Markups and Public Procurement: Evidence from Czech Construction Tenders

Marek Chadim*

Abstract

This paper analyzes the effect of public procurement on firm markups, using a panel dataset of Czech construction firms from 2006 to 2021. Markups are estimated through a structural framework, while biases are addressed using a selection-on-observables design and models for unobserved factors. Propensity score-based estimations indicate that firm markups increase by approximately 15% during contract years. A temporal analysis, employing synthetic control and matrix completion methods, reveals that treatment effects decline from around 30% in 2006 to 10% in 2021. These patterns are consistent with institutional improvements in the Czech Republic and offer empirical evidence of increasing efficiency in public spending.

Keywords Models with Panel Data; Firm Behavior: Empirical Analysis; Procurement; Construction

JEL C23, D22, H57, L74

*MSc thesis submitted to the Department of Economics, Stockholm School of Economics. 42624@student.hhs.se. Replication files are available at github.com/marek-chadim/Markups-and-Public-Procurement. I am grateful to my thesis supervisor, Jaakko Meriläinen, for invaluable guidance; to Matěj Bajgar, Mitch Downey, and David Schönholzer for their helpful feedback; and to Jiří Skuhrovec of Datlab s.r.o. for providing administrative data on government procurement. This thesis builds on my earlier work at the Institute of Economic Studies, Charles University Prague (DOI: dspace.cuni.cz/handle/20.500.11956/184831), where I accessed firm financial statement data via MagnusWeb.

1 Introduction

Public procurement accounts for approximately 12% of GDP in OECD countries and represents a significant portion of total government expenditure (OECD, 2021). Ensuring efficiency in public procurement is essential to preventing resource waste and maximizing taxpayer value. The role of discretion, political favoritism, and limited competition are some of the main topics analyzed in the literature. Much of the existing research on public procurement focuses on comparing tenders, while the differences between public procurement and private markets have received little attention.

Models incorporating heterogeneous producers and firm-specific markups suggest that more productive firms can afford the costs of entering public procurement, which may explain why government contractors tend to have higher markups. In addition, differences in quality are significant. Government contractors may charge higher markups if they supply higher-quality goods produced using higher-quality inputs. However, the dynamics of markups as firms enter and exit public procurement markets remain unexplored.

This paper presents evidence on the dynamics of markups as firms enter and exit public procurement markets. To the best of my knowledge, this is the first study to present robust econometric evidence on the relationship between markups and engagement in public procurement. My findings contribute to debates on public procurement efficiency, competition, and the broader implications of government contracts for market power, particularly when these contracts account for a substantial share of economic activity.

First, I examine the evolution of market power in the Czech construction sector using firm-level data from 2006 to 2021. I estimate markups under the assumption that cost-minimizing producers choose optimal variable input expenditure using the method of De Loecker and Warzynski (2012). This approach provides plant-level markup estimates without requiring assumptions about how firms compete in the product market. The results of the structural estimation show that aggregate markups in the Czech construction sector declined from 40% above marginal cost in 2006 to 30% in 2021, driven primarily by firms at the upper end of the markup distribution.

Second, I apply recent methodological advances from Imbens and Xu (2024) to estimate credible non-experimental estimates of the public procurement premium under unconfoundedness, conditioning on observable lagged markups and financial statement data. To ensure sufficient covariate overlap, I trim the sample and apply propensity score methods to generate doubly robust estimators, which I validate using placebo tests and sensitivity analyses.

Finally, I employ recent panel data methods to estimate causal effects of binary interventions in longitudinal settings, drawing on practical guidance from Arkhangelsky and Imbens (2024). These methods extend difference-in-differences and two-way fixed-effects models by incorporating factor models, interactive fixed-effects models, and synthetic control methods. I emphasize heterogeneity in causal effects, employ robust inference techniques tailored for settings with a limited number of treated units, and evaluate both absorbing and non-absorbing treatment structures.

Drawing causal inferences from observational data inherently relies on strong assumptions. The absence of natural experiments, instrumental variables, and regression discontinuity designs limits the availability of exogenous variation in this study. I rely on selection based on either observables or unobservables, using the potential outcomes framework to conceptualize the problem as modeling the counterfactual markups for government contractors had they not engaged in public procurement. The stable unit treatment value assumption (SUTVA) rules out interference between units, as well as the dynamic effects of past treatments and outcomes. In the competitive environment of public procurement, interpreting causal effects becomes problematic under SUTVA violations. Nevertheless, I prioritize internal validity through design-based balanced panel results and also present findings that leverage the full dataset to enhance generalizability.

1.1 Literature Review

Public Procurement One major challenge in public procurement is balancing the benefits of discretion with the risks of rent-seeking. If the interests of procuring officials and the public are not perfectly aligned, bureaucratic discretion increases the risk of corruption and the misallocation of contracts to politically connected winners. On the other hand, auctions are typically more time-consuming and expensive to organize compared to less formalized procedures, such as direct negotiations. In the Czech Republic, Titl (2023) shows that about 23% of public contracts are awarded through single-bid procedures, and that a 2012 reform aimed at reducing single-bidding resulted in a 10% drop in procurement. Kang and Miller (2022) find that single-bidding is prevalent in the U.S. as well, but show that information asymmetries and noncontractible quality dimensions may make a certain level of discretion necessary to incentivize effective contract execution, resulting in ambiguity regarding the net effect of discretion on procurement outcomes.

Palguta and Pertold (2017) document that officials in the Czech Republic often manipulate procurement thresholds by adjusting contract values to avoid competitive bidding. However, due to identification challenges caused by bunching, the existing empirical evidence is also inconclusive.

Szucs (2024) uses a variation of a policy reform and finds that granting 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, particularly those with fewer bidders, increase corruption risks and often benefit politically connected firms.

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 aligns with findings from Italy, where Bandiera et al. (2009) report that weaker governance leads public bodies to overpay for comparable goods.

Markups and Market Power Markups, here defined as the ratio of price to marginal cost,

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

play a central role in understanding market power in both theoretical and empirical economics. Recent research examined how markups reflect competitive dynamics, shaping welfare, pricing strategies, and policy interventions. Loecker and Syverson (2021) emphasize the dual role of markups. Beyond price-setting, markups encapsulate key features of imperfect competition, allowing researchers to measure deviations from competitive pricing and the deadweight loss associated with market power, as firms unwilling or unable 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, reflecting a growing concentration of market power. Hall (2018) further corroborate evidence of rising markups. Autor et al. (2020) document how increased market power has affected labor markets, particularly with the emergence of "superstar firms" that capture a disproportionate share of profits while employing fewer workers, contributing to the decline in labor's share of GDP. They argue that this concentration exacerbates income inequality, as capital captures a larger portion of economic gains.

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. 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 experience declining competition, others may become more competitive due to technological progress. Similarly, Miller (2024) attribute the rise to technological advancements, suggesting that productivity gains, reductions in marginal costs, and improvements in product quality have enabled firms to increase markups without necessarily harming consumer welfare.

Markup Econometrics Two methods are commonly used to estimate markups: one demand-based and one production-based. For measuring markups and conducting retrospective studies, the two approaches are arguably substitutes (De Loecker and Scott, 2016). However, the production approach cannot be used for counterfactual analysis, which requires a detailed understanding of demand. Recall the formula for monopoly pricing:

$$\frac{P}{c} = \frac{1}{1 + E_D^{-1}},$$

where E_D^{-1} is the inverse elasticity of demand. In more complex settings, such as those involving differentiated products, one can still solve for markups as a function of demand elasticities. The demand-based approach has been the standard method for analyzing pricing behavior, but it depends on several restrictive assumptions. First, it typically assumes static Nash-Bertrand competition (or some other form of imperfect competition that allows for a tractable equilibrium solution). Second, it requires instruments to identify demand elasticities accurately. Finally, the approach involves functional form assumptions on the demand system as well as a model of consumer heterogeneity, potentially limiting its applicability in complex or dynamic markets.

De Loecker and Warzynski (2012) (DLW) introduced a 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 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. 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 account for firm-specific price variations.

Despite these criticisms, De Ridder et al. (2024) demonstrate that revenue-based markups still reveal important trends over time and across firms, making the DLW approach valuable for studying market power patterns, even if markup level estimates require careful interpretation.

The DLW method relies on a first-order condition for the cost minimization of a variable input in production to express markups based on the output elasticity of the variable input and its revenue share. The assumption is that, in each period, producers minimize costs by choosing inputs optimally, free from frictions. The sketch of the main idea follows.

Let P_{it}^v denote the price of input v, and P_{it} the price of output. The production function is given by $Q_{it} = Q_{it}(X_{it}^1, \ldots, X_{it}^V, K_{it}, \omega_{it})$, where $v = 1, 2, \ldots, V$ indexes variable inputs. Assuming that variable inputs are set each period to minimize costs, the Lagrangian for the cost minimization problem is as follows: $\mathcal{L}(X_{it}^1, \ldots, X_{it}^V, K_{it}, \lambda_{it}) = \sum_{v=1}^V P_{it}^v X_{it}^v + r_{it}K_{it} + \lambda_{it}(Q_{it} - Q_{it}(\cdot))$.

The first-order condition for input v is $P_{it}^v - \lambda_{it}\frac{\partial Q_{it}(\cdot)}{\partial X_{it}^v} = 0$, where λ_{it} is the marginal cost of production at the production level Q_{it}. Multiplying this condition by X_{it}^v/Q_{it} leads to the following equation:

$$\frac{\partial Q_{it}(\cdot)}{\partial X_{it}^v}\frac{X_{it}^v}{Q_{it}} = \frac{1}{\lambda_{it}}\frac{P_{it}^v X_{it}^v}{Q_{it}}.$$

With $\mu_{it} \equiv P_{it}/\lambda_{it}$, this becomes

$$\frac{\partial Q_{it}(\cdot)}{\partial X_{it}^v}\frac{X_{it}^v}{Q_{it}} = \mu_{it}\frac{P_{it}^v X_{it}^v}{P_{it}Q_{it}}, \tag{1}$$

where we have multiplied and divided by P_{it}. This derivation produces a simple expression for the markup: $\mu_{it} = \theta_{it}^v(\alpha_{it}^v)^{-1}$, where θ_{it}^v is the output elasticity with respect to input v, which is generally specific to the producer and time period, and α_{it}^v represents expenditures on input v as a share of revenues. On its own, this formula is nothing new. What's new about DLW is how flexible they are about estimating θ_{it}^V and how they base their inferences about markups on careful production function estimation.

DLW adopt the control function approach, pioneered by Olley and Pakes (1996) and later refined by Levinsohn and Petrin (2003), to link the production function to the economic model describing firm behavior and the competitive environment in which firms operate. Ackerberg et al. (2015) argue that these approaches suffer from identification issues and consolidate them by introducing modified assumptions on the timing of input decisions, moving the identification of all coefficients of the production function to the second stage of the estimation, which I follow and outline in Methods Section A.

2 Data

I utilize a dataset obtained through licensed[1] access to the MagnusWeb database[2], which provides full company accounts for an unbalanced panel of 1,297 firms active in the Czech construction sector. The dataset contains 7,261 observations covering the period from 2006 to 2021. The sample size accounts for the markup estimation requirement of having at least one lag of data. For robustness of the production function estimates, the top and bottom percentiles of firms are trimmed based on the sales-to-cost-of-goods-sold ratio.

In the Czech Republic, public procurement contracts are awarded under nationwide regulations requiring procurers to publish contract details. The dataset includes contracts awarded by central, regional, and municipal governments, as well as government-owned enterprises. A private company[3] corrected and cleaned the raw data to ensure their accuracy. By linking firm financial statements with public procurement records, I construct an indicator of government tender sales to classify firms into four groups: those operating exclusively in the private sector, those that have entered public procurement, those that have exited, and those that remain government contractors.

Year	No. Firms	Share Contractors
2006	227	0.38
2007	290	0.33
2008	348	0.32
2009	412	0.32
2010	497	0.28
2011	506	0.26
2012	457	0.33
2013	235	0.41
2014	243	0.40
2015	245	0.48
2016	338	0.51
2017	660	0.57
2018	708	0.55
2019	764	0.53
2020	769	0.55
2021	562	0.53
Total	7,261	0.44

[1]Charles University Prague, Faculty of Social Sciences, Institute of Economic Studies
[2]Dun & Bradstreet Czech Republic, a.s.
[3]Datlab s.r.o.

3 Results

Section 3.1 documents the main patterns of markups in the Czech construction sector and the relationship between markups and public procurement. I validate these results using two treatment effect identification strategies. Section 3.2 employs a design-based approach relying on unconfoundedness, or selection based on observable financial statement information and pre-procurement markups. In contrast, Section 3.3 focuses on imputing the counterfactual markup evolution for government contracting firms, addressing selection on latent unobservable factors and incorporating synthetic control methods to improve robustness.

3.1 Markups and Public Procurement in the Czech Construction Sector

I use the empirical framework of De Loecker and Warzynski (2012)[4] to correlate markups with a firm's public procurement status and evaluate whether they change upon entering public procurement, controlling for input usage and aggregate trends. While I remain agnostic about a specific theoretical model, I draw on various economic mechanisms to interpret the findings. To estimate markups in the context of public procurement, I incorporate a firm's public procurement status into input demand equations, allowing it to directly affect 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 leading to Equation 1.

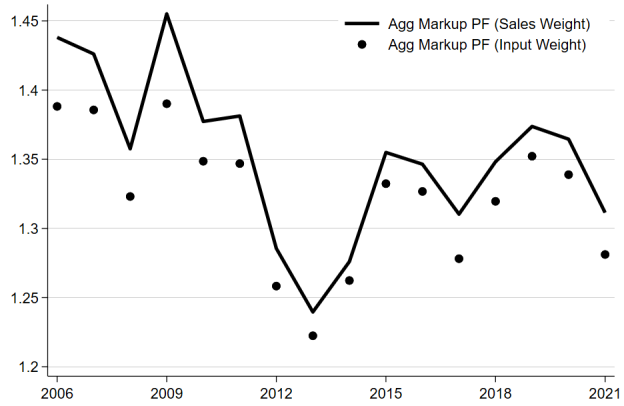
The Evolution of Markups in Czech Construction Equation 1 defines the markup as the product of the output elasticity, θ, and the inverse of the variable input's revenue share, $\frac{PQ}{P^v X^v}$. The revenue share is directly observed in firms' income statements, while the output elasticities are estimated. These elasticities are both firm- and time-specific, which reflect 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 compare it with total costs as the input weight. Figure 1 shows the evolution of average sales-weighted and input-weighted aggregate markups in the Czech construction sector.

[4]See Methods A for markup and production function estimation details.

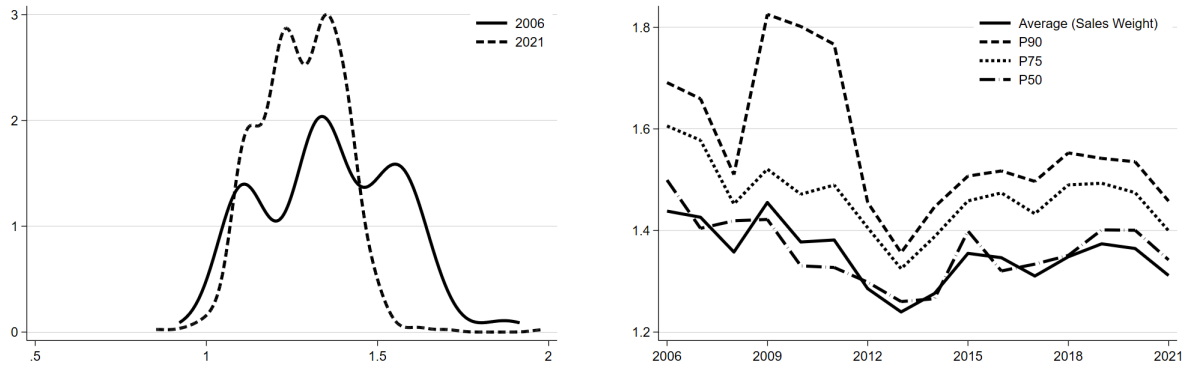
Figure 1: Average Markups. Estimated output elasticities $\hat{\theta}_{it}^{V}$ are time- and firm-specific.



Early in the sample period, markups were stable at around 1.45, declined to approximately 1.25 during 2012–2014, and rose slightly to just above 1.35 by the early 2020s. In 2006, the average markup stood at 44% above marginal cost, compared to 31% in 2021. However, averages alone fail to capture the changes in the markup distribution. The strength of the De Loecker and Warzynski (2012) method lies in its ability to estimate firm-specific markups, which enables an analysis of the entire distribution. A key finding is that a small subset of firms drive the overall decline in markups, while most firms see only modest decreases.

Figure 2a illustrates the shifts in the distribution by showing the kernel density of the unweighted markups for 2006 and 2021. Over time, the variance of markups has narrowed, driven primarily by a thinning and shortening of the upper tail. Appendix Table 5 provides the unweighted markup distribution by year.

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



(a) Kernel Density (unweighted)

(b) Percentiles (revenue-weighted)

Because the kernel density does not account for firm weights, Figure 2b shows the moments of the sales-weighted markup distribution over time. Firms are ranked by markup, and percentiles are weighted by each firm's market share. This weighting ensures comparability between the percentiles and the share-weighted average. The rankings are updated annually, meaning that the top firms may vary from year to year. Firms in the upper half of the markup distribution primarily drive the decline in average markups. The median (P50) and lower percentiles remain mostly stable over time. The sharpest drop occurs at the 90th percentile, particularly before 2012, when markups fell significantly from 1.8 to 1.4. This indicates that the overall reduction in average markups is due to a few firms experiencing substantial declines compared to earlier periods. This finding provides preliminary evidence of the evolving relationship between markups and public procurement. It is particularly relevant when viewed alongside Titl (2023), who demonstrated that the 2012 Czech public procurement reform, which eliminated single-bid contracts, reduced prices relative to estimated costs for these contracts.

Do Government Contractors Have Different Markups? I examine whether government contractors systematically have different markups compared to private-sector firms by using firm-specific markups in a regression framework. Specifically, I estimate the percentage difference in markups between government contractors and private-sector firms.

This approach ensures robustness even if the variable inputs used to compute markups are subject to adjustment costs, provided that such costs do not disproportionately affect government contractors. After estimating the regression, I convert the percentage differences into absolute markup differences.

The regression specification is as follows:

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

where pp_{it} is a dummy variable indicating a firm's public procurement status, and δ_1 measures the percentage markup premium for government contractors. The vector $\mathbf{b}_{it}$ includes control variables[5], with σ as their corresponding coefficients.

It is important to note that δ_1 is not interpreted as a causal parameter. Instead, this specification identifies whether government contractors, on average, have different markups.

[5]Control variables include variable input and capital use to account for differences in size and factor intensity, along with 47 year–subindustry interaction dummies to capture aggregate markup trends.

To the best of my knowledge, this relationship has not previously been documented, making these findings a novel contribution. While the coefficients of the control variables are not the primary focus of this analysis, I revisit the correlation between markups and other economic factors later in the paper. The regression is estimated at the manufacturing level with full year–subindustry interaction dummies. Once δ_1 is estimated, I compute the absolute markup difference by applying the percentage difference to the constant term, which represents the average markup for firms that are inactive in public procurement. This calculation is expressed as: $\hat{\mu}_{PP} = \hat{\delta}_1 \exp(\hat{\delta}_0)$. The estimated markup premium in levels is 0.327 with a delta-method standard error of 0.018. The percentage premium parameter estimate, $\hat{\delta}_1$, is 0.149, with a standard error of 0.003.

These results suggest that government contractors, on average, charge 14.9% higher markups than private-sector firms. They align with microeconomic models in which government contractors charge higher markups due to greater productivity, allowing them to outcompete rivals in tender processes. A comparison of the average markups between government contractors and private-sector firms supports this prediction and highlights its potential policy implications. Models with heterogeneous firms suggest that market share reallocates from less efficient producers to more efficient ones, with government contractors expected to exhibit higher productivity. This higher productivity enables them to cover the fixed costs associated with entering public procurement markets. However, because measured productivity is a residual of a sales-generating production function, it likely also captures unobserved differences between government contractors and private-sector firms. As such, caution is warranted when interpreting the public procurement-productivity relationship.

Public Procurement Entry and Markup Dynamics My dataset allows me to examine how markups evolve for firms entering public procurement markets, both before and after becoming government contractors. I focus on three distinct groups of government contractors identified in the sample: starters, quitters, and continuous contractors[6]. To analyze these groups, I estimate the following regression to compare markups before and after public procurement entry and exit and to measure the markup differential for firms that consistently receive government contracts:[7]

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

[6] $\text{Entry}_{it} = \mathbb{1}\{\Delta pp_{it} = 1\}$, $\text{Exit}_{it} = \mathbb{1}\{\Delta pp_{it} = -1\}$, $\text{Always}_i = \mathbb{1}\{\frac{1}{16}\sum_{t=1}^{16} pp_{it} = 1\}$

[7] I exclude 152 firms that enter and exit public procurement markets more than once during the sample period to ensure cleaner comparisons.

where $\text{Entry}_{it} = 1$ if a firm becomes a government contractor in year t, and $\text{Exit}_{it} = 1$ if a firm stops contracting with the government. The constant term, γ_0, represents the average log markup for private-sector firms, including those that enter or exit public procurement markets. The primary coefficients of interest are: - γ_1, which captures the percentage change in markups for starters (firms entering public procurement) between the pre- and post-entry periods. - γ_2, which measures the impact of public procurement exit on markups for quitters. - γ_3, which reflects the markup premium for continuous contractors—firms consistently engaged in public procurement throughout the sample period.

I compute the implied markup-level effect of public procurement entry as:

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

My results indicate that public procurement entry is associated with a significant markup increase of approximately 12%, while controlling for aggregate markup dynamics with year and subindustry dummies. All coefficients are reported in Appendix Table 8. The 95% delta-method confidence interval for the level increase ranges from 0.23 to 0.35.

These results suggest that firms entering public procurement experience substantial markup increases, highlighting the role of public contracts in shaping firm behavior. The markup premium for continuous contractors, captured by γ_3, further suggests that sustained public procurement engagement allows firms to maintain higher pricing power over time.

Discussion I report two main findings: (i) in the cross-section, government contractors exhibit 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, controlling for costs of goods sold, capital, and aggregate demand and supply effects via subindustry-year dummies. Several hypotheses may account for these findings.

More efficient producers are likely to charge lower prices, sell more in private markets, and outperform rivals in tenders for government contracts. Their cost advantage enables them to set higher markups under certain conditions, particularly when considering 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 and services differ from those in the private sector.

However, the exact mechanism driving these results cannot be conclusively identified with the available data, as I lack firm-specific price information to disentangle markup differences arising from cost versus price effects. Nonetheless, the observed within-firm increase in markups for Czech government contractors suggests that factors beyond cost differences influence pricing behavior. In summary, firms supplying construction projects to the Czech government charge higher markups than firms operating solely in the private sector, and their markups increase significantly upon entering the public procurement market. This aligns with findings by Titl (2023), Baránek and Titl (2024), and Szucs (2024), who emphasize the roles of discretion and favoritism in public procurement.

Heterogeneity Remarks The Czech construction sector, as classified by the NACE system, is divided into three primary segments: construction of buildings, civil engineering, and specialized activities (e.g., electrical and plumbing installations). Firms in all three divisions participate in public tenders to supply projects for the government. To test whether markups differ for contractors involved in civil engineering projects such as road and railway construction, I estimate Equation 2 with an interaction term for public procurement and the NACE 2-digit civil engineering division. The interaction term has a point estimate of 0.037 (standard error = 0.01), indicating that government contracts in civil engineering are associated with a 4% higher markup compared to contracts in building construction and specialized activities. Due to data limitations, I cannot test additional hypotheses, such as those related to differential quality. However, I also investigate whether markups differ for government contracts supplied by sole proprietors versus companies or cooperatives. My dataset identifies 12 sole proprietors, and the interaction term for sole proprietors is significant at 0.082, suggesting an additional markup premium of 8.2% for this group. These findings may point to potential discretion in procurement processes, as discussed in Baránek and Titl (2024). Finally, I interact the public procurement dummy with yearly indicators for 2006–2021, using 2006 as the reference year. The results reveal a declining markup premium over time, with differences relative to 2006 decreasing from around -1% to more than -4% by the end of the period. These findings, consistent with those reported in Appendix Table 9, obtained by estimating Equation 2 separately for each year, suggest that the declining public procurement markup has substantially contributed to the overall decline in aggregate markups.

Decomposing the Public Procurement Effect Markup estimates reflect a combination of private-sector and public procurement markups. While my data does not capture hours worked or employee allocation between these segments, I analyze the share of public procurement sales in total sales, interacting it with the public procurement entry dummy to assess changes in private-sector markups. I find a significant coefficient of 0.117 for γ_1 in Equation 3, corresponding to a level effect of 0.27, consistent with previous estimates. However, for firms with less than 1% of their sales from public procurement, markups increase by only 0.117%, indicating no significant change in private-sector markups. This approach assumes inputs are used in proportion to final sales, which may not hold in practice. Fully decomposing market-specific markups would require modeling a demand system and cost functions for each market. While the DLW approach avoids such assumptions, it enables meaningful cross-producer comparisons and captures how markups evolve with public procurement entry, even without disentangling market-specific effects within firms.

Robustness to Omitted Prices In estimating output elasticities,[8] I use deflated sales as a proxy for physical quantities, which may introduce omitted variable bias from unobserved prices, as noted by Klette and Griliches (1996). For price-taking firms, this is not an issue, as Equation 1 can be rewritten in terms of sales, $\partial R_{it}(\cdot)/\partial X^v_{it} = P_t \cdot \partial Q_{it}(\cdot)/\partial X^v_{it}$. However, for firms with market power, unobserved prices can bias output elasticity estimates: $\partial R_{it}(\cdot)/\partial X^v_{it} = \partial Q_{it}(\cdot)/\partial X^v_{it} \cdot (P_{it} + \partial P_{it}/\partial Q_{it})$. Such bias may arise if increased input usage reduces input prices under standard demand and cost specifications, leading to downward bias in elasticities and underestimated markups. However, this primarily affects level estimates rather than the relationship between markups and public procurement status. The average percentage markup difference is consistent as long as deviations between estimated and true elasticities are not systematically correlated with public procurement status. To mitigate this, I estimate markups using a productivity proxy, $h(\cdot)$, which accounts for price variation correlated with productivity. While unrelated demand shocks may still bias input coefficients, these do not affect the core relationship. In regressions, I use the dependent variable $\ln \hat{\theta}^V_{it} - \ln \hat{\alpha}^V_{it}$, where $\hat{\alpha}^V_{it}$ is the expenditure share, and price bias affects output elasticities $\hat{\theta}^V_{it}$. Using a translog production function, $\hat{\theta}^V_{it} = \theta^V_{it} + \rho(v_{it}, k_{it})$, where v_{it} and k_{it} are log expenditures on variable inputs and capital, and $\rho(\cdot)$ captures unobserved price deviations. I control for inputs v_{it} and k_{it} in both Equations 2 and 3, as in Bond et al. (2021, Appendix B.2).

[8] See Methods A.

3.2 Estimation of the Public Procurement Effect Under Unconfoundedness

In this section, I reanalyze data from the Czech construction sector, building on the structural inference of the underlying markup distribution, to evaluate the effect of public procurement engagement between 2010 and 2021 on firm pricing power. The primary outcome of interest is the annual contract year markup. I use multiple methods to estimate the causal effects of public procurement engagement under unconfoundedness. I adopt the potential outcomes framework, relying on two key assumptions—unconfoundedness and overlap—and address their plausibility using placebo and sensitivity analyses. [9]

The analysis focuses on a binary treatment, examining the efficiency of government tenders in the Czech construction sector compared to private markets during 2006–2021. A key challenge is the lack of detailed information on the treatment assignment mechanism. However, the availability of lagged outcomes allows for validation of the unconfoundedness assumption. The analysis is restricted to a balanced panel as a design-based decision. This approach ensures consistency over time and facilitates intuitive matching procedures, though it limits the analysis to 26 firms observed over the full 16-year period.

The control group includes 161 firm-year observations from firms without any government tender revenues between 2010 and 2021. The treatment group includes 151 firm-year observations from firms receiving government construction tender payments between 2010 and 2019. Treated firms include both recurring contractors and one-time providers. The total contract value amounts to 25.9 billion Czech Crowns in 2021, equivalent to approximately 1.2 billion USD based on the 2021 exchange rate (1 USD = 21.5 CZK).

	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**Total**
Contract Value (Billions CZK)	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
Number of Contracting Firms	7	8	9	13	11	14	12	17	17	15	15	13	151

Engagement in government contracts is not randomly assigned and cannot ensure comparability between treatment and control groups at the time of public procurement entry. However, I present results under the assumption that unconfoundedness holds when conditioned on a set of observable covariates, including the year in which the contract was awarded. Notably, I have data on past markups spanning 4 years (2006–2009) to 15 years (2006–2020) before a firm won a contract. These historical outcomes are used as either conditioning variables or placebo outcomes.

[9] See Methods B.

In the subsequent analysis, I use the natural logarithm of markups from three post-contract periods as the outcome variables, denoted as $Y_{i,0}, Y_{i,1}, Y_{i,2}$, where $t = 0$ represents the year the contract was awarded. The log markups from the three years immediately preceding the contract win $Y_{i,-3}, Y_{i,-2}, Y_{i,-1}$ and their average will serve as placebo outcomes. Additionally, the log markups from the three years preceding the placebo period $Y_{i,-6}, Y_{i,-5}, Y_{i,-4}$ will be included as covariates for adjustment. This adjustment will also incorporate 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 contract year and the NACE 2-digit division. Lagged values of the public procurement indicator are included to compare firms with similar pre-treatment histories.

Table 1: 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 26 firms already active in public procurement in 2006. The period from 2006 to 2009 is unavailable due to matching and placebo requirements of at least four years. This setup results in $N \times T = 312$ observations during 2010–2021. The measure diff/sd represents the absolute mean difference, standardized by the pooled standard deviation:

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

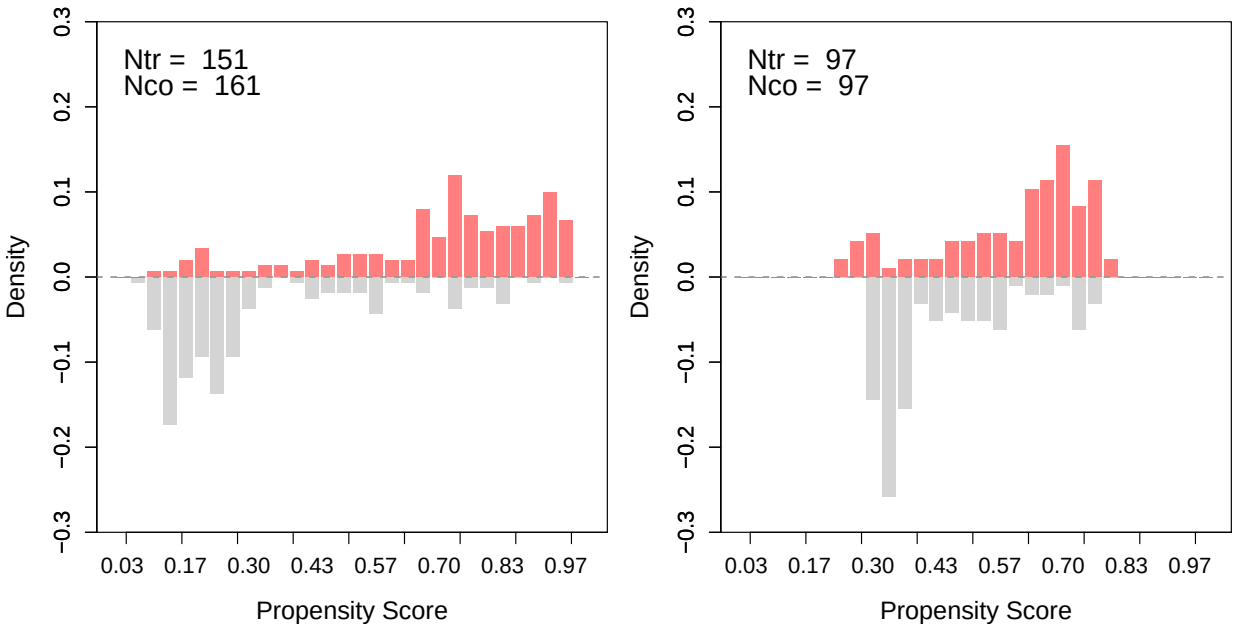
In Table 1, I present the averages and standard deviations for the observations used in the estimation sample. The sample includes firms observed over the full 16-year period, excluding those engaged in public procurement as of 2006. Additionally, I condition the sample on four lagged observations of the pre-treatment variable.

The final dataset consists of 26 firms tracked across 10 periods of instantaneous treatment effects between 2010 and 2021. This includes 161 firm-year observations for firms exclusively active in the private sector ($i, t; W_{it} = 0$) and 151 for firms engaged in government contracts ($i, t; W_{it} = 1$). Average covariate values by treatment status are reported, normalized by their respective standard deviations.

Table 1 shows that the baseline difference in average markups between treated firms (future contractors) and control firms (those active only in the private sector during 2010–2021) amounts to 0.66 standard deviations. I do not report the t-statistic for this difference, as it partially reflects sample size. Larger normalized differences clearly indicate overlap issues, with values exceeding 0.25 standard deviations generally considered substantial (Imbens and Wooldridge, 2009).

To estimate the treatment effects, the controls are used to calculate the conditional mean, $\mu_0(x) \equiv [Y_i(0)|X_i = x]$, which is then used to predict the missing control outcomes for the treated units. The significant disparity in covariate distributions between the two groups—up to 0.99 standard deviations from zero—suggests that linear regression may rely heavily on extrapolation, making results highly sensitive to the model's functional form.

Figure 3: Assessing Overlap

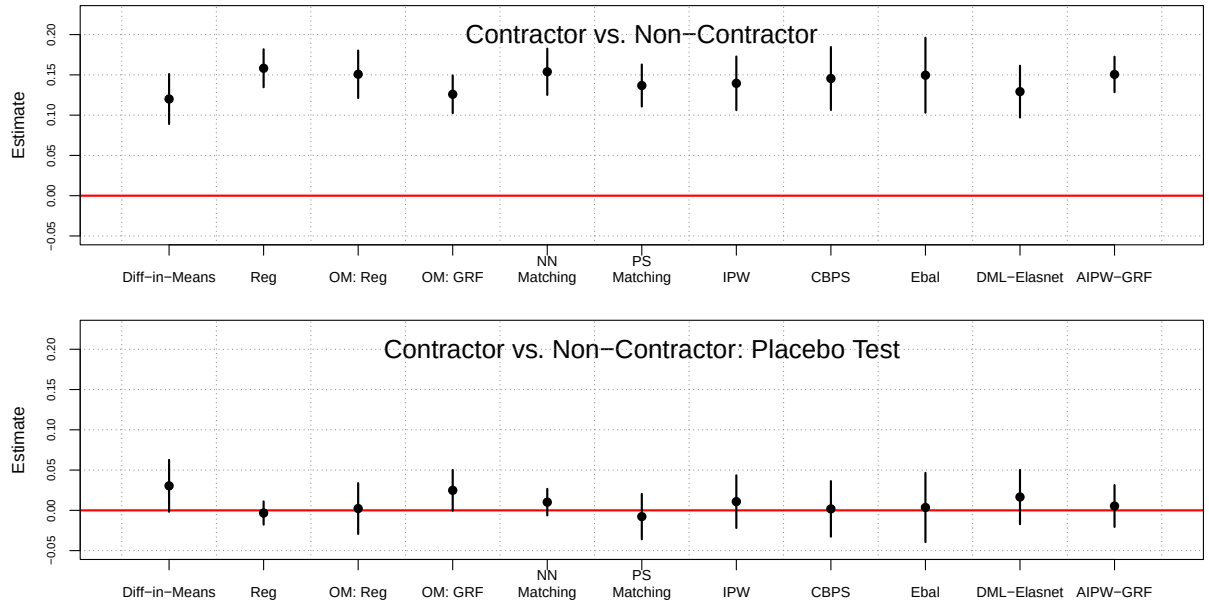


(a) Full sample (b) Trimmed sample

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 while the propensity score distribution for individuals in the treatment group differs from that of the control group, the propensity scores of the treatment group remain within the range of the control group's distribution. To improve overlap, I exclude observations with propensity scores above 0.8, removing 8 control units and 53 contractors. I refine the control group using 1:1 propensity score matching without replacement.

Figure 4: ATT Given Unconfoundedness and Placebo Estimates



Note: The top figure displays ATT (percentage) estimates comparing government contractors to non-contractors for the average annual log markups in Year 0, with 95% confidence intervals. The bottom figure shows ATT estimates for the placebo outcome (average markups from Year –3 to Year –1), with 95% confidence intervals. This analysis employs eleven estimators, 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 (using `DoubleML`), and AIPW with GRF (using `grf`).

Table 2: ATT Given Unconfoundedness and Placebo Estimates

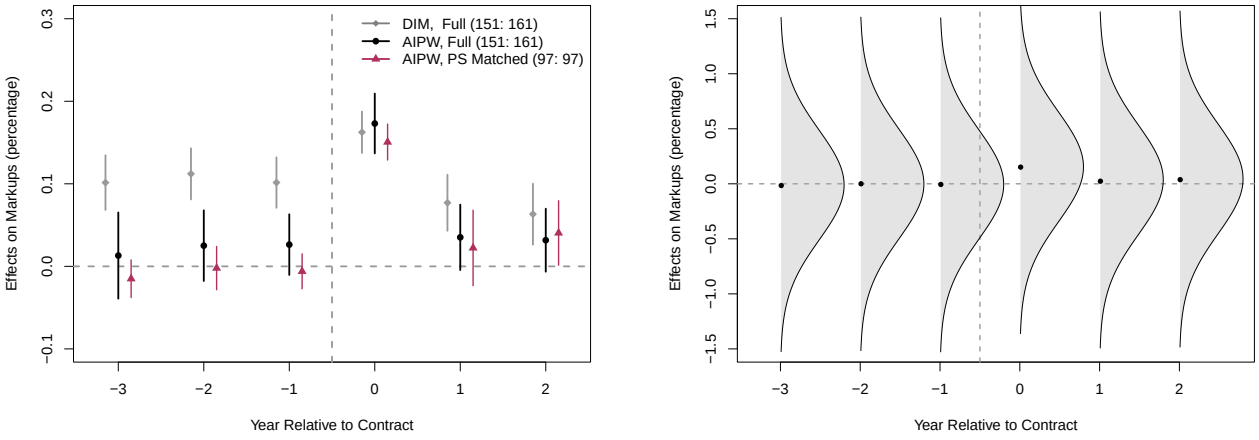
Effect on Markups	**Contract**	**Pre-Contract Average**
Difference-in-Means	0.12 (0.02)	0.03 (0.02)
Regression	0.16 (0.01)	-0.00 (0.01)
Oaxaca Blinder	0.15 (0.01)	0.00 (0.02)
GRF	0.13 (0.01)	0.03 (0.01)
NN Matching	0.15 (0.01)	0.01 (0.01)
PS Matching	0.13 (0.01)	-0.00 (0.01)
IPW	0.14 (0.02)	0.01 (0.02)
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 the placebo outcome (average markups from Year –3 to Year –1) are based on the trimmed dataset, excluding extreme propensity scores (above 0.8). Robust standard errors are in parentheses. The outcome is measured in natural logarithms and presented as percentage effects. These estimates are visualized in Figure 4.

Figure 4 displays the ATT estimates from various estimators for the real outcome (log markup in Year 0) and the placebo outcome (average log markups from Year -3 to Year -1). The figure demonstrates that different covariate adjustment methods yield consistent results, extending the descriptive results from the previous section: participation in public procurement leads to a significant increase in markups, with an average increase of up to 15% across methods. Applying doubly robust estimators, such as AIPW-GRF, yields nearly zero estimates for the placebo outcomes, further supporting the unconfoundedness assumption.

I estimate the ATT and CATT for log markups from Year -3 to Year 3, using both difference-in-means and AIPW-GRF methods. These results are presented in Figure 5. This figure resembles an event study plot often used in panel data analysis; however, the primary identification assumption here is unconfoundedness. The estimates from AIPW-GRF and difference-in-means differ substantially, with AIPW-GRF producing 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 provide strong corroborative evidence from the placebo tests: the CATT estimates are close to zero in the pre-contract years and indicate minimal treatment effect heterogeneity among government contractors.

Figure 5: ATT and CATT Estimates

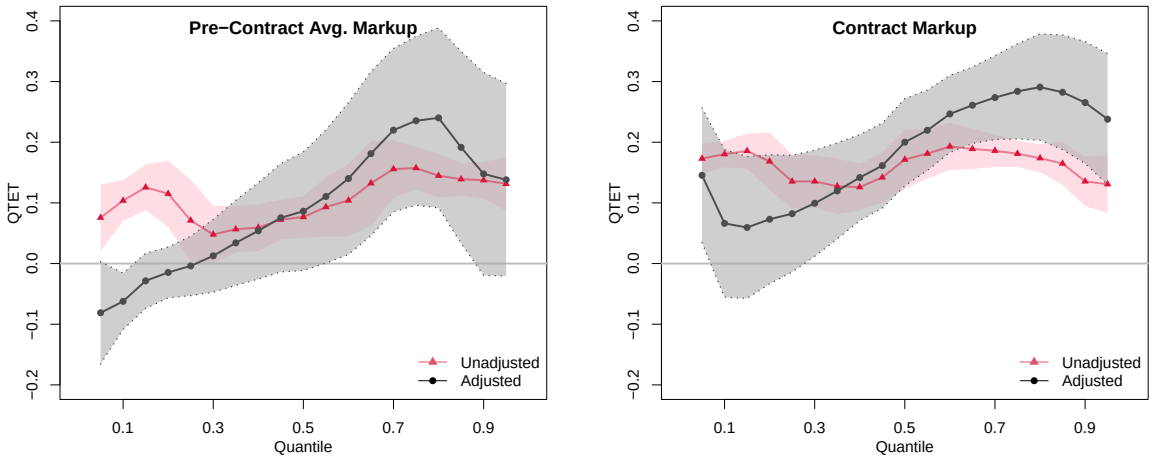


(a) ATT (b) CATT

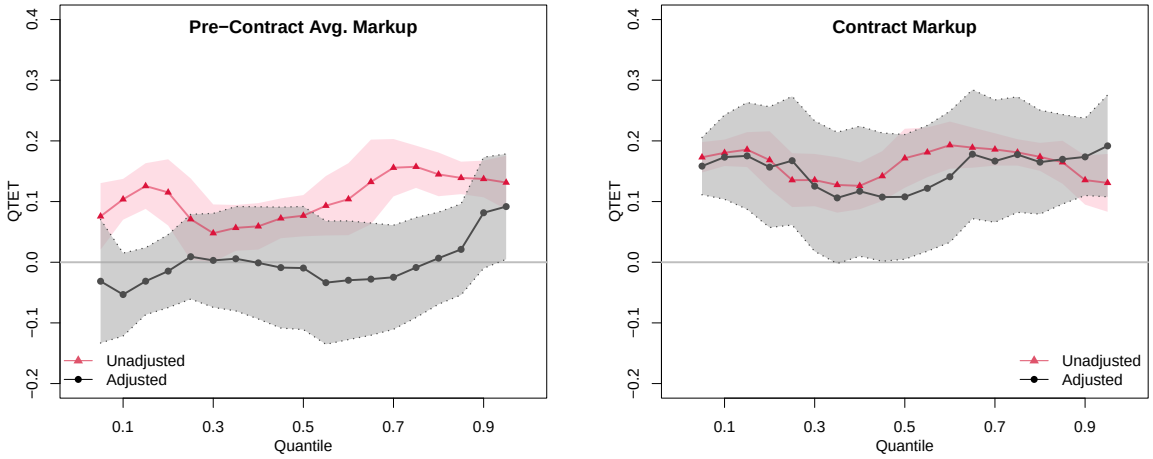
Note: Figures show the ATT and CATT estimates. The outcome variables include markups from 3 years before winning to 6 years after winning (Years –3 to 6). Estimates for pre-contract outcomes (Years –3 to –1) are used as placebo tests, validating the unconfoundedness assumption.
Subfigure A displays ATT estimates using 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 shows the distribution of CATT estimates using AIPW-GRF for the trimmed sample, with black dots representing the corresponding ATT estimates.

I also estimate the quantile treatment effects on the treated (QTET) using the IPW approach (Firpo, 2007). Figure 6 plots the QTET estimates for both the full sample and the trimmed sample.

Figure 6: Quantile Treatment Effects



(a) Contractors vs Controls (Full Sample)



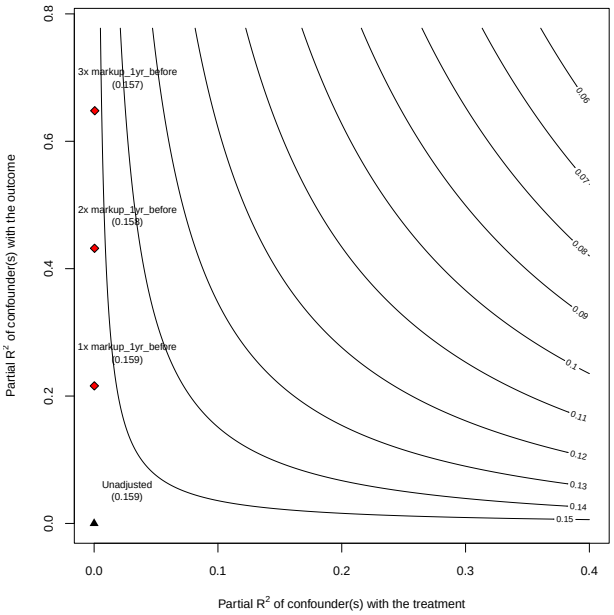
(b) Contractors vs Controls (Trimmed Sample)

Note: Figures show the quantile treatment effects on the treated (QTET) with or without adjusting for covariates (in gray and pink, respectively). Each dot represents a QTET estimate for a specific quantile, with gray or pink areas indicating bootstrapped 95% confidence intervals. Unadjusted models exclude covariates, while adjusted models incorporate the full set of covariates to estimate propensity scores using a logit model.
Subfigure A: Results for the full sample. **Subfigure B**: Results for the trimmed sample.

The figure illustrates how trimming the sample enhances the plausibility of the unconfoundedness assumption. In the trimmed sample, the adjusted pre-contract average placebo treatment effects are consistently non-significant across quantiles. However, in the full sample, significant adjusted placebo treatment effects are observed for quantiles above the median prior to public procurement engagement, which may be attributable to selection bias.

Finally, Figure 7 presents results from a sensitivity analysis. The estimated causal effect of public procurement is robust to potential confounders. For instance, the estimate remains substantial and positive (0.157) even when a confounder's correlation with treatment and outcome is tripled relative to that of Y_{-1}.

Figure 7: Unconfoundedness Sensitivity Analysis



Note: This contour plot illustrates the treatment effect estimate, $\hat{\tau}_{OLS}$, based on the sensitivity analysis framework introduced by Imbens (2003) and extended by Cinelli and Hazlett (2020). The benchmark covariate is log markup one year before the contract (Y_{-1}). The model is a linear regression including all available long-term covariates.

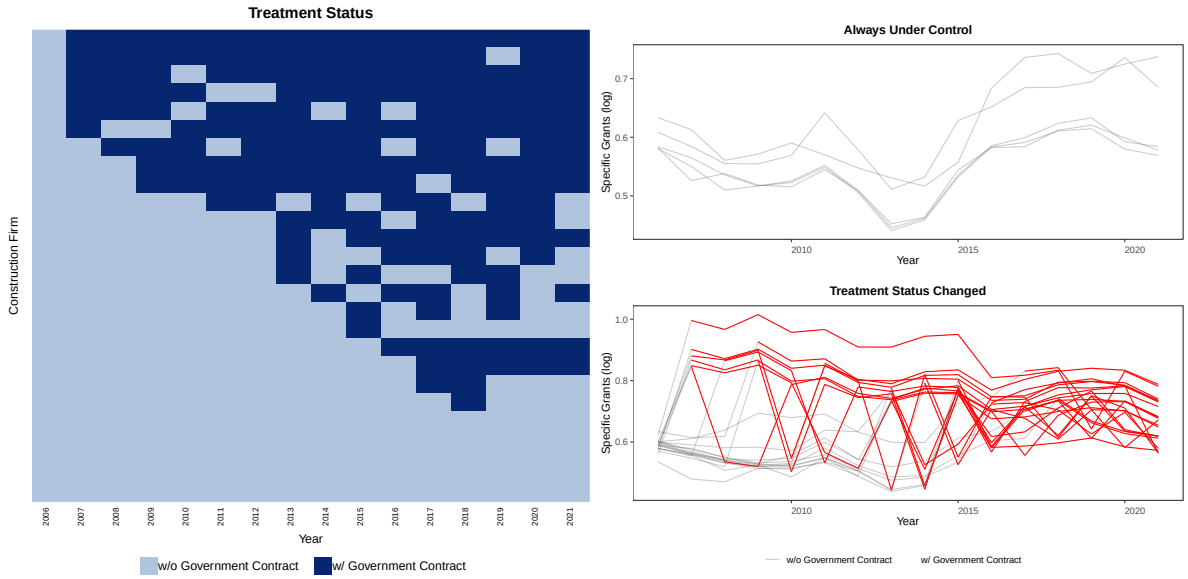
Overall, placebo tests provide strong evidence supporting the unconfoundedness assumption, enhancing the credibility of the causal estimates. Although limited knowledge of the treatment assignment process poses challenges for justifying unconfoundedness, lagged outcomes likely capture both selection and outcome variables, serving as reliable placebo outcomes.

In sum, the propensity score was estimated using flexible methods and overlap assessed through propensity score distributions, which lead to trimming the data accordingly. Modern techniques, including doubly robust estimators, were applied to estimate average causal effects. To investigate treatment effect heterogeneity, alternative estimands were considered, including the conditional average treatment effect (CATT) and quantile treatment effects (QTET). Placebo tests with pretreatment outcomes validate the unconfoundedness assumption, while sensitivity analyses strengthen the robustness of the findings.

3.3 Model-based Estimation of the Public Procurement Effect

I extend my analysis of whether switching from private-sector operations to government contracts increases markups for construction firms in the Czech Republic during 2006–2021 by employing a panel data framework outlined in Methods Section C. The outcome variable is the natural logarithm of the markup for firm i during year t. Figure 8 visualizes the patterns of the treatment and outcome variables. I focus on results from the balanced panel, excluding always-treated observations. This yields a sample of $N = 26$ firms observed over $T = 16$ years, allowing me to compare this approach to the unconfoundedness analysis while retaining the time series structure of the panel data. The treatment is redefined as a dummy variable indicating whether a firm has ever received revenue from government tenders, and therefore employs a staggered adoption design.

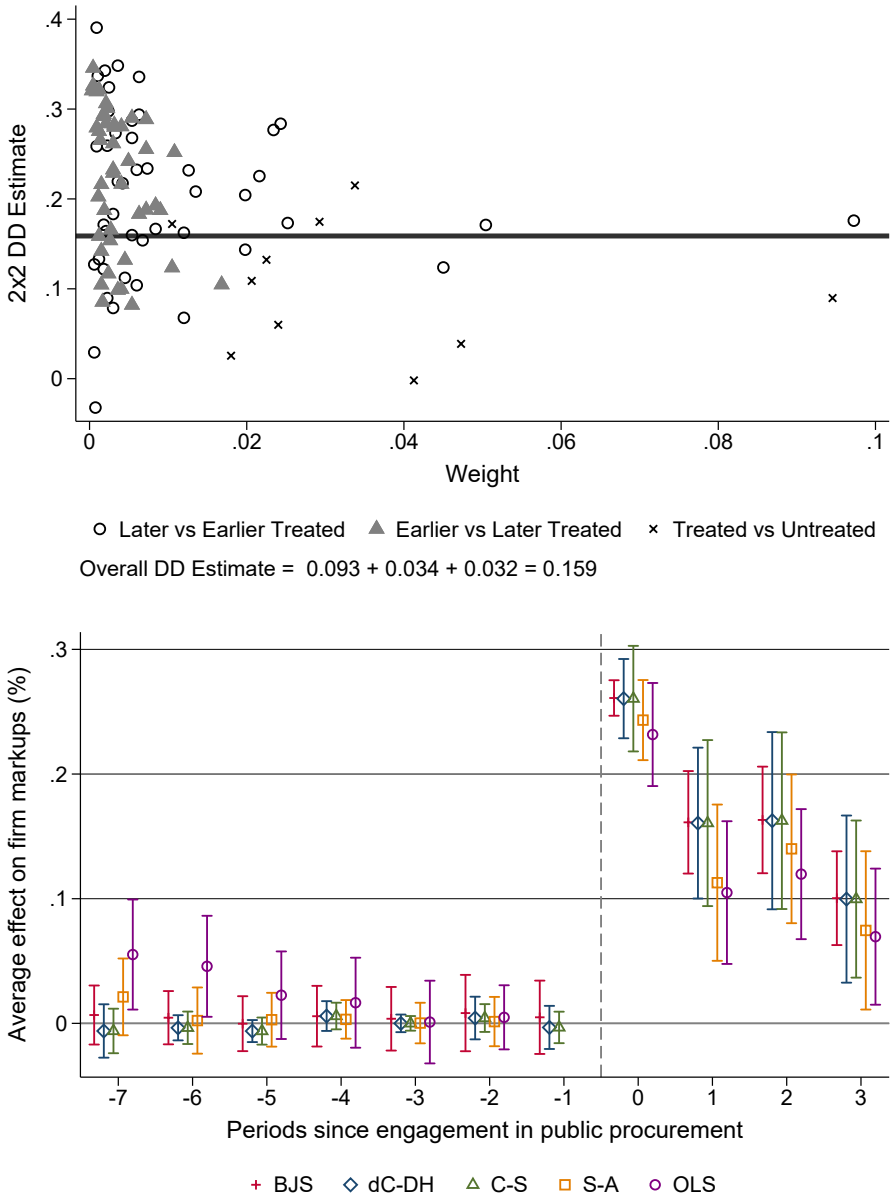
Figure 8: Balanced Panel Redefined Treatment and Outcome Visualization



Note: After excluding firms already active in public procurement in 2006, I visualize the absorbing treatment status using the `panelView` package (Mou et al., 2023). Treated observations are shown in deep blue, while control status observations are displayed in a lighter shade of blue. I depict the trajectory of the outcome variable over the study's time window for each individual unit in the balanced panel, with control units shown in gray and treated units in red.

This may be problematic, as treatment reversals occur within the dataset. However, it allows for the decomposition proposed by Goodman-Bacon (2021) and the comparison of alternative DID-type estimators, which are robust to time- and cohort-specific heterogeneity in the effects of public procurement on markups under a staggered adoption setting. Figure 9 presents the results, including the estimated TWFE coefficient of 0.159, with a standard error of 0.023.

Figure 9: Balanced Panel TWFE Decomposition and Alternative DID-type Estimators



Note: Top figure: The decomposition from Goodman-Bacon (2021), which breaks down the TWFE estimate into a weighted average of all possible 2×2 DID estimates across different cohorts. Bottom figure: ATT estimates for the average log markups from Year –5 to Year 3, with 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.

Goodman-Bacon (2021) demonstrate that the TWFE estimator in a staggered adoption setting can be represented as a weighted average of all possible 2×2 difference-in-differences (DID) estimates between different cohorts. However, when treatment effects change heterogeneously over time across cohorts, the "forbidden" comparisons, which use post-treatment data from early adopters as controls for late adopters, may introduce bias into the TWFE estimator. The decomposition shows that the estimates from the DIDs comparing ever-treated cohorts switching into treatment and other ever-treated cohorts still in their pre-treatment 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 (the crosses, labeled "Treated vs. Untreated") rank second in terms of contribution (weight 0.34). The "forbidden" DIDs, which compare ever-treated cohorts entering treatment with other ever-treated cohorts already treated (the circles, labeled "Later vs. Earlier Treated"), receive the most weight at 0.49.

I first use the dynamic TWFE regression, which includes interaction terms between a dummy indicating whether a unit is treated and each lead/lag indicator relative to treatment. This specification allows treatment effects to vary over time and uses the period immediately preceding treatment as the reference period. If the regression is saturated and there are no heterogeneous treatment effects across cohorts, this dynamic TWFE can consistently estimate the dynamic treatment effect.

Additionally, to address heterogeneous treatment effects across cohorts, I apply methods designed for this purpose, such as those 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 these heterogeneity-robust estimators are substantively similar to those from conventional TWFE models. However, Sun and Abraham (2020) emphasize issues with negative weights that become apparent when visually validating the two-way model. Testing for parallel trends in two-way specifications with treatment leads commonly involves pre-treatment data comparisons that include negative weights. In my application, OLS "pre-trends" become significant at further treatment leads, unlike the alternative estimators.

As Roth (2024) highlights, some HTE-robust estimators produce asymmetry between pre- and post-treatment coefficients, requiring cautious interpretation. This asymmetry, therefore, motivates the following approach to obtain robust confidence intervals.

First I use the `fect` package (Liu et al., 2022) and the imputation method proposed by Borusyak et al. (2021), Gardner (2022), and Dube et al. (2023) to estimate an ATT of 0.142 with a standard error of 0.018.

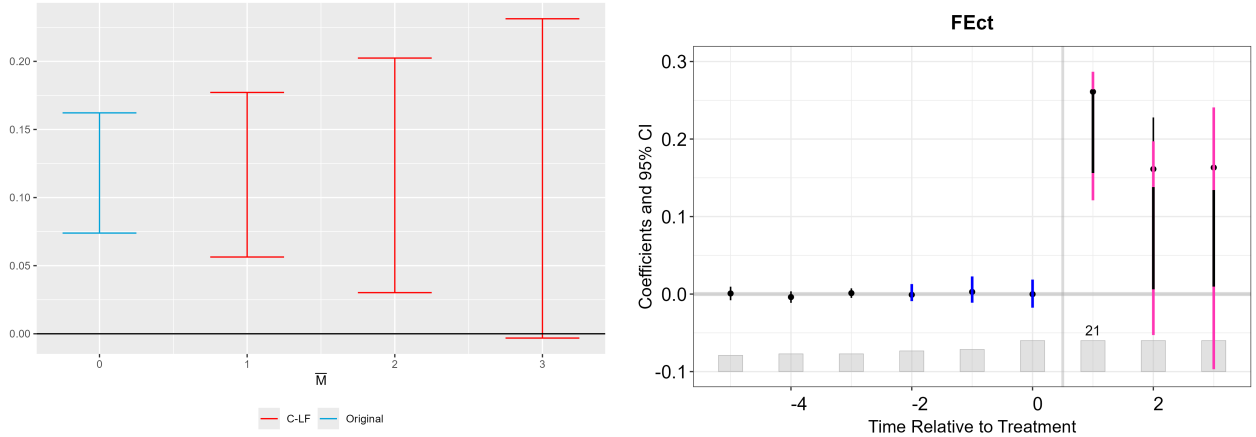
The contemporaneous average treatment effect in the year of receiving a government contract (averaged over first-time treatments among 21 firms that engaged during the sample period) is 0.261 (see BJS in Figure 9, $t = 0$) with a cluster bootstrap standard error of 0.021.

Next, I exclude the three placebo periods immediately preceding treatment onset from the model fitting process, impute counterfactuals for placebo and post-treatment periods, and calculate event-study coefficients. This adjustment excludes observations from first years of the sample, aligning it with the sample analyzed in Section 3.2, lowers the contract-year ATT point estimate to 0.2, and ensures symmetric coefficients for post-treatment and placebo periods.

I then compute robust confidence sets (CSs) proposed by Rambachan and Roth (2023), which allow PTA violations under relative magnitude (RM) restrictions. These restrictions limit post-treatment PTA violations to at most $\bar{M}$ times the maximum pre-treatment placebo period violation. Robust confidence sets ensure uniform validity for partially identified treatment effects under the RM restriction. Rambachan and Roth (2023, p. 2653) suggest $\bar{M} = 1$ as a "natural benchmark," assuming post-treatment PTA violations are no worse than observed pre-trends.

In Figure 10, the robust confidence set for the ATT in the three post-treatment periods, estimated by `fect`, is shown in red. The sensitivity analysis yields a breakdown value of $\tilde{M} \approx 3.0$, indicating that the public procurement effect is robust to PTA violations under the RM restriction. I also calculate separate robust confidence intervals for the three post-treatment periods. Placebo periods are shown in blue and robust confidence intervals in pink.

Figure 10: Balanced Panel Absorbing Treatment Sensitivity Analysis



Note: The right panel shows robust confidence intervals for the ATTs in three post-treatment periods, calculated using RM restrictions by Rambachan and Roth (2023). Placebo periods are in blue, and robust confidence intervals are in pink. The left panel shows the sensitivity analysis breakdown $\tilde{M}$, indicating the average effect across three post-contract years is robust to three times as large PTA violations as observed in the three years prior to the contract.

As is well known (Ashenfelter and Card, 1985), DID-type estimators assume that, absent public procurement, log markups in private sector-only firms would have evolved in parallel. To test the robustness of my findings, I analyze the balanced panel using Synthetic Difference in Differences (SDID) and Synthetic Control (SC) estimators adapted to staggered treatment settings. Figure 11 compares the ATT for a first-time government contractor in 2018 using SDID, DID, and SC.

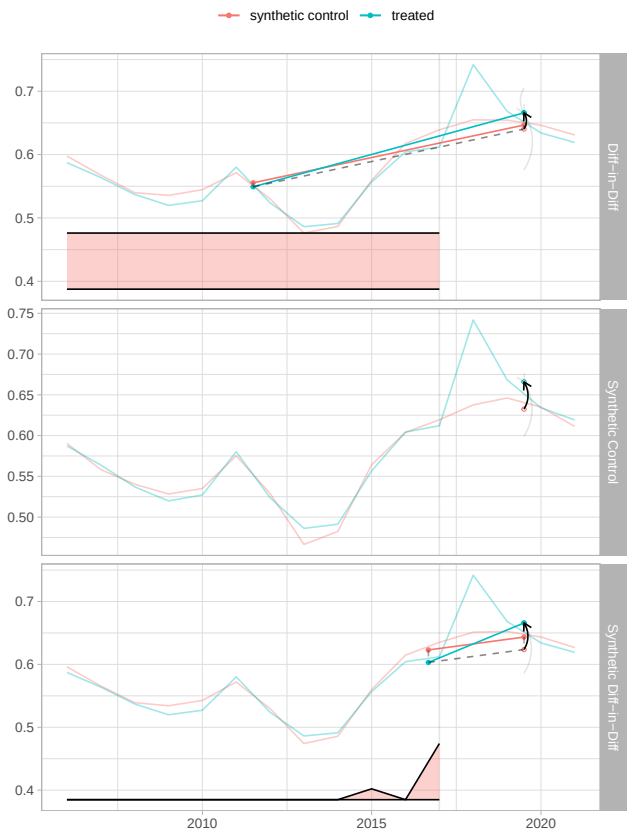


Figure 11: Comparison of Difference-in-Differences, Synthetic Control, and Synthetic Difference-in-Differences estimates for the effect of government contracts on firm markups (in logs). The markup trend for the firm treated in 2018 and the relevant weighted average of control firms are indicated by an arrow.

In this case, pre-intervention trends are not parallel, making the DID estimate suspect. SC re-weights unexposed firms to match government contractors' pre-intervention trends, attributing post-intervention divergence to the treatment. SDID re-weights control units to align their time trends with those of government contractors (while allowing for level differences) pre-intervention and then applies DID to the re-weighted panel. Additionally, SDID uses time weights, focusing only on a subset of pre-intervention periods to ensure the weighted average of historical outcomes predicts control outcomes during the treatment period ($\lambda_{2017} = 0.838$, $\lambda_{2015} = 0.162$ in Fig. 11).

Figure 12 shows the trends and time weights for the 2011–2017 cohorts.

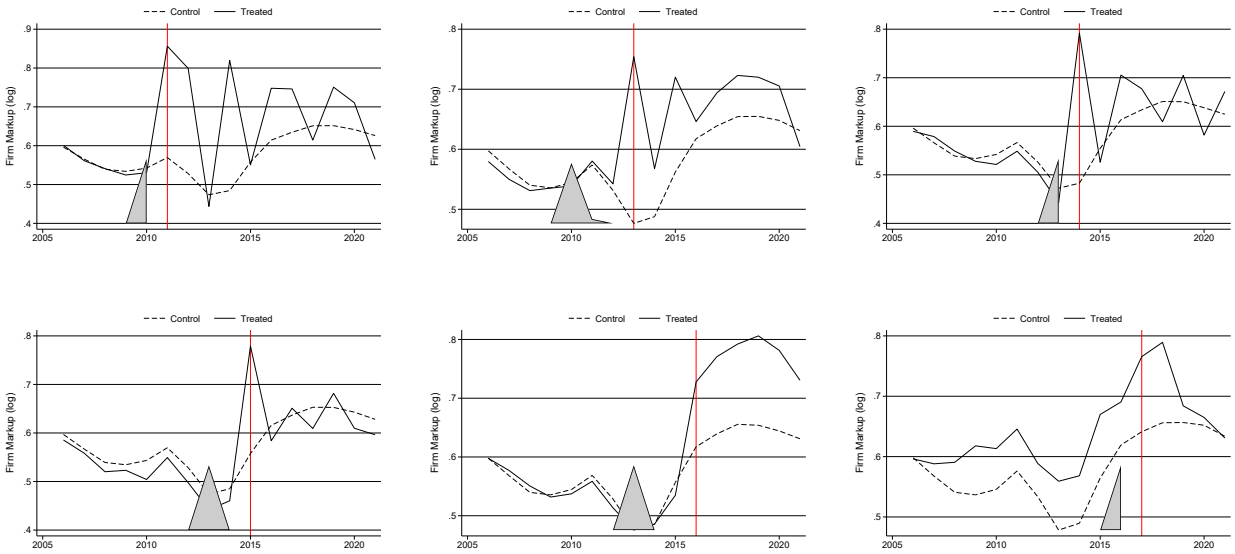


Figure 12: Synthetic Difference-in-Differences trends and time weights for cohorts (2011–2017).

It is useful to contrast the data-driven SDID approach for selecting the time weights with both DID, where all pre-treatment periods are given equal weight, and event studies, where typically the last pre-treatment period is used as a comparison and so implicitly gets all the weight (Freyaldenhoven et al., 2019; Borusyak et al., 2021).

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 treated units, and it emphasizes periods that are on average similar to the treated periods. This localization provides two advantages compared to the standard DID estimator. Intuitively, using only similar units and similar periods makes the estimator more robust. Less intuitively, the weights can improve the estimator's precision by removing systematic thus predictable parts of the outcome. Together, these weights make the DID strategy more plausible.

Researchers often adjust for covariates or select appropriate time periods to address lack of parallel time trends for treated and control units (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).

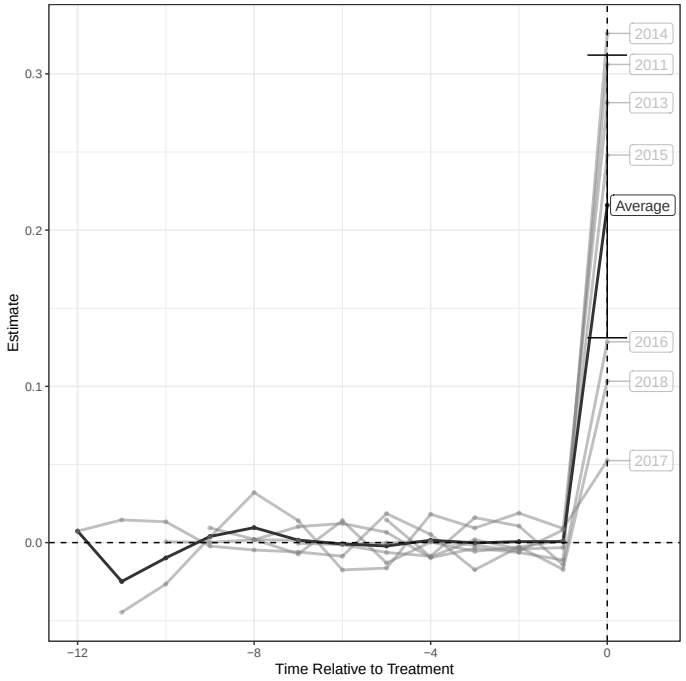
Table 3 presents the cohort-specific ATT[10] estimates and standard errors, calculated using the algorithms described in Clarke et al. (2023) (Appendix Algorithms 1 and 2).

Table 3: SDID Overall Cohort Level ATTs

	2011	2013	2014	2015	2016	2017	2018
$\hat{\tau}_a^{sdid}$	0.118	0.090	0.082	0.050	0.128	-0.012	0.042
Standard Error	(0.013)	(0.035)	(0.016)	(0.025)	(0.023)	(0.020)	(0.005)
No. Treated	1	4	1	2	1	2	1

To highlight notable features of the $\hat{\tau}_a^{sdid}$ estimates, Figure 13 visualizes the pre-treatment fit and the "on-impact" effect of public procurement using the Augmented SC approach (Ben-Michael et al., 2022). The figure corroborates the SDID results and shows a pronounced downward trend in the markup premium over time.

Figure 13: Augmented Synthetic Control: Cohort Aggregated On-impact ATTs $\hat{\tau}_{a,1}^{augsc}$

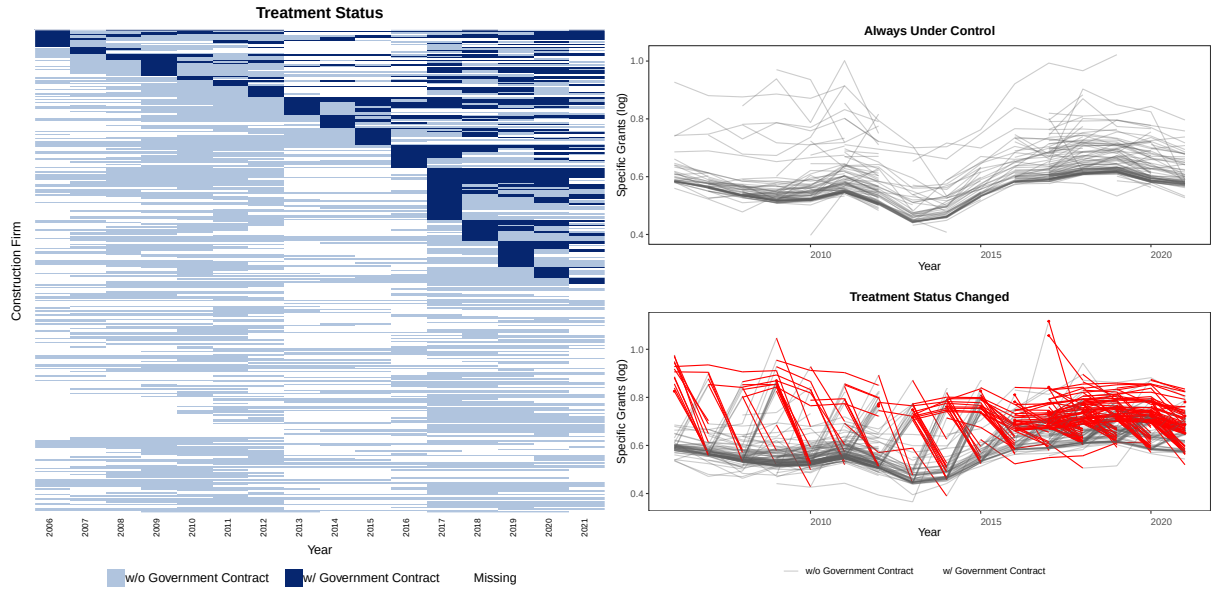


My balanced panel analysis improves the internal validity of the estimates as it ensures that there are no compositional changes or missing data patterns. It also allows for direct comparison with results from Section 3.2. Yet, sample restrictions limit the generalizability of my findings.

[10]Appendix Figure 17 reports the aggregated overall SDID ATT of 0.075, with a standard error of 0.019, along with the individual event-study estimates $\hat{\tau}_\ell^{sdid}$ for firms in the balanced panel that operated exclusively in the private sector for at least five periods before receiving a government contract. To improve the synthetic control aspect of pre-period matching, I exclude firms in treatment cohorts before 2011.

To address this limitation, I conclude my analysis by considering the full unbalanced panel with a general assignment pattern. Figure 14 visualizes the outcomes and treatment statuses for the estimating sample, which I restrict to firms with at least five periods of operations exclusively in the private sector to enable more credible identification of individual fixed effects.

Figure 14: Visualization of Treatment and Outcomes: Firms with at Least Five Untreated Periods

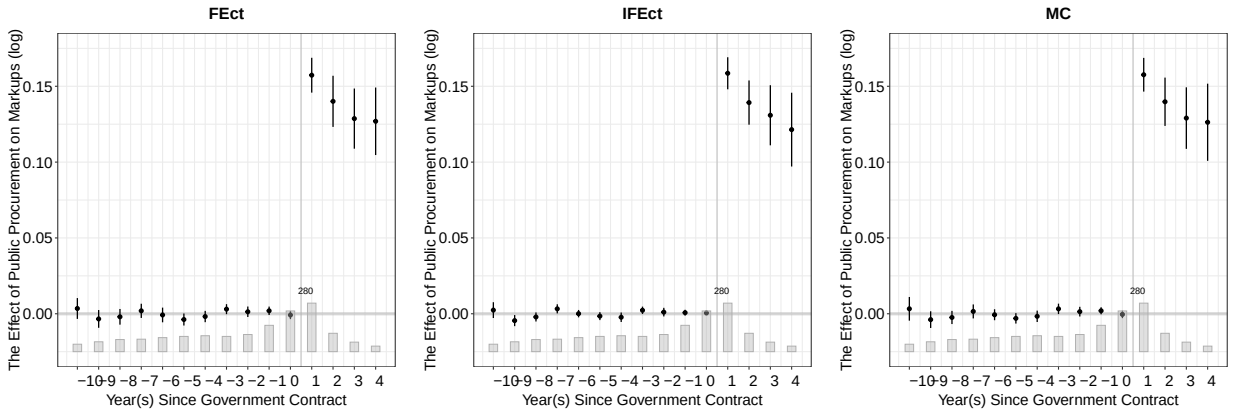


Note: On the left, firms are aggregated by unique treatment histories. Treated statuses are shown in deep blue, and control statuses in lighter blue. On the right, the trajectories of the natural logarithm of markups (2006–2021) are shown for firms with five or more untreated periods. Control units are displayed in gray, and treated units in red.

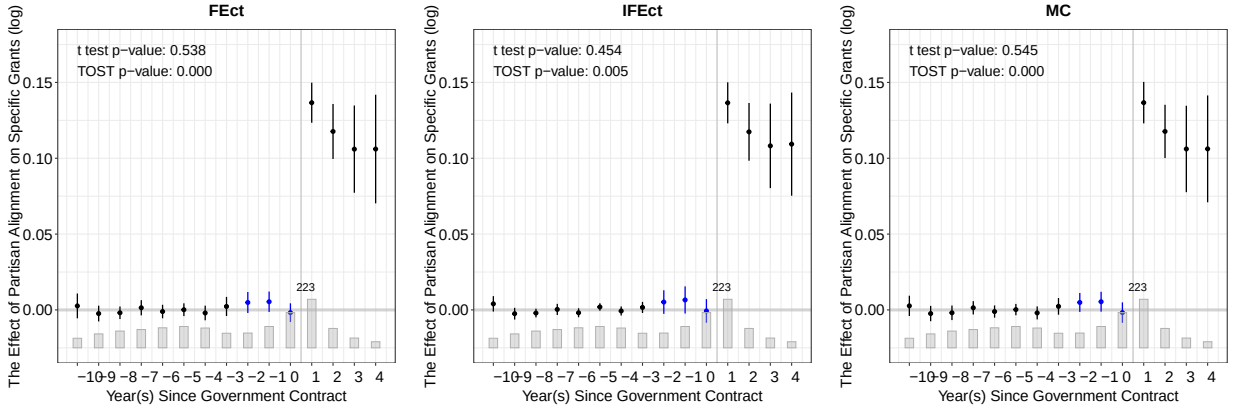
I use the imputation TWFE (`fect`), interacted factor (`ifect`), and matrix completion (`MC`) estimators, employing three pre-treatment periods for placebo tests and obtaining uncertainty estimates with a clustered bootstrap at the unit level. Figure 15(a) presents the estimated dynamic treatment effects, with residual averages flat and centered around zero during the pre-treatment periods. The ATT point estimates and standard errors, rounded to three decimals, are 0.150 and 0.006, respectively, across all three methods. In contrast, the classic TWFE estimator, which uses all observations and assumes already-treated units as controls, yields a point estimate of 0.186 with a standard error of 0.004. Figure 15(b) reports placebo test results: I fail to reject the null hypothesis of no placebo effect (t-test, $p > 0.4$) and reject the null of a placebo effect exceeding the equivalence threshold (TOST, $p \leq 0.005$). Finally, Figure 15(c) shows no evidence of carryover effects three years after the end of public procurement engagement. [11]

[11] See Methods C for the statement of assumptions and a description of the diagnostic tests.

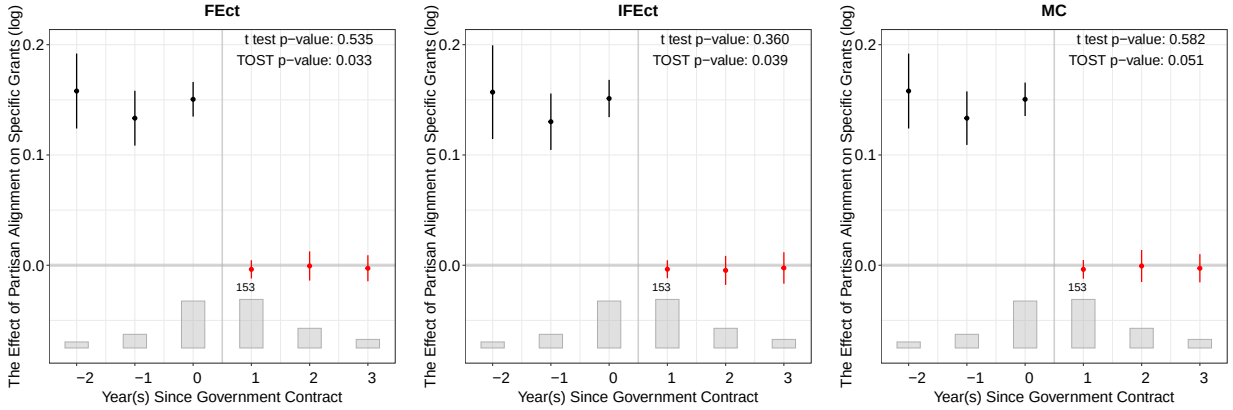
Figure 15: Non-absorbing Treatment Effect of Public Procurement on Markups: Counterfactual Estimators



(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 the TOST results, are shown at the top corners of panels (b) and (c).

Overall, the results lend support to the validity of the identifying assumptions. I specify 2 factors to be extracted from principal components of residuals for the `IFEct` model. The `fect` model assumes rank = 0 for factors beyond unit and time fixed effects and slightly outperforms `IFEct` in the diagnostic tests. For matrix completion (`MC`), the optimal λ from cross-validation exceeds 15 of the 16 singular values of the matrix L, whose rank equals the number of periods, which is the minimum of total firms N and periods T. The matrix completion approach employs data-driven regularization to estimate a rank-1 factor model for the outcome, which is residualized from fixed two-way effects in the pre-treatment test sample.[12] In the test sample, this method reduces residual variation more effectively than both `IFEct` and `FEct`, and leads to lower standard errors. [13]

I present the year-aggregated ATT estimates from the matrix completion method in Table 4. The results provide further evidence of a decreasing trend in the public procurement premium over time for firms in the Czech construction sector. Specifically, my estimates suggest that, compared to firms operating only in the private sector, government contractors had markups approximately 30% higher from 2006 to 2009, over 20% higher during the period 2010–2015, and about 10% higher during 2016–2021—only one-third of their initial level.

Table 4: Matrix Completion Year-Aggregated ATT Estimates

Year	ATT	Standard Err.	No. Treated
2006	0.294	0.016	12
2007	0.282	0.022	7
2008	0.300	0.014	8
2009	0.337	0.010	18
2010	0.273	0.017	9
2011	0.241	0.012	11
2012	0.236	0.013	18
2013	0.262	0.010	18
2014	0.301	0.008	20
2015	0.219	0.008	31
2016	0.092	0.008	34
2017	0.105	0.008	77
2018	0.100	0.006	68
2019	0.101	0.006	74
2020	0.118	0.007	78
2021	0.098	0.008	58

[12] See Methods C for visualization.

[13] Appendix Figure 16 presents the balanced panel equivalent of Figure 15, where specifying the rank with `IFEct` outperforms `MC` in diagnostics, though the treatment effect estimates are larger in magnitude but less precise.

4 Conclusion

Using firm-level data on publicly traded firms in the Czech construction sector, this study employs a method that accommodates diverse price-setting strategies and nonconstant returns to scale to estimate firm-specific markups. It describes the evolution of market power through the distributional characteristics of markups. Between 2006 and 2021, markups declined from 40% to nearly 30%, a decline primarily driven by firms with the highest initial markups. Over time, the distribution of markups became less skewed, with a thinner upper tail.

The decline in Czech construction aggregate markups is linked to the dynamics of the public procurement markup premium. The relationship is validated using policy evaluation designs to account for both observable and unobservable selection bias.

In the balanced sample, the year-of-contract-treatment effect is estimated to be 15% under selection on observable lagged variables, while the same-sample fixed-effects analysis yields a point estimate of 20%. Using both the unconfoundedness approach and the TWFE model to test robustness to alternative identifying assumptions leads to the bracketing property bounds on the estimated causal effect (Angrist and Pischke, 2009, pp. 185).

The analysis also integrates Synthetic Difference-in-Differences re-weighting and matching on pre-exposure trends to reduce reliance on parallel-trend assumptions. This approach accounts for additive unit-level shifts while utilizing data-driven synthetic control methods.

Finally, the full sample estimation employs interactive fixed effects and matrix completion methods. The framework is validated using diagnostic tests of out-of-sample predictions for untreated potential outcomes, which are robust to model misspecification and overfitting. Subgroup analysis indicates that the overall effect is primarily driven by government contracts awarded in the early years of the sample. The estimated average public procurement effect on government contractors is 30% in 2006 and decreases to 10% by 2021.

These findings contribute to the public procurement literature by establishing a relationship between markups and public procurement. They also provide a benchmark for assessing the competitiveness of government projects relative to those in the private sector. The results provide empirical support for reforms aimed at eliminating single-bidding practices and increasing transparency in the tendering process that prevents favoritism in public procurement. They offer policy-relevant insights, highlighting the role of institutional improvements in enhancing the efficiency of public spending.

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. (2009). *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 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. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

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.

Blundell, R. and Bond, S. (1998). *Initial conditions and moment restrictions in dynamic panel data models*. em Journal of Econometrics 87(1):.

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., 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.

Clarke, D., Pailanir, D., Athey, S., and Imbens, G. (2023). Synthetic difference in differences estimation. *arXiv preprint*.

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 Scott, P. T. (2016). Estimating market power evidence from the us brewing industry. Working Paper 22957, National Bureau of Economic Research.

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.

Dube, A., Girardi, D., Jordà, , and Taylor, A. M. (2023). A local projections approach to difference-in-differences. Working Paper 31184, 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*. em arXiv preprint arXiv:2406.00827.

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.

Loecker, J. D. and Syverson, C. (2021). An industrial organization perspective on productivity. In *Handbook of Industrial Organization, Volume 4*, pages 141–223. Elsevier.

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.

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*.

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.

Methods and Additional Results

A Markup Estimation

As in De Loecker et al. (2020), 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_{it} 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 variable inputs adjust frictionlessly within each period, while capital is subject to frictions. The firm's cost minimization problem can be represented by the following Lagrangian objective function:

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

where $P_{V_{it}}$ is the price of the variable input, r_{it} is the user cost of capital, F_{it} is the fixed cost, andλ_{it} is the Lagrange multiplier, which represents the marginal cost of production. The first-order condition with respect to the variable input V_{it} is:

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

Multiplying through by $\frac{V_{it}}{Q_{it}}$ and rearranging terms, the output elasticity of input V_{it} is:

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

The Lagrange multiplier λ_{it} represents marginal cost, and the markup is defined as the ratio of price to marginal cost:

$$\mu_{it} = \frac{P_{it}}{\lambda_{it}}.$$

Substituting for λ_{it} in terms of the output elasticity θ_{it}^V, the markup can be expressed as:

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

Production Function Estimation Following De Loecker and Warzynski (2012), consider the 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 common technology parameters, and ω_{it} is a firm-specific productivity shock.

A log-linear version of the production function, incorporating unobserved productivity shocks ω_{it} and measurement error ϵ_{it}, is:

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

wherey_{it} is logged output, $f(v_{it}, k_{it}; \beta)$ is the log production function, v_{it} represents the vector of variable inputs, and ϵ_{it} includes unanticipated production shocks and measurement error.

To estimate the parameters of the production function (β) and compute θ_{it}^V, I use the demand for variable inputs as a proxy for productivity:

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

where z_{it} are additional state variables (e.g., input prices or public procurement status). By inverting the input demand function $v_t(\cdot)$, I recover estimates of unobserved productivity ω_{it}.

Using a second-order translog production function, the specification is:

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

In the first stage, I estimate expected output as: $y_{it} = \varphi_t(v_{it}, k_{it}, z_{it}) + \epsilon_{it}$, where $\varphi_t(\cdot)$ captures the systematic component of output. The first-stage estimate of expected output, $\hat{\varphi}_{it}$, is:

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

where $h_t(\cdot)$ accounts for productivity variations driven by inputs and firm-specific factors.

The second stage models the law of motion for productivity as a first-order Markov process:

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

where ξ_{it} is the productivity innovation, and p_{it-1} represents lagged decision variables (e.g., public procurement status). By estimating ω_{it} as:

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

I recover the residual $\xi_{it}(\beta)$ through a non-parametric regression of $\omega_{it}(\beta)$ on $\omega_{it-1}(\beta)$ and p_{it-1}.

Identification and Moment Conditions To estimate the production function parameters, I form moment conditions based on the productivity innovations:

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

where v_{it-1} and k_{it} are lagged variable and capital inputs. These moment conditions are estimated using Generalized Method of Moments (GMM). The identification strategy assumes that capital is predetermined (decided in advance) and uncorrelated with ξ_{it}, while variable inputs respond flexibly to productivity shocks within the period.

Output Elasticities and Markups Under the translog specification, the output elasticity of the variable input V is: $\hat{\theta}_{it}^V = \hat{\beta}_v + 2\hat{\beta}_{vv} v_{it} + \hat{\beta}_{vk} k_{it}$. The markup is then computed as:

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

where $\hat{\alpha}_{it}^{V}$ is the expenditure share of input V in firm i's total revenue. To address measurement error in observed output, I correct the expenditure share as:

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

where $\tilde{Q}_{it}$ is observed output, and $\hat{\epsilon}_{it}$ represents the estimated measurement error or unanticipated shocks. Finally, I use block bootstrapping to estimate standard errors, ensuring that the error structure reflects autocorrelation across firms and time.

Table 5: Unweighted Markup Distribution 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 6: Unweighted Markup Distribution 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 7: Unweighted Markup Distribution 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

Table 8: **Time Series 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.

Table 9: **Cross Sectional 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.

B Selection on Observables

I follow Imbens and Xu (2024) and adopt the potential outcome framework introduced by Rubin (1974). For each firm i, where $i = 1, \ldots, N$, two potential outcomes exist: $Y_i(0)$ represents the

Table 10: **Cross Sectional 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

outcome (log markups) had firm i not participated in public procurement, and $Y_i(1)$ represents the outcome had it participated in public procurement. The difference between these two potential outcomes,

$$\tau_i \equiv Y_i(1) - Y_i(0),$$

is the causal effect of 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 augmented the covariate vector to include lagged markups. 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), defined as:

$$ATE \equiv \frac{1}{N} \sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\}.$$

I cannot directly estimate the ATT because I do not observe the control outcomes for the treated units. Therefore, I define my estimand as the covariate-adjusted difference between treated and control groups:

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

This quantity can be consistently estimated. However, it is equal to the ATT only under two key assumptions: unconfoundedness and overlap (see, for example, Abadie and Cattaneo, 2018).

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 requires only a weaker version of unconfoundedness:

$$W_i \perp\!\!\!\perp Y_i(0) \mid X_i$$

. This assumption contrasts with traditional econometric assumptions of exogeneity, which are articulated in terms of residuals defined by functional forms. Unconfoundedness separates the functional form component of the assumptions from their essence. Essentially, it is sufficient for researchers to understand the treatment assignment mechanism without full knowledge of the data-generating process for the potential outcomes. A key result in Rosenbaum and Rubin (1983) shows that Assumption 1 implies:

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

where $e(X_i) \equiv \Pr(W_i = 1 \mid X_i)$ is the propensity score for unit i. This result is important because it reduces the dimensionality 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 (see, for example, Imbens, 2004). While I acknowledge concerns about its validity in the absence of a clear understanding of the treatment assignment mechanism, supplementary analyses such as placebo tests and sensitivity analyses can help assess the plausibility of this assumption and improve the credibility of results based on it. I will illustrate these methods in the next section. In the context of my data, it is evident that appropriate pretreatment variables should be controlled (see, for example, Cinelli et al., 2022).

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

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

If the ATT is of interest, only a weaker overlap assumption is required: $\Pr(W_i = 1 \mid X_i) < 1$. Overlap is crucial for identifying the ATT when I am unwilling to make functional form assumptions about the conditional means of the potential outcomes. When X_i only a few covariates, inspecting pairs of 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, inspecting the distribution of the propensity scores by treatment status, estimated using a flexible method, is a more practical approach.

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 allow for interpolation or extrapolation of treatment effects across all covariate levels and combinations, formally eliminating the need for overlap. Ensuring overlap or improving balance typically involves dropping some units from the full sample. While this reduces information, the improvement in robustness and reduction of bias often outweigh the loss in precision. An approach well-suited to settings focused on the ATT is to create a matched sample where each treated unit is matched to a control unit based on the estimated propensity score. Beyond ensuring overlap, this method produces a sample with much better balance in the covariate distributions. Trimming to ensure overlap is likely more important than the choice of specific estimation strategies.

Estimation Given Unconfoundedness and Overlap Denote the conditional means of the two potential outcomes as:

$$\mu_w(x) \equiv \mathbb{E}[Y_i(w) \mid X_i = x], \quad \text{for } w \in \{0, 1\}.$$

The most commonly used method is a simple linear regression using the treatment indicator and covariates as regressors. The regression method models the conditional means of potential outcomes parametrically. Specifically, $\mu_0(x) = x^\top \beta$ assumes a linear relationship between covariates and outcomes, 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 approach, also known as the Oaxaca-Blinder decomposition (Kline, 2011), allows for flexibility in modeling the outcome semiparametrically or nonparametrically (Heckman et al., 1998; Athey et al., 2019). Methods that directly adjust for covariate imbalance between the treatment and control groups include blocking on covariates, covariate matching, and weighting approaches 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 instance, one of the inverse propensity weighting (IPW) estimators for the ATT is the Hájek estimator:

$$\hat{\tau}_{IPW} = \frac{1}{N_{tr}} \sum_{i:W_i=1} Y_i - \sum_{i:W_i=0} \hat{\omega}_i Y_i,$$

with weights:

$$\hat{\omega}_i = \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} \Big/ \sum_{j:W_j=0} \frac{\hat{e}(X_j)}{1-\hat{e}(X_j)}.$$

Here, $\hat{e}(x)$ is the estimated propensity score.

The augmented inverse propensity weighting (AIPW) estimator, which combines weighting and regression, can be written as:

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

where $\hat{\omega}_i$ are the IPW (or balancing) weights. The AIPW estimator can be viewed as combining an outcome model with an adjustment term, consisting of an IPW estimator applied to the residuals from the outcome model. Recent machine learning estimators (Chernozhukov et al., 2018) adopt the form of an AIPW estimator. They achieve stability in the moment conditions used to identify the causal parameter using Neyman orthogonal score functions that are robust to small perturbations in nuisance functions: the conditional mean $\mu_w(x)$ and the propensity score $e(x)$.

Alternative Estimands I study heterogeneous treatment effects by estimating the conditional average treatment effect for the treated (CATT):

$$\tau(x) \equiv \frac{1}{N_x} \sum_{i:X_i=x, W_i=1} \tau_i,$$

where N_x is the number of treated units whose covariate values equal x. Quantile treatment effects are defined as the differences between the quantiles of the treated and untreated potential outcome distributions for the population or treated group. Because Assumptions 1 and 2 allow for the identification of the full marginal distributions of $Y_i(0)$ and $Y_i(1)$, quantile treatment effects are identified under these assumptions. The range of the CATT or quantile treatment effects estimates can also provide evidence regarding the plausibility of the unconfoundedness assumption.

Placebo Analyses I assess unconfoundedness indirectly using a conditional independence restriction that is formally testable. This restriction differs from unconfoundedness in two ways. First, it conditions on a subset of the covariates used 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. Specifically, I use a placebo test that estimates the treatment effect on a pretreatment variable, known to be unaffected by the treatment. A lagged outcome is particularly appealing as a pseudo-outcome because it is typically a good proxy for the target outcome. Imbens (2015) discusses testing additional implications of the conditional independence relationship.

Sensitivity Analyses By assuming unconfoundedness holds only conditional on observed covariates X_i and an unobserved confounder U_i, a causal relationship can be considered insensitive to unobserved confounding if the estimated effect remains robust to strong associations with U_i.

I follow Imbens (2003), who benchmarks the associations between the unobserved U_i and the outcome and treatment, adjusting for observed covariates. Contour plots introduced by Imbens (2003) help interpret the sensitivity of results to violations of the unconfoundedness assumption.

C Selection on Unobservables

I present a summary of Arkhangelsky and Imbens (2024), who survey recent advancements in the causal panel data literature. This body of work has extended earlier methods on difference-in-differences and two-way fixed-effect estimators, incorporating factor models and interactive fixed effects. Additionally, it has introduced novel methods that leverage synthetic control approaches.

The Econometrics Panel Data Literature The earlier econometric panel data literature focused heavily on the dynamics of the outcome process. It distinguished between state dependence and unobserved heterogeneity (Chamberlain, 1984) and explored various dynamic forms of exogeneity (Engle et al., 1983). A key area of study in this literature involved models that combined unit-fixed effects with lagged dependent variables. These models raised concerns about biases in least squares estimators when applied to short panels (Nickell, 1981). To address these biases, instrumental variable approaches were proposed (Arellano and Bond, 1991; Blundell and Bond, 1998).

In contrast to this earlier focus, a prominent theme in the current literature is the presence of general heterogeneity in causal effects. This heterogeneity manifests both over time and across units and is often associated with observed and unobserved characteristics. Recognizing the importance of heterogeneity has revealed that many previously popular estimators, including traditional difference-in-differences and two-way fixed-effects estimators, are sensitive to its presence. This sensitivity has spurred the development of more robust alternatives.

For instance, recent methods incorporate interactive fixed effects or factor models to account for unobserved common shocks that vary over time and across units. These approaches better accommodate heterogeneous treatment effects, which traditional methods often fail to address adequately. Additionally, the introduction of synthetic control methods has allowed researchers to construct counterfactual outcomes in cases with limited treated units, offering a more flexible framework for addressing selection on unobservables.

Data Setup 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 indexed by unit and time.

Collect the outcomes and treatment assignments into two $N \times T$ matrices:

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

Ideally, the focus is on a balanced panel, where for all units $i = 1, \ldots, N$, outcomes are observed for all $t = 1, \ldots, T$. In practice, however, panels are often unbalanced due to incomplete observations—either because units are observed for different lengths of time or because data is missing for some periods. In the most general case, treatment may vary both across units and over time,

with units switching in and out of the treatment group:

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

With this type of data, causal effects can be identified using both within-unit and within-time variation in treatment. In particular, the presence of both types of variation allows for richer identification strategies. However, in settings with dynamic effects, assuming their absence without proper justification can lead to results that are difficult to interpret.

A key focus of the recent difference-in-differences (DID) and two-way fixed-effects (TWFE) literature has been on the staggered adoption case, where units remain in the treatment group once they adopt the treatment. In this setting, the time of treatment adoption varies across units—some adopt earlier, while others adopt later:

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

This scenario is also referred to as the absorbing treatment setting, in which a unit remains treated once it adopts the treatment. It provides richer information about the potential presence of dynamic effects. Under certain assumptions, it enables researchers to distinguish dynamic effects (e.g., lagged or cumulative treatment effects) from heterogeneity across calendar time. By leveraging this structure, researchers can identify treatment effects more robustly than in general cases.

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

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

and let $\mathbf{w}_i$ denote the vector of treatment values for unit i. Similarly, let $\mathbf{w}^t$ denote the t-component column vector of treatment assignments up to time t:

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

so that $\mathbf{w}^T = \mathbf{w}$, and similarly for $\mathbf{w}_i^t$. The potential outcomes for unit i in period t are indexed by the full T-component vector of assignments $\mathbf{w}$:

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

This notation implicitly assumes the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1978; Imbens and Rubin, 2015), which states that there are no spillovers or interference between units. While SUTVA is a strong assumption and may be violated in some applications, much of the recent causal panel data literature does not explicitly account for interference.

Without further restrictions, this setup describes 2^T potential outcomes for each unit and each time period as a function of the T-component vector of treatment assignments. Unit-level treatment effects for any pair of assignment vectors $\mathbf{w}$ and $\mathbf{w}'$ are defined as:

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

with the corresponding population average effect:

$$\tau_t^{\mathbf{w},\mathbf{w}'} \equiv \mathbb{E}\left[Y_{it}(\mathbf{w}') - Y_{it}(\mathbf{w})\right].$$

These unit-level and average causal effects form the basic building blocks of many estimands considered in the literature. A key challenge is the number of possible treatment effects: there are $2^{T-1} \times (2^T - 1)$ distinct average causal effects of the form $\tau_t^{\mathbf{w},\mathbf{w}'}$. With $T = 2$, there are six distinct average causal effects, and this number grows rapidly in T. In practice, this necessitates focusing on summary measures of these effects, such as averages over time or specific subpopulations.

Assumption 3 (No Anticipation)**.** *The potential outcomes satisfy:*

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

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

Under this assumption, outcomes for period t depend only on the treatment path up to and including period t, not on future treatments. This reduces the number of potential treatment effects from $2^{T-1} \times (2^T - 1)$ to $(\sum_{t=1}^{T} 2^{t-1})(\sum_{t=1}^{T} 2^t - 1)$. In observational studies, the no-anticipation assumption may not hold. To address this, one approach is to allow for limited anticipation, where treatments are assumed to be anticipated for a fixed number of periods. Algorithmically, this corresponds to redefining $\mathbf{w}$ by shifting it forward by the fixed number of periods.

The reduced structure allows for distinguishing between static and dynamic treatment effects. Static treatment effects measure the response of the current outcome to the current treatment while holding past treatments fixed:

$$\tau_{it}^{(\mathbf{w}^{t-1},0),(\mathbf{w}^{t-1},1)}.$$

Dynamic treatment effects capture the effect of past treatments on the current outcome while holding the current treatment fixed:

$$\tau_{it}^{(\mathbf{w}^{t-1},w^t),(\mathbf{w}'^{t-1},w^t)}.$$

A stronger assumption is the absence of any dynamic or carryover effects.

Assumption 4 (No Dynamic / Carryover Effects)**.** *The potential outcomes satisfy:*

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

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

Under this assumption, outcomes depend only on the contemporaneous treatment assignment. This reduces the total number of potential treatment effects for each unit to T (one per period). The treatment effects simplify to:

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

where τ_{it} has no superscripts because there are only two possible treatment states $w \in \{0, 1\}$.

An alternative assumption is staggered adoption, where treatment is absorbing, meaning once a unit adopts treatment, it remains treated.

Assumption 5. (STAGGERED ADOPTION)

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

Define the adoption date A_i as the period when unit i first receives treatment:

$$A_i \equiv T + 1 - \sum_{t=1}^{T} W_{it},$$

for units that are treated within the sample, and $A_i \equiv \infty$ for units that are never treated. Under staggered adoption, the potential outcomes can be written in terms of the adoption date a, with some abuse of notation:

$$Y_{it}(a), \quad \text{for } a = 1, \ldots, T, \infty.$$

The realized outcome for each unit is $Y_{it} = Y_{it}(A_i)$.

Without imposing the no-anticipation or no-dynamics assumptions, the building blocks are:

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

with the corresponding population average:

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

For subpopulations defined by adoption date a'', one can define the conditional average treatment effect:

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

This estimand is conceptually similar to the average treatment effect on the treated in cross-sectional settings, but here selection operates over both units and time. As in the cross-sectional setting, this matters for interpretation in observational studies, in which the researcher does not control the assignment process. In settings with variation in the adoption date there are many such average effects, depending on when the units adopted, and which period one measures the effect.

The Two-Way Fixed Effects Characterization Consider a panel data setting with no anticipation, no dynamics, and the parallel trends assumption:

Assumption 6 (Two-Way Fixed Effects Model)**.** *The control outcome $Y_{it}(0)$ satisfies:*

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

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

Assumption 7 (Parallel Trends Assumption)**.** *The potential outcomes satisfy:*

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

Combining these two assumptions gives the following model for the realized outcome:

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

This model can be estimated via least squares:

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

To avoid perfect collinearity of the regressors, one must impose a normalization (e.g., fixing one α_i or one β_t to zero). However, this normalization does not affect the value of the treatment effect estimator $\hat{\tau}^{\text{TWFE}}$.

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 address the issue of negative weights, researchers have recently proposed several modifications to the TWFE estimator. These methods maintain the TWFE assumption for control outcomes while avoiding additional assumptions about treatment effect heterogeneity.

Callaway and Sant'Anna (2020) propose two methods to address negative weights. The first method focuses on a group with an adoption date a. It compares $\overline{Y}_{t|a}$, the group's average outcomes in any post-adoption period $t \geq a$, with $\overline{Y}_{a-1|a}$, the group's average outcomes immediately before adoption. The difference in outcomes is then adjusted by subtracting the difference for the same time periods in the never-treated group. As an alternative, Callaway and Sant'Anna (2020) suggest using the average of groups that adopt treatment after period t as the control group.

Sun and Abraham (2020) recommend reporting the unweighted average of the same components as Callaway and Sant'Anna (2020) across all post-treatment periods t. Within a period, weights are assigned based on the fraction of units with adoption dates prior to t, excluding first-period adopters. An additional issue highlighted by Sun and Abraham (2020) is related to validating the two-way model. They show that common implementations of parallel-trend tests using pre-treatment data often involve comparisons with negative weights. Consequently, they caution against such procedures.

de Chaisemartin and d'Haultfœuille (2020) tackle negative weights by focusing on one-period-ahead double differences, using later adopters ($a > t$) as control groups. These differences are aggregated by averaging across all later-adopting groups and time periods, weighted by the fraction of adopters in each period. A challenge with the de Chaisemartin and d'Haultfœuille (2020) approach is the potential for larger standard errors due to fewer comparisons. While the additivity assumption may hold better over shorter horizons, this method is more sensitive to dynamic effects.

Borusyak et al. (2021) propose a model for baseline outcomes that generalizes the TWFE framework:

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

where A_{it} and X_{it} are observed covariates, resulting in a factor-type structure. The TWFE model is a special case of this framework, with $A_{it} \equiv 1$ and $X_{it} \equiv (\mathbf{1}_{t=1}, \ldots, \mathbf{1}_{t=T})$. They estimate λ_i and δ using least squares on control units only, then construct unit- and period-specific imputations for unobserved control outcomes in treated units. This leads to unit- and period-specific treatment effect estimates:

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

Robust Confidence Set and Sensitivity Analysis Rambachan and Roth (2023) evaluate the robustness of significant results under varying degrees 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 captures the differences in trends between treated and comparison groups that would have occurred in the absence of treatment.

Consider the restriction that posttreatment violations of parallel trends between consecutive periods are no greater than $\bar{M}$ times the maximum difference in trends between consecutive placebo periods:

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

The ATT of interest, denoted θ, can be expressed as a weighted average of post-treatment trends: $\theta = l'_{att}\delta_{\text{post}}$, where

$$l_{att} = \left[\frac{n_1}{\sum_{t=1}^{T_{\text{post}}} n_t}, \cdots, \frac{n_{T_{\text{post}}}}{\sum_{t=1}^{T_{\text{post}}} n_t} \right]',$$

and n_t represents the number of observations in the posttreatment period t.

The set of values for θ consistent with a given β under the restriction $\delta \in \Delta^{RM}(\bar{M})$ is defined 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, $\hat{\beta} \sim \mathcal{N}(\tau+\delta, \Sigma_n)$, Rambachan and Roth (2023) define confidence sets for θ as: $\mathcal{C}_n\left(\hat{\beta}_n, \Sigma_n\right)$, satisfying:

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

To evaluate the robustness of the significant ATT estimate, one can determine the threshold value of $\bar{M}$ such that the confidence interval $\mathcal{C}_n(\hat{\beta}_n, \hat{\Sigma}_n)$ just includes 0.

Relaxing the Two-Way Fixed Effect Structure A key strand of recent causal panel data literature originates from Abadie and Gardeazabal (2003), who introduced the Synthetic Control (SC) method. This approach focuses on imputing missing potential outcomes by constructing synthetic versions of treated units, which are represented as convex combinations of control units. Unlike model-based approaches, the SC method relies on algorithmic procedures, and the introduction of this aspect has inspired extensive new research.

One prominent set of methods focuses directly on factor models, where the control outcome is assumed to take the form:

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

This specification generalizes the TWFE structure. By setting the rank to $R = 2$, $\alpha_{i2} = 1$ for all i, and $\beta_{t1} = 1$ for all t, the factor model reduces to the standard TWFE setup, where outcomes depend on additive unit and time fixed effects.Athey et al. (2021) propose modeling the entire matrix of potential control outcomes as:

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

where ε_{it} is random noise, uncorrelated with the other components. The matrix $\mathbf{L}$, with elements L_{it}, is a low-rank matrix. While the unit and time fixed effects (α_i and β_t) could theoretically be subsumed into $\mathbf{L}$, keeping them separate improves the estimator's performance in practice. Athey et al. (2021) propose the Nuclear-Norm Matrix-Completion (NNMC) estimator, which minimizes the following objective function:

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

over $\mathbf{L}$, α, and β. The missing values $Y_{it}(0)$ are then imputed using the estimated parameters. In this framework, the nuclear norm $\|\mathbf{L}\|_*$, defined as the sum of the singular values $\sigma_i(\mathbf{L})$ of the matrix $\mathbf{L}$, serves as a regularization technique to encourage low-rank solutions. The singular values are derived from the singular value decomposition (SVD) of $\mathbf{L}$, where $\mathbf{L} = \mathbf{S}\Sigma\mathbf{R}$. Here, $\mathbf{S}$ and $\mathbf{R}$ are matrices of left and right singular vectors, and Σ is the diagonal matrix of singular values. The penalty parameter λ is selected via out-of-sample cross-validation, ensuring optimal shrinkage of $\mathbf{L}$ toward a low-rank solution, analogous to how LASSO induces sparsity in linear regression.

Xu (2017) offer a related but distinct approach, focusing on the direct estimation of factor models as an alternative to synthetic control methods. In this approach, the number of factors is either pre-specified or estimated from the data, and the model is directly estimated after appropriate normalization. This work builds on the econometric literature on factor models (e.g., Bai, 2009). Using the factor model, researchers can impute missing potential outcomes for treated unit-time pairs and thereby estimate the average treatment effect for the treated (ATT).

Imputation Estimators Framework This section presents the framework of Liu et al. (2022). Denote $\mathbf{U}_{it}$ as the unobservable attributes and ε_{it} as the idiosyncratic error term.

Assumption 1 (*Functional form*) $Y_{it}(0) = h(\mathbf{U}_{it}) + \varepsilon_{it}$, where $h(\cdot)$ is a known parametric function.

Assumption 2 (*Strict exogeneity*) $\varepsilon_{it} \perp\!\!\!\perp \{W_{js}, \mathbf{U}_{js}\}, \forall i, j \in \{1, 2, \ldots, N\}, \forall s, t \in \{1, 2, \ldots, T\}$.

Assumption 3 (*Low-dimensional decomposition*) There exists a low-dimensional decomposition of $h(\mathbf{U}_{it})$: $h(\mathbf{U}_{it}) = L_{it}$, where $\text{rank}(\mathbf{L}_{N\times T}) \ll \min\{N, T\}$. For example, $\mathbf{L} = \mathbf{\Lambda F}$, where $\mathbf{\Lambda}$ is an $(N \times r)$ matrix of factor loadings, $\mathbf{F}$ is an $(r \times T)$ matrix of factors, and $r \ll \min\{N, T\}$.

This class of models is scale-dependent (Athey and Imbens, 2006), and their ability to control for unobserved confounders relies on the functional form assumption. The setup excludes temporal and spatial interference, including anticipation effects and carryover effects. Borusyak et al. (2021) demonstrate that anticipation effects can lead to underidentification of causal effects, and the same logic applies to carryover effects.

Together with Assumption 1, Assumption 2 implies quasi-randomization conditional on $\mathbf{U}$, such that:

$$Y_{is}(0) \perp\!\!\!\perp W_{it} \mid \mathbf{U}_{i,1\ldots T}, \quad \forall i, s, t,$$

where $\mathbf{U}_{i,1\ldots T}$ denotes the time series of $\mathbf{U}_{it}$. Assumption 2 further implies:

$$\mathbb{E}[Y_{it}(0) \mid \mathbf{U}_{it}] - \mathbb{E}[Y_{is}(0) \mid \mathbf{U}_{is}] = \mathbb{E}[Y_{jt}(0) \mid \mathbf{U}_{jt}] - \mathbb{E}[Y_{js}(0) \mid \mathbf{U}_{js}], \quad \forall i, j, \forall t, s,$$

which states that, conditional on the unobserved attributes, the average change in untreated potential outcomes between periods s and t is the same across units i and j.

Assumption 3 enables one to condition on $\mathbf{U}_{it}$. For example, if $\mathbf{U}_{it} = f_t \cdot \lambda_i$ is one-dimensional, it can be interpreted as the effect of a common time trend f_t that varies heterogeneously across units, where the heterogeneity is captured by λ_i. When f_t is constant, $\mathbf{U}_{it}$ reduces to unit fixed effects; when λ_i is constant, it reduces to time fixed effects. In the presence of unobserved confounders $\mathbf{U}_{it}$, treatment assignment depends on observed untreated outcomes. Assumption 3 allows researchers to approximate $\mathbf{U}_{it}$ using data, breaking this dependency and making the model feasible. The estimators proposed by Liu et al. (2022) and Athey et al. (2021) rely on Assumptions 1–3.

Difference between IFEct and MC Regularizing the singular values of the residual matrix $\mathbf{L}$ is a key step in imputation approaches. The IFEct (Interactive Fixed Effects) method uses a "best subset" approach, selecting the r largest singular values, where r is pre-specified such that $r < \min\{N, T\}$. In contrast, the MC (Matrix Completion) method imposes an L_1 penalty on all singular values, controlled by a tuning parameter λ_L.

$$\begin{pmatrix} \sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & \sigma_T \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}_{N\times T} \quad \begin{pmatrix} |\sigma_1 - \lambda_L|_+ & 0 & 0 & \cdots & 0 \\ 0 & |\sigma_2 - \lambda_L|_+ & 0 & \cdots & 0 \\ 0 & 0 & |\sigma_3 - \lambda_L|_+ & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ & & & & |\sigma_T - \lambda_L|_+ \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}_{N\times T}$$

Note: The figure above from Liu et al. (2022) illustrates how regularization works under IFEct and MC.

Placebo Test To assess the validity of the model, assume the treatment starts S periods earlier than its actual onset for each unit in the treatment group. Apply the same counterfactual estimator to obtain estimates of ATT_s for $s = -(S-1), \ldots, -1, 0$. Then estimate the overall ATT for the S pre-treatment periods. If this "ATT" estimate is statistically different from zero, it suggests that some or all of the identifying assumptions are invalid. For instance, a feedback effect (where past outcomes, such as Y_{t-1}, influence current treatment assignment, D_t) would indicate a violation of the strict exogeneity assumption, which can often be detected with sufficient data.

Since a placebo test is effectively a test for equivalence, a simple difference-in-means approach may lack statistical power. Specifically, when the number of observations is small, failing to reject the null hypothesis of zero placebo effects does not necessarily imply that equivalence holds. Instead, researchers can employ an equivalence test with reversed null hypotheses:

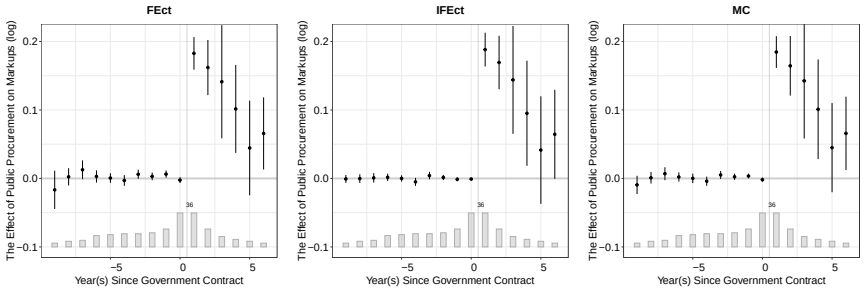
$$ATT^p < -\theta_2 \quad \text{or} \quad ATT^p > \theta_1,$$

where $-\theta_2 < 0 < \theta_1$ are pre-specified equivalence thresholds. Rejection of the null hypothesis implies $-\theta_2 \leq ATT^p \leq \theta_1$, providing evidence in favor of the identifying assumptions. The Two One-Sided Tests (TOST) procedure is used to check the equivalence of ATT^p to zero.

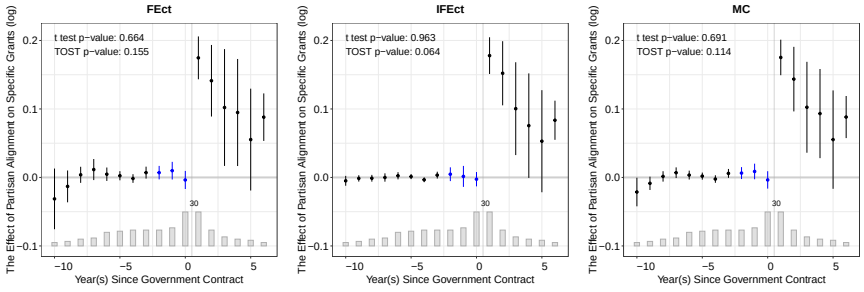
I follow Liu et al. (2022) and set $\theta_1 = \theta_2 = 0.36\hat{\sigma}_\varepsilon$, where $\hat{\sigma}_\varepsilon$ is the standard deviation of the residualized untreated outcomes.

A Test for No Carryover Effects Using the same framework, predict $Y_{it}(0)$ for periods immediately following the end of treatment. If carryover effects are absent, the average prediction error in these periods should be close to zero. Both t-tests and equivalence (TOST) tests can be applied.

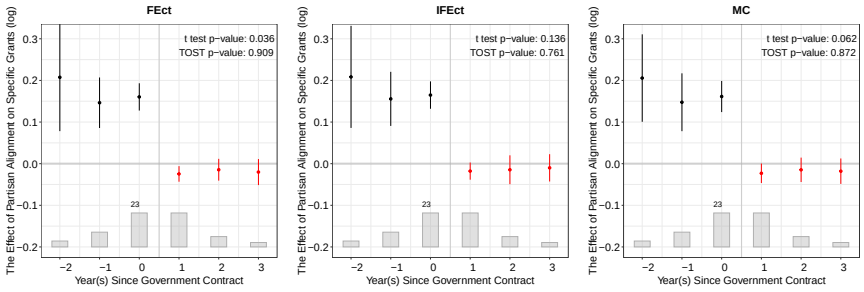
Figure 16: Balanced Panel Non-absorbing Treatment Effect of Public Procurement on Markups



(a) Dynamic Treatment Effects



(b) Placebo Test



(c) Test for Carryover Effects

Hybrid Methods Recent hybrid methods combine the benefits of the Synthetic Control (SC) approach with features of two-way fixed effects (TWFE) or unconfoundedness-based methods. These methods are particularly appealing because they nest TWFE within more flexible outcome models. One can generalize the outcome model or use a localized version of the TWFE model. The SC estimator for the treatment effect can be expressed as a weighted least squares regression:

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

By contrast, the TWFE estimator relies on a slightly different least squares regression:

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

There are two key differences between these methods: 1. The SC regression incorporates weights $\hat{\omega}_i$, while the TWFE regression does not. 2. The TWFE regression includes unit-specific fixed effects α_i, which are omitted in the SC regression. Arkhangelsky et al. (2021) address the absence of unit fixed effects in the SC regression by proposing the Synthetic Difference-in-Differences (SDID) estimator. This method integrates unit fixed effects α_i and SC weights $\hat{\omega}_i$, along with analogous time weights $\hat{\lambda}_t$. The SDID estimator minimizes:

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

The time weights $\hat{\lambda}_t$ are calculated similarly to the unit weights. Specifically, they are chosen to minimize the weighted squared prediction error for the untreated units in the post-treatment period: $\min_{\lambda} \sum_{i=1}^{N-1} \left(Y_{iT} - \sum_{s=1}^{T-1} \lambda_s Y_{is}\right)^2$, subject to the constraints $\lambda_s \geq 0$ and $\sum_{s=1}^{T-1} \lambda_s = 1$.

Ben-Michael et al. (2021) propose augmenting the SC estimator by regressing outcomes in the treatment period on lagged outcomes using data from control units. For simplicity, consider a setup where unit N and period T represent the only treated unit/time-period pair. Ridge regression is used in this first step:

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

where the ridge penalty parameter λ is selected via cross-validation. Under a standard unconfoundedness approach, the potential control outcome for the treated unit/time-period pair is predicted as:

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

The augmented SC estimator modifies this prediction by incorporating SC weights, thereby correcting bias in either the unconfoundedness estimator or the SC estimator. Specifically:

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

Equivalently, this can be expressed as: $\hat{Y}_{NT} = \sum_{i=1}^{N-1} \omega_i Y_{iT} + \sum_{s=1}^{T-1} \hat{\eta}_s \left(Y_{Ns} - \sum_{j=1}^{N-1} \omega_j Y_{js} \right)$. This formulation highlights the dual role of SC weights and regression adjustments in improving predictive accuracy. Ben-Michael et al. (2022) further extend this approach to accommodate settings with staggered treatment adoption, broadening its applicability to more realistic scenarios.

Implementing Staggered Synthetic Difference-in-Differences This section presents the estimation procedure for event-study Synthetic Difference-in-Differences estimators, as outlined by Ciccia (2024). Consider a setting with N units observed over T periods, where $N_{tr} < N$ units receive treatment D starting in period a, with $1 < a \leq T$. The treatment D is binary ($D \in \{0, 1\}$) and affects an outcome of interest Y. Both the outcome and treatment are observed for all (i, t) cells, ensuring a balanced panel structure. The cohort-specific SDID estimator is expressed as:

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

where: - λ_t and ω_i are optimal weights chosen to approximate the pre-treatment outcome evolution of treated and synthetic control units, - N_{tr}^a is the number of treated units in cohort a, - T_{tr}^a is the number of post-treatment periods for cohort a, - I^a is the set of treated units in cohort a, and - N_{co} is the number of control units. The weights ω_i are related to those in Abadie et al. (2010)'s synthetic control method, but with two key differences.

First, Arkhangelsky et al. (2021) include an intercept term ω_0, allowing weights to relax the need to perfectly match pre-treatment trends, instead ensuring parallel trends. This flexibility is possible because unit fixed effects α_i absorb constant differences across units.

Second, following Doudchenko and Imbens (2016), a regularization penalty is applied to ω_i to increase weight dispersion and ensure uniqueness. Importantly, the unit weights include regularization, but the time weights do not. This distinction accounts for the correlated observations within units over time but not across units within the same time period, beyond what is captured by the latent factor model.

The treatment effect ℓ periods after treatment adoption ($\ell \in \{1, \ldots, T^a_{post}\}$) can be estimated by disaggregating τ^{sdid}_a into 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 generalizes those proposed by Borusyak et al. (2021), Liu et al. (2022), Gardner (2022), and Dube et al. (2023) for canonical Difference-in-Differences (de Chaisemartin and D'Haultfoeuille, 2023) designs, but it introduces unit-time-specific weights.

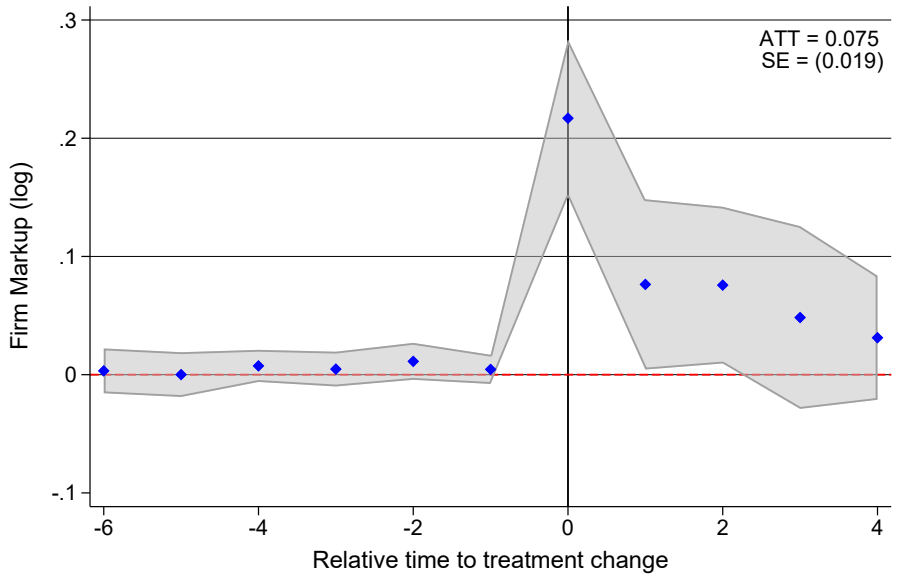
Let A_ℓ be the subset of cohorts in A where $a - 1 + \ell \leq T$, meaning their ℓ-th dynamic effect is defined. Define $N^\ell_{tr} = \sum_{a \in A_\ell} N^a_{tr}$ as the total number of treated units in cohorts where the ℓ-th dynamic effect can be estimated. The weighted average of cohort-specific treatment effects ℓ periods after treatment is:

$$\hat{\tau}^{sdid}_\ell = \sum_{a \in A_\ell} \frac{N^a_{tr}}{N^\ell_{tr}} \hat{\tau}^{sdid}_{a,\ell}.$$

This estimator aggregates cohort-specific effects, upweighting cohorts with more treated units.

Finally, let $T_{post} = \sum_{a \in A} N^a_{tr} T^a_{tr}$ denote the total number of post-treatment periods across all cohorts, and let $T_{tr} = \max_{a \in A} T^a_{tr}$ denote the maximum number of post-treatment periods. The overall average treatment effect on the treated (ATT) is: $\widehat{ATT}_{sdid} = \frac{1}{T_{post}} \sum_{\ell=1}^{T_{tr}} N^\ell_{tr} \hat{\tau}^{sdid}_\ell$.

Figure 17: $\hat{\tau}^{sdid}_\ell$ and $\widehat{ATT}_{sdid}$ for cohorts 2011-2018 in the balanced panel



Algorithm 1: SDID Estimation with Staggered Adoption (Clarke et al., 2023)

Data: $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{A}$.

Result: Adoption-specific values $\hat{\tau}_a^{sdid}$, $\hat{\omega}_a^{sdid}$, and $\hat{\lambda}_a^{sdid}$ for all $a \in \mathbf{A}$.

for $a \in \mathbf{A}$ **do**

1. Subset $\mathbf{Y}$ and $\mathbf{W}$ to units that are pure controls or first adopt treatment in period $t = a$. ;
2. Compute the regularization parameter ζ. ;
3. Compute unit weights $\hat{\omega}_a^{sdid}$. ;
4. Compute time weights $\hat{\lambda}_a^{sdid}$. ;
5. Compute the SDID estimator via weighted DID regression:

$$\left(\hat{\tau}_a^{sdid}, \hat{\mu}_a, \hat{\alpha}_a, \hat{\beta}_a\right) =_{\tau,\mu,\alpha,\beta} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\right)^2 \widehat{\omega}_{a,i}^{sdid} \hat{\lambda}_{a,t}^{sdid} \right\}.$$

end

Algorithm 2: SDID Bootstrap Inference with Staggered Adoption (Clarke et al., 2023)

Data: $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{A}$, B.

Result: Variance for each adoption-specific estimate $\widehat{V}_{\tau_a}^{cb}$ for all $a \in \mathbf{A}$.

for $b \leftarrow 1$ *to* B **do**

1. Construct a bootstrap dataset $(\mathbf{Y}^{(b)}, \mathbf{W}^{(b)}, \mathbf{A}^{(b)})$ by sampling N rows of $(\mathbf{Y}, \mathbf{W})$ with replacement, and generating $\mathbf{A}$ as the corresponding adoption vector. ;
2. **if** *the bootstrap sample has no treated or control units* **then**
   Discard the resample and return to Step 1. ;

   end
3. Generate a vector of adoption-date-specific resampled SDID estimates $\tau_a^{(b)}$ for all $a \in \mathbf{A}^{(b)}$ using Algorithm 1. ;

end

4. Define:

$$\widehat{V}_{\tau_a}^{cb} = \frac{1}{B} \sum_{b=1}^{B} \left(\tau_a^{(b)} - \frac{1}{B} \sum_{b=1}^{B} \tau_a^{(b)} \right)^2.$$