted data. SDID makes this process automatic and applies a similar logic to weighting both units and time periods, all while retaining statistical guarantees. From this point of view, SDID addresses pretesting concerns expressed in Roth (2022).

It is possible to estimate the treatment effect ℓ periods after the adoption of the treatment, with $\ell \in \{1, ..., T^a_{post}\}$, via a simple disaggregation of τ^{sdid}_a into the following event-study estimators:

$$\hat{\tau}^{sdid}_{a,\ell} = \frac{1}{N^a_{tr}} \sum_{i \in I^a} Y_{i,a-1+\ell} - \sum_{i=1}^{N_{co}} \omega_i Y_{i,a-1+\ell} - \sum_{t=1}^{a-1} \left(\frac{1}{N^a_{tr}} \sum_{i \in I^a} \lambda_t Y_{i,t} - \sum_{i=1}^{N_{co}} \omega_i \lambda_t Y_{i,t} \right)$$

This estimator is very similar to those proposed by Borusyak et al. (2021), Liu et al. (2022) and Gardner (2022), when the design is a canonical DiD (de Chaisemartin and D'Haultfoeuille, 2023). The only difference lies in the fact that the outcomes are weighted via unit-time specific weights.

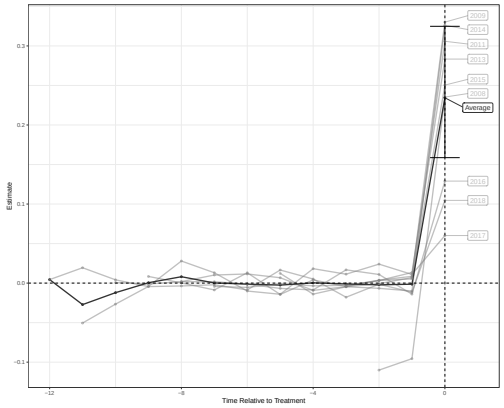
Table 5 reports the cohort-specific ATT the estimates of the contamporaneous effect of treatment $\hat{\tau}^{sdid}_{a,1}$. Figure 12 vizualizes the pre-treatment fit and the same "on-impact" effect of public procurement using the Augmented SC approach (Ben-Michael et al., 2022) and shows robustness across the results obtained from SDID, as the relative ordering of cohorts across methods.

Table 5: Synthetic Difference in Differences: On Impact ATTs - Cohort level

Cohort	2007	2008	2009	2011	2013	2014	2015	2016	2017	2018
$\hat{\tau}^{sdid}_{a,1}$	.33	.34	.35	.30	.28	.34	.25	.11	.05	.11

The most notable feature of the $\hat{\tau}^{sdid}_a$ estimates in is the heterogeneity in treatment effects across groups, notably the declining trend in the public procurement premium over time. Specifically, firms that engaged in public procurement in 2007 are estimated to have, on average, markups 33% greater than they would have if operating solely in the private sector. In contrast, firms have first records of public procurement contract in their financial statements in 2018 show an effect that is three times smaller, 11%, relative to their private sector only counterfactuals.

Figure 12: Augmented Synthetic Control: On-impact cohort ATTs

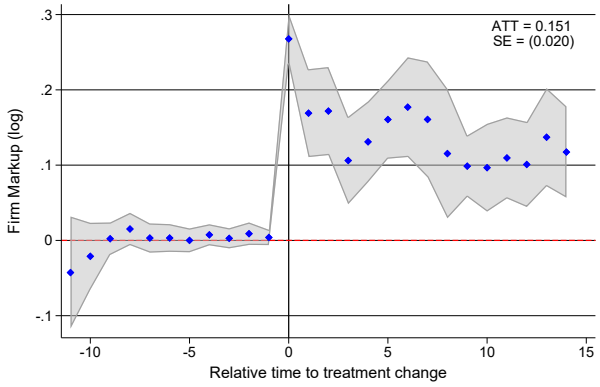


Let A_ℓ be the subset of cohorts in A such that $a - 1 + \ell \leq T$, i.e. such that their ℓ-th dynamic effect can be computed, and let $N_{tr}^{\ell} = \sum_{a \in A_\ell} N_{tr}^{a}$ denote the number of units in cohorts where the ℓ-th dynamic effect can be estimated. Let

$$\hat{\tau}_{\ell}^{sdid} = \sum_{a \in A_\ell} \frac{N_{tr}^{a}}{N_{tr}^{\ell}} \hat{\tau}_{a,\ell}^{sdid}$$

denote the weighted sum of the cohort-specific treatment effects ℓ periods after the onset of the treatment, with weights corresponding to the relative number of groups participating into each cohort. This estimator aggregates the cohort-specific treatment effects, upweighting more representative cohorts in terms of units included. Finally, let $T_{post} = \sum_{a \in A} N_{tr}^{a} T_{tr}^{a}$ and $T_{tr} = \max_{a \in A} T_{tr}^{a}$ be the maximum number of post-treatment periods across all cohorts. Figure 13 reports the the overall $\widehat{ATT} = \frac{1}{T_{post}} \sum_{\ell=1}^{T_{tr}} N_{tr}^{\ell} \hat{\tau}_{\ell}^{sdid}$, equal to 0.15 with a standard error of 0.02, together with the individual event-study estimates $\hat{\tau}_{\ell}^{sdid}$ for $\ell \in \{1, ..., T_{tr}\}$.

Figure 13: Synthetic Difference in Differences: Event Study



5.6 Conclusion

The recent literature has greatly expanded the set of methods available to empirical researchers in social sciences in settings that are important in practice. This section is an attempt to put these methods in context and show the close relationship between various approaches, including two-way-fixed-effect and synthetic control methods, to follow Arkhangelsky and Imbens (2024) guidance on the use of various methods.

Although the standard TWFE estimator continues to be widely used, there are now methods available that have superior properties in settings with both cross-section and time dimensions at least modestly large. (In cases with few units and few time periods, there may not be enough information in the data to go beyond the simpler methods.) These methods relax the parallel trends assumption that is unattractive both from a conceptual perspective (because it is tied to a particular functional form) and from a practical perspective (because it is unlikely to hold over long periods of time). Some of the new methods allow for factor structures that generalize the two-way fixed effect setup. Others use synthetic control approaches, sometimes in combination with fixed effects. While none of these methods is likely to dominate uniformly, preliminary simulation evidence (*e.g.,* Arkhangelsky et al. (2021)) suggests that many of them dominate TWFE in realistic settings. Recent results in Arkhangelsky and Hirshberg (2023) also suggest that some of these methods, in particular those based on synthetic control, dominate TWFE in settings with more complicated selection mechanisms.

The staggered adoption case, common in empirical work, opens up new opportunities for estimation strategies (exploiting the variation in adoption times), but also forecloses some options (the standard synthetic control estimator). Some of the recent proposals modify the TWFE estimator and relax the parallel trends assumptions by limiting the comparisons between treated and control outcomes to a subset of the set of possible comparisons. This subset may avoid comparisons distant in time, avoid the use of units that are to be treated at a future date as controls, or, in contrast, avoid the use of units that are never treated.

Much of the discussion on unconfoundedness and the TWFE model has been framed in terms of a choice. Here I took inspiration from Angrist and Pischke (2008), page 184: ”So what is an applied guy to do? One answer, as always, is to check the robustness of your findings using alternative identifying assumptions, ... Fixed effects and lagged dependent variables estimates also have a useful bracketing property ,.., bounding the causal effect you are after.”

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Abadie, A. and Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(-):113–132.

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification Properties of Recent Production Function Estimators. *Econometrica*, 83(6):2411–2451.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.

Arkhangelsky, D. and Hirshberg, D. (2023). *Large-sample properties of the synthetic control method under selection on unobservables*. em arXiv preprint arXiv:2311.13575.

Arkhangelsky, D. and Imbens, G. (2024). Causal models for longitudinal and panel data: a survey. *The Econometrics Journal*, 27(3):C1–C61.

Ashenfelter, O. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4):648–660.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., and Van Reenen, J. (2020). The Fall of the Labor Share and the Rise of Superstar Firms*. *The Quarterly Journal of Economics*, 135(2):645–709.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bandiera, O., Prat, A., and Valletti, T. (2009). Active and Passive Waste in Government Spending: Evidence from a Policy Experiment. *American Economic Review*, 99(4):1278–1308.

Baránek, B. and Titl, V. (2024). The Cost of Favoritism in Public Procurement. *The Journal of Law and Economics*, 67(2):445–477.

Ben-Michael, E., Feller, A., and Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803.

Ben-Michael, E., Feller, A., and Rothstein, J. (2022). Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):351–381.

Berry, S., Gaynor, M., and Scott Morton, F. (2019). Do Increasing Markups Matter? Lessons from Empirical Industrial Organization. *Journal of Economic Perspectives*, 33(3):44–68.

Bond, S., Hashemi, A., Kaplan, G., and Zoch, P. (2021). Some Unpleasant Markup Arithmetic: Production Function Elasticities and Their Estimation from Production Data. *Journal of Monetary Economics*, 121:1–14.

Borusyak, K. and Jaravel, X. (2016). Revisiting event study designs.

Borusyak, K., Jaravel, X., and Spiess, J. (2021). *Revisiting event study designs: Robust and efficient estimation*. em arXiv preprint arXiv:2108.12419.

Callaway, B. and Sant'Anna, P. H. (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*.

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1).

Ciccia, D. (2024). A short note on event-study synthetic difference-in-differences estimators.

Cinelli, C., Forney, A., and Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*, page 00491241221099552.

Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, pages 187–199.

de Chaisemartin, C. and d'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96.

de Chaisemartin, C. and D'Haultfoeuille, X. (2023). Difference-in-differences for simple and complex natural experiments. *Working textbook under contract with Princeton University Press*.

De Loecker, J., Eeckhout, J., and Unger, G. (2020). The Rise of Market Power and the Macroeconomic Implications. *The Quarterly Journal of Economics*, 135(2):561–644.

De Loecker, J. and Syverson, C. (2021). An Industrial Organization Perspective on Productivity. In *Handbook of Industrial Organization, Volume 4*, pages 141–223. Elsevier.

De Loecker, J. and Warzynski, F. (2012). Markups and Firm-Level Export Status. *The American Economic Review*, 102(6):2437–2471.

De Ridder, M., Grassi, I., and Morzenti, A. (2024). The hitchhiker's guide to Markup Estimation: Assessing Estimates from Financial Data. Discussion Paper 17532, CEPR Press.

Decarolis, F., Fisman, R., Pinotti, P., and Vannutelli, S. (2020). Rules, Discretion, and Corruption in Procurement: Evidence from Italian Government Contracting. Working Paper 28209, National Bureau of Economic Research.

Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

Engle, R. F., Hendry, D. F., and Richard, J.-F. (1983). Exogeneity. *Econometrica*, pages 277–304.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276.

Freyaldenhoven, S., Hansen, C., and Shapiro, J. M. (2019). Pre-event trends in the panel event-study design. *American Economic Review*, 109(9):3307–38.

Gardner, J. (2022). Two-stage differences in differences. *arXiv preprint*.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Hall, R. E. (2018). New Evidence on the Markup of Prices over Marginal Costs and the Role of Mega-Firms in the US Economy. NBER Working Paper 24574.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294.

Imbens, G. and Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Imbens, G. and Xu, Y. (2024). LaLonde (1986) after Nearly Four Decades: Lessons Learned.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review, Papers and Proceedings*, 93(2):126–132.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, pages 1–29.

Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2):373–419.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kang, K. and Miller, R. A. (2022). Winning by Default: Why is There So Little Competition in Government Procurement? *The Review of Economic Studies*, 89(3):1495–1556.

Klette, T. J. and Griliches, Z. (1996). The inconsistency of common scale estimators when output prices are unobserved and endogenous. *Journal of Applied Econometrics*, 11(4):343–361.

Kline, P. (2011). Oaxaca-blinder as a reweighting estimator. *American Economic Review*, 101(3):532–537.

Levinsohn, J. and Petrin, A. (2003). Estimating Production Functions Using Inputs to Control for Unobservables. *The Review of Economic Studies*, 70(2):317–341.

Liu, L., Wang, Y., and Xu, Y. (2022). A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 68(1):160–176.

Miller, N. H. (2024). Industrial Organization and The Rise of Market Power. Working Paper 32627, National Bureau of Economic Research.

Mou, H., Liu, L., and Xu, Y. (2023). Panel data visualization in r (panelview) and stata (panelview). *Journal of Statistical Software*.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49(6):1417–1426.

OECD (2021). Public Procurement. https://www.oecd.org/en/topics/policy-issues/public-procurement.html. Accessed: October 3, 2024.

Olley, G. S. and Pakes, A. (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*, 64(6):1263–1297.

Palguta, J. and Pertold, F. (2017). Manipulation of Procurement Contracts: Evidence from the Introduction of Discretionary Thresholds. *American Economic Journal: Economic Policy*, 9(2):293–315.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *The Review of Economic Studies[*, 90(5):2555–2591.

Raval, D. (2023). Testing the Production Approach to Markup Estimation. *The Review of Economic Studies*, 90(5):2592–2611.

Richard Blundell and Stephen Bond (1998). *Initial conditions and moment restrictions in dynamic panel data models*. em Journal of econometrics 87(1):.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, pages 550–560.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–22.

Roth, J. (2024). Interpreting Event-Studies from recent Difference-in-Differences methods.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Shapiro, C. and Yurukoglu, A. (2024). Trends in Competition in the United States: What Does the Evidence Show? Working Paper 32762, National Bureau of Economic Research.

Sun, L. and Abraham, S. (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

Syverson, C. (2024). Markups and Markdowns. Working Paper 32871, National Bureau of Economic Research.

Szucs, F. (2024). Discretion and Favoritism in Public Procurement. *Journal of the European Economic Association*, 22(1):117–160.

Titl, V. (2023). The One and Only: Single-Bidding in Public Procurement. Working Papers, Utrecht School of Economics.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

Zubizarreta, J. R., Stuart, E. A., Small, D. S., and Rosenbaum, P. R. (2023). *Handbook of Matching and Weighting Adjustments for Causal Inference*. CRC Press.

Appendix

A The De Loecker and Warzynski Framework

Consider an economy with N firms, indexed by i. Firms are heterogeneous in terms of their productivity Ω_{it} and production technology $Q_{it}(\cdot)$. In each period t, firm i minimizes the contemporaneous cost of production given the following production function:

$$Q_{it} = Q_{it}(\Omega_{it}, V_{it}, K_{it}),$$

where V is the vector of variable inputs of production (e.g., labor, intermediate inputs, materials, etc.), K_{it} is the capital stock, and Ω_{it} represents productivity. The key assumption is that within each period, variable inputs adjust frictionlessly, while capital is subject to adjustment costs and other frictions. For simplicity, we treat the vector V as a scalar V for exposition purposes. We consider the Lagrangian objective function associated with the firm's cost minimization problem:

$$L(V_{it}, K_{it}, \lambda_{it}) = P_{V_{it}} V_{it} + r_{it} K_{it} + F_{it} - \lambda_{it} \left(Q(\cdot) - Q_{it} \right),$$

where P_V is the price of the variable input, r is the user cost of capital, F_{it} is the fixed cost, and λ_{it} is the Lagrange multiplier. The first-order condition with respect to the variable input V is given by:

$$\frac{\partial L_{it}}{\partial V_{it}} = P_V - \lambda_{it} \frac{\partial Q(\cdot)}{\partial V_{it}} = 0.$$

Multiplying all terms by $\frac{V_{it}}{Q_{it}}$ and rearranging, we obtain the expression for the output elasticity of input V:

$$\theta_{it}^V \equiv \frac{\partial Q(\cdot)}{\partial V_{it}} \cdot \frac{V_{it}}{Q_{it}} = \frac{1}{\lambda_{it}} \cdot \frac{P_V V_{it}}{Q_{it}}.$$

The Lagrange multiplier λ_{it} is a direct measure of marginal cost, and we define the markup as the ratio of price to marginal cost: $\mu_{it} = \frac{P}{\lambda_{it}}$. Substituting marginal cost into the markup expression, we derive the formula for the markup:

$$\mu_{it} = \theta_{it}^V \cdot \frac{P_{it} Q_{it}}{P_V V_{it}}.$$

Consider the following general production function:

$$Q_{it} = F(V_{it}^1, V_{it}^2, \ldots, V_{it}^V, k_{it}; \beta) \exp(\omega_{it}),$$

where Q_{it} is the output of firm i at time t, V_{it}^v are the variable inputs, k_{it} is the capital stock, β represents the common technology parameters, and ω_{it} is the firm-specific productivity shock. Next, we consider the log-linear version of the production function, incorporating unobserved productivity shocks ω_{it} and measurement error ϵ_{it}:

$$y_{it} = f(v_{it}, k_{it}; \beta) + \omega_{it} + \epsilon_{it},$$

where y_{it} is the logged output, $f(x_{it}, k_{it}; \beta)$ is the production function in log form, v_{it} represents the vector of variable inputs, k_{it} is capital, and ϵ_{it} includes unanticipated production shocks and measurement error. We explicitly allow for measurement error in output and assume that ϵ_{it} is i.i.d. and uncorrelated with input choices. To obtain consistent estimates of the parameters of the production function β, and thus compute θ_{it}^V, we use the demand for variable inputs as a proxy for productivity:

$$v_{it} = v_t(k_{it}, \omega_{it}, z_{it}),$$

where z_{it} represents additional state variables that influence input demand, such as input prices or a firm's public procurement status. By inverting the input demand function $v_t(\cdot)$, we can recover an estimate of unobserved productivity ω_{it}. We start by specifying a second-order approximation translog production function, which is given by:

$$y_{it} = \beta_v v_{it} + \beta_k k_{it} + \beta_{vv} v_{it}^2 + \beta_{kk} k_{it}^2 + \beta_{vk} v_{it} k_{it} + \omega_{it} + \epsilon_{it},$$

where y_{it} represents the log of output for firm i at time t, v_{it} and k_{it} are the logged values of variable and capital inputs, and ω_{it} denotes unobserved productivity. The term ϵ_{it} captures measurement error and unanticipated production shocks. In the first stage, we estimate expected output using the following equation:

$$y_{it} = \varphi_t(v_{it}, k_{it}, z_{it}) + \epsilon_{it},$$

where $\varphi_t(\cdot)$ captures the systematic component of output, which depends on variable inputs, capital, and additional controls (z_{it}) such as demand or market conditions that affect input demand.

These control variables help ensure that the productivity estimates are robust even in the presence of imperfect competition or heterogeneous demand across firms. The first-stage estimate of expected output, $\hat{\varphi}_{it}$, is obtained as follows:

$$\hat{\varphi}_{it} = \beta_v v_{it} + \beta_k k_{it} + \beta_{ll} v_{it}^2 + \beta_{kk} k_{it}^2 + \beta_{vk} v_{it} k_{it} + h_t(v_{it}, k_{it}, z_{it}),$$

where $h_t(\cdot)$ accounts for variations in productivity driven by variable inputs, capital, and other firm-specific factors. The second stage relies on the law of motion for productivity, modeled as a first-order Markov process:

$$\omega_{it} = g_t(\omega_{it-1}, p_{it-1}) + \xi_{it},$$

where ξ_{it} represents the productivity innovation. To account for firm-level decisions that affect future productivity, such as public procurement status, we allow for additional lagged and observable decision variables p_{it-1} in the estimation of the productivity process. This adjustment addresses the potential issues raised regarding the limitations of assuming exogenous productivity processes. Once we have estimated $\hat{\varphi}_{it}$, we compute the firm's productivity ω_{it} for any given parameter set $\beta = (\beta_v, \beta_k, \beta_{vv}, \beta_{kk}, \beta_{vk})$ using the following expression:

$$\omega_{it}(\beta) = \hat{\varphi}_{it} - \beta_v v_{it} - \beta_k k_{it} - \beta_{vv} v_{it}^2 - \beta_{kk} k_{it}^2 - \beta_{vk} v_{it} k_{it}.$$

We then estimate the productivity innovation $\xi_{it}(\beta)$ by non-parametrically regressing $\omega_{it}(\beta)$ on its lag $\omega_{it-1}(\beta)$ and the public procurement indicator, and recovering the residuals. This process enables us to estimate all of the production function coefficients and to account for the role of productivity in the input-output relationship. To obtain the production function parameters, we form moment conditions based on the innovations to productivity. These moments rely on a timing assumption that capital decisions are made one period ahead, meaning that capital should not be correlated with the innovation to productivity. We use the following moment condition:

$$E\left[\xi_{it}(\beta)\begin{pmatrix} v_{it-1} \\ k_{it} \\ v_{it-1}^2 \\ k_{it}^2 \\ v_{it-1}k_{it} \end{pmatrix}\right] = 0,$$

where $\xi_{it}(\beta)$ represents the productivity innovation, and v_{it-1} and k_{it} are the lagged variable and capital inputs, respectively. These moment conditions allow us to identify the parameters of the production function using standard Generalized Method of Moments. The identification strategy exploits the fact that capital is predetermined, meaning that it is decided in advance and should not be correlated with the contemporaneous productivity innovation ξ_{it}. By contrast, variable inputs are assumed to respond flexibly to productivity shocks within the period, and thus the expectation of $v_{it}\xi_{it}$ is expected to be nonzero. In order for lagged variable inputs to serve as a valid instrument for current variable inputs, it is necessary to assume that input prices are serially correlated over time. Finally, we use block bootstrapping to estimate standard errors, ensuring that the error structure reflects the autocorrelation across firms and time.

Under a translog production function, the output elasticity for variable input (V) is given by:

$$\hat{\theta}_{it}^{V} = \hat{\beta}_v + 2\hat{\beta}_{vv} v_{it} + \hat{\beta}_{vk} k_{it},$$

With the estimated output elasticities in hand, we use the first-order condition on input demand and our to compute markups as follows:

$$\mu_{it} = \frac{\hat{\theta}_{it}^{V}}{\hat{\alpha}_{it}^{V}},$$

where $\hat{\theta}_{it}^{V}$ is the output elasticity of input V, and $\hat{\alpha}_{it}^{V}$ is the expenditure share of input V in firm i's total revenue. However, we do not directly observe the true expenditure share of input V_{it}, as we only observe firm-level output $\tilde{Q}_{it}$, which is measured with error. The observed output is given by:

$$\tilde{Q}_{it} = Q_{it} \exp(\epsilon_{it}),$$

where Q_{it} is the true output and ϵ_{it} represents measurement error or unanticipated shocks to output. The first stage of our procedure provides an estimate of ϵ_{it}, which we use to correct the observed expenditure share:

$$\hat{\alpha}_{it}^{V} = \frac{P_{it}^{V} V_{it}}{P_{it} \tilde{Q}_{it}} \exp(\hat{\epsilon}_{it}),$$

This correction eliminates any variation in expenditure shares not related to variables that affect input demand, including input prices, productivity, technology parameters, and market characteristics, such as the elasticity of demand and income levels.

B Unweighted Markup Distribution

Table 6: Summary Statistics by Year

Year	p10	p25	p50	p75	p90	Mean	SD	N
2006	1.09	1.19	1.35	1.51	1.60	1.36	0.20	227
2007	1.04	1.10	1.35	1.50	1.58	1.32	0.20	290
2008	1.06	1.11	1.27	1.42	1.49	1.27	0.17	348
2009	1.05	1.10	1.28	1.42	1.49	1.27	0.18	412
2010	1.03	1.09	1.25	1.37	1.45	1.25	0.18	497
2011	1.04	1.10	1.28	1.40	1.49	1.27	0.19	506
2012	1.02	1.11	1.24	1.36	1.44	1.24	0.17	457
2013	1.04	1.12	1.20	1.30	1.35	1.21	0.13	235
2014	1.03	1.13	1.23	1.35	1.42	1.23	0.14	243
2015	1.09	1.20	1.29	1.42	1.47	1.31	0.19	245
2016	1.08	1.19	1.31	1.44	1.50	1.31	0.17	338
2017	1.13	1.23	1.30	1.42	1.48	1.32	0.16	660
2018	1.13	1.23	1.32	1.46	1.53	1.34	0.17	708
2019	1.15	1.26	1.33	1.45	1.51	1.35	0.16	764
2020	1.12	1.23	1.31	1.43	1.50	1.33	0.16	769
2021	1.11	1.19	1.28	1.37	1.42	1.28	0.14	562

Table 7: Summary Statistics by NACE 2-digit division

NACE 2	p10	p25	p50	p75	p90	Mean	SD	N
41: Construction of Buildings	1.24	1.28	1.39	1.46	1.53	1.38	0.14	3950
42: Civil Engineering	1.14	1.21	1.29	1.36	1.46	1.30	0.14	681
43: Specialised Activities	1.03	1.07	1.14	1.25	1.32	1.17	0.14	2630

Table 8: Summary Statistics by Public Procurement

Public Procurement	p10	p25	p50	p75	p90	Mean	SD	N
0: Private-sector-only firms	1.04	1.10	1.21	1.29	1.35	1.21	0.15	4034
1: Government-contractors	1.25	1.31	1.42	1.48	1.54	1.41	0.13	3227

C Cross Sectional Results by Year and NACE 2-digit divison

Table 9: **Percentage Public Procurement Markup Premia by Year**

Year	N	Adjusted R^2	δ_1
2006	227	0.876	0.181 (0.008)
2007	290	0.917	0.161 (0.007)
2008	348	0.865	0.166 (0.007)
2009	412	0.821	0.160 (0.008)
2010	497	0.756	0.169 (0.008)
2011	506	0.754	0.188 (0.009)
2012	457	0.761	0.168 (0.007)
2013	235	0.730	0.151 (0.010)
2014	243	0.835	0.146 (0.007)
2015	245	0.657	0.156 (0.013)
2016	338	0.764	0.136 (0.007)
2017	660	0.673	0.149 (0.006)
2018	708	0.718	0.138 (0.005)
2019	764	0.643	0.135 (0.005)
2020	769	0.712	0.134 (0.005)
2021	562	0.614	0.136 (0.006)

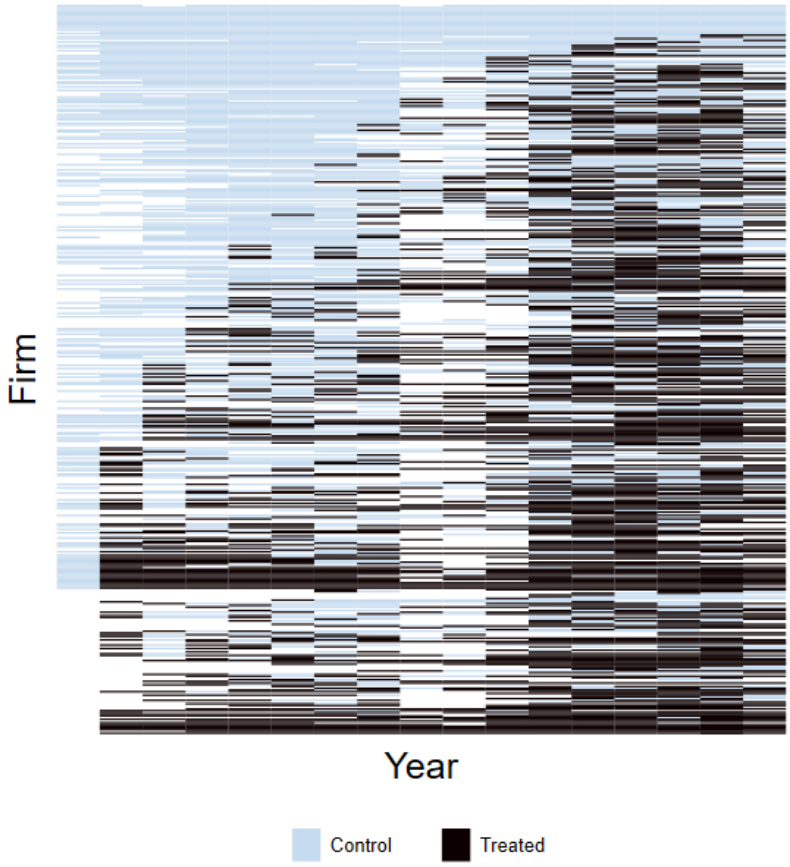
Note: Estimates are obtained after running equation 2 by year. Cluster robust standard errors in parentheses.

Table 10: **Percentage Public Procurement Markup Premia by NACE 2-digit**

	41 Construction of Buildings	42 Civil Engineering	43 Specialised Activities
δ_1	0.144	0.154	0.158
	(0.004)	(0.008)	(0.006)
N	3950	681	2630
Adjusted R^2	0.620	0.723	0.601

Note: Estimates are obtained after running equation 2 by sub-industry. Cluster robust standard errors in parentheses

D Data Structure Visualization



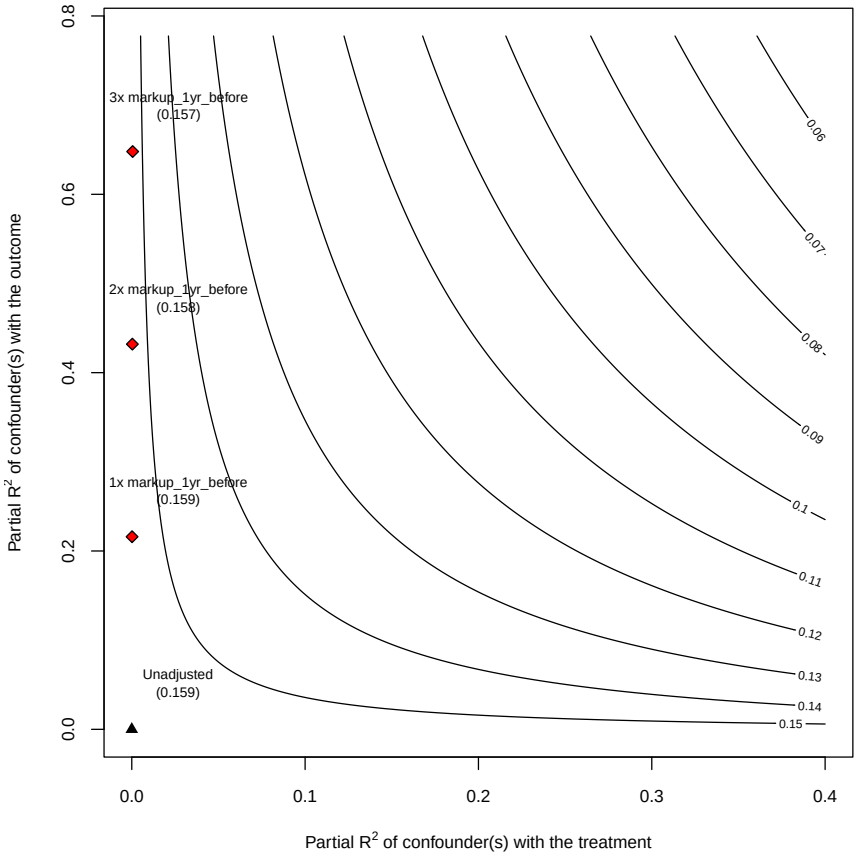
Note: This figure shows all 342 unique public procurement histories for the 7,261 firms with estimated markups in the sample. Black: Government Contractors, Blue: Firms inactive in Public Procurement, White: Missing.

E Time Series Results: Parameter Estimates

γ_0 (Constant)	γ_1 (Entry)	γ_2 (Exit)	γ_3 (Always)
0.831	0.120	-0.038	0.153
(0.054)	(0.006)	(0.009)	(0.005)

Estimates are obtained after running equation 3.
Cluster robust standard errors in parentheses.
N = 5744. Adjusted R^2 = 0.736.

F Unconfoundedness Results: Sensitivity Analysis



Note: Contour plot for the treatment effect coefficient $\hat{\tau}_{OLS}$ based on sensitivity analysis first proposed by Imbens (2003) and then modified by Cinelli and Hazlett (2020). The benchmark covariate is log markup one year before the contract (Y_{-1}). The model is a linear regression with all available long-term covariates included.

G Synthetic DID: Cohort Trends and Time Weights