A Tale of Two Panel Data Regressions

By Dennis Shen\textsuperscript{1}, Peng Ding\textsuperscript{1}, Jasjeet Sekhon\textsuperscript{2}, and Bin Yu\textsuperscript{1}

\textsuperscript{1}University of California, Berkeley
\textsuperscript{2}Yale University

Abstract

A central goal in social science is to evaluate the causal effect of a policy. In this pursuit, researchers often organize their observations in a panel data format, where a subset of units are exposed to a policy (treatment) for some time periods while the remaining units are unaffected (control). The spread of information across time and space motivates two general approaches to estimate and infer causal effects: (i) unconfoundedness, which exploits time series patterns, and (ii) synthetic controls, which exploits cross-sectional patterns. Although conventional wisdom decrees that the two approaches are fundamentally different, we show that they yield numerically identical estimates under several popular settings that we coin the symmetric class. We study the two approaches for said class under a generalized regression framework and argue that valid inference relies on both correlation patterns. Accordingly, we construct a mixed confidence interval that captures the uncertainty across both time and space. We illustrate its advantages over inference procedures that only account for one dimension using data-inspired simulations and empirical applications. Building on these insights, we advocate for panel data agnostic (PANDA) regression—rooted in model checking and based on symmetric estimators and mixed confidence intervals—when the data generating process is unknown.

Keywords—synthetic controls, unconfoundedness, causal inference, principal component regression, minimum norm estimators

\textsuperscript{*}email: dshen24@berkeley.edu. The data and code are available at https://github.com/deshen24/panel-data-regressions. We sincerely thank Avi Feller for his thoughtful comments and insightful feedback. We also gratefully acknowledge partial support from NSF grants 1945136, 1953191, 2022448, 2023505 on Collaborative Research: Foundations of Data Science Institute (FODSI), and ONR grant N00014-17-1-2176.
1 Introduction

In a seminal paper, Abadie and Gardeazabal (2003) set out to investigate the economic impact of terrorism in Basque Country. Prior to the outset of terrorist activity in the early 1970’s, Basque Country was considered to be one of the wealthiest regions in Spain. After thirty years of turmoil, however, its economic activity dropped substantially relative to its neighboring regions. Although intuition affirms that Basque Country’s economic downturn can be attributed, at least partially, to its political and civil unrest, it is difficult to quantitatively isolate the economic costs of conflict. In response to this challenge, Abadie and Gardeazabal (2003) introduced the synthetic controls framework. At its core, the synthetic controls method constructs a synthetic Basque Country from a weighted composition of control regions that are largely unaffected by the instability to estimate Basque Country’s economic evolution in the absence of terrorism. This novel concept has inspired an entire subliterature within econometrics that is “arguably the most important innovation in the policy evaluation literature in the last 15 years” (Athey and Imbens, 2017).

Historically, researchers have often tackled problems of this flavor using repeated observations of units across time, i.e., panel data, where a subset of units are exposed to a treatment of interest during some time periods while the other units remain unaffected. In the study above, the per capita gross domestic product (GDP) of 17 Spanish regions are measured from 1955–1998. Basque Country is the sole treated unit and the remaining regions are the control units; the pre- and post-treatment periods are defined as the time horizons before and after the first wave of terrorist activity, respectively.

In recent years and across many fields, synthetic controls has emerged as the de facto approach in panel studies. Prior to its introduction, the unconfoundedness approach (Rosenbaum and Rubin, 1983; Rubin, 2006; Imbens and Wooldridge, 2009; Wooldridge, 2010; Imbens and Rubin, 2015) served as a common workhorse. In contrast to synthetic controls, which posits a relation between treated and control units that is stable over time, unconfoundedness posits a relation between treated and pretreatment periods that is stable across units. Therefore, while synthetic controls exploit cross-sectional (unit-level) correlation patterns, unconfoundedness exploits time series correlation patterns. Given their conceptual and methodological distinctions, the two approaches are widely considered to be fundamentally different (Abadie, 2021; Athey et al., 2021).

Yet, contrary to conventional wisdom, it turns out that synthetic controls and unconfoundedness can yield numerically identical point estimates. As shown in Figure 1a, when the regression models are learned via ordinary least squares (OLS) or principal component regression (PCR), then the two approaches produce the same counterfactual economic evolution for Basque Country in the absence of terrorism. Figure 1b, on the other hand, demonstrates that when the regression models are learned via lasso ($\ell_1$ regularization) or are enforced to lie within the simplex—the original proposition of Abadie and Gardeazabal (2003)—then the two approaches output contrasting economic trajectories. Curiously, Figure 2 suggests that even though synthetic controls and unconfoundedness can arrive at the same estimates, their respective uncertainty levels can be markedly different. The juxtaposition of Figures 1 and 2 beg the following questions:

Q1: “When do synthetic controls and unconfoundedness estimates agree?”
Q2: “When the estimates agree, why are the uncertainty levels different?”
Q3: “In general, how do we decide between the two approaches?”
**Contribution.** This article tackles Q1–Q3 from first principles. To this end, we analyze the finite-sample estimation and large-sample inference properties of unconfoundedness and synthetic controls. We first classify the most widely studied regression formulations into (i) a symmetric class that yields numerically identical estimates and (ii) an asymmetric class that yields contrasting estimates. We then consider a generalized regression framework and establish that the statistical uncertainty of the counterfactual prediction is governed by time series and cross-sectional patterns. Accordingly, we construct a mixed confidence interval that accounts for randomness across both time and space. We illustrate its advantages over inference procedures that only consider one dimension using data-inspired simulations and empirical applications. In this view, we find the one-approach-fits-all mindset that is trending in several fields to be unwise. Instead, we advocate for panel data agnostic (PANDA) regression, which is rooted in model checking and based on symmetric estimators and mixed confidence intervals, when the data generating process is unbeknown to the researcher.
Y can only observe one economic state, each year, one that is immune to terrorism and another that is affected by terrorism. In reality, however, we the potential outcomes framework posits that each region possesses two possible levels of economic activity corresponding to its outcome in the absence and presence of a binary treatment, respectively. Philosophically, each unit $i$ is characterized by two potential outcomes (Neyman, 1923; Rubin, 1974), $Y_{it}(0)$ and $Y_{it}(1)$, which correspond to its outcome in the absence and presence of a binary treatment, respectively. Philosophically, the potential outcomes framework posits that each region possesses two possible levels of economic activity each year, one that is immune to terrorism and another that is affected by terrorism. In reality, however, we can only observe one economic state, $Y_{it}(0)$ or $Y_{it}(1)$—this is the fundamental challenge of causal inference.

Throughout, let $Y_{it}$ be the observed outcome. In a standard panel data study, we observe all $N$ units without treatment (control) for $T_0$ time periods, i.e., $Y_{it} = Y_{it}(0)$ for all $i \leq N$ and $t \leq T_0$. For the remaining $T_1 = T - T_0$ time periods, $N_1$ units receive control treatment while the remaining $N_0 = N - N_1$ units remain under control, i.e., if we arbitrarily label the first $N_0$ units as the control group, then $Y_{it} = Y_{it}(1)$ for all $i > N_0$ and $t > T_0$, and $Y_{it} = Y_{it}(0)$ for all $i \leq N_0$ and $t > T_0$. In our study, Basque Country is the single treated unit, thus $N_1 = 1$ and $N_0 = 16$. The first wave of terrorist activity came in 1970, thus the duration of the pre- and post-treatment periods are $T_0 = 15$ and $T_1 = 28$ years, respectively.

For ease of exposition, the remainder of this article considers the simplified setting of a single treated unit and single treated period indexed by the $N$th unit and $T$th time period, respectively. However, our analysis to follow is applicable for any $(i,t)$ pair where $i > N_0$ is a treated unit and $t > T_0$ is a treated period. Under this setting, we organize our observed control data into an $N \times T$ matrix, $Y = [Y_{it}]$, as shown in Figure 3. In our example, $y_N = [Y_{Nt} : t \leq T_0] \in \mathbb{R}^{T_0}$ represents Basque Country’s economic evolution prior to the onset of terrorist activity; $Y_0 = [Y_{it} : i \leq N_0, t \leq T_0] \in \mathbb{R}^{N_0 \times T_0}$ represents the control regions’ economic evolution prior to the onset of terrorist activity; and $Y_T = [Y_{iT} : i \leq N_0] \in \mathbb{R}^{N_0}$ represents the control regions’ economic evolution after the onset of terrorist activity. Our object of interest is Basque Country’s counterfactual GDP in the absence of terrorism, $Y_{NT}(0)$.

Two panel data regressions. Given the spread of information across space and time, there are two natural ways to impute the missing $(N,T)$th entry. These perspectives are explored in two large and mostly separate bodies of work (Athey et al., 2021): (i) unconfoundedness and (ii) synthetic controls.
**I: Unconfoundedness.** Unconfoundedness operates on the concept that “history is a guide to the future”. As such, unconfoundedness expresses outcomes in the treated period as a weighted composition of outcomes in the pretreatment periods. This is carried out by regressing the control units’ treated period outcomes, $y_T$, on its lagged outcomes, $Y_0$; the learned regression coefficients are then applied to the treated unit’s lagged outcomes, $y_N$, to predict the missing $(N,T)$th entry. In other words, the predictive model is learned on the control units’ outcomes across all time and applied on the treated unit’s pretreatment outcomes.

**II: Synthetic controls.** Synthetic controls is built on the conception that “similar units behave similarly”. Therefore, synthetic controls expresses the treated unit’s outcomes as a weighted composition of control units’ outcomes. This is carried out by regressing the treated unit’s lagged outcomes, $y_N$, on the control units’ lagged outcomes, $Y_0'$; the learned regression coefficients are then applied to the control units’ treated period outcomes, $y_T$, to predict the missing $(N,T)$th entry. In other words, the predictive model is learned on pretreatment outcomes across all units and applied on the control units’ treated period outcomes.

A curious duality. The two approaches simultaneously exhibit a symmetric and an asymmetric relationship. Symmetrically, both approaches appear to implicitly rely on time series and cross-sectional structures to impute the missing counterfactual. Asymmetrically, each approach explicitly exploits one of the two structures to learn a regression model. Because of the asymmetry, Athey et al. (2021) refers to unconfoundedness as horizontal (HZ) regression and synthetic controls as vertical (VT) regression. We explore this curious duality.

### 3 Finite-Sample Point Estimation

**Q1:** “When do HZ and VT estimates agree?”

We tackle Q1 by studying the finite-sample estimation properties of HZ and VT regressions. We denote the singular value decomposition of $Y_0$ as $Y_0 = \sum_{\ell=1}^{R} s_{\ell} u_{\ell} v_{\ell}' = USV'$, where $u_{\ell} \in \mathbb{R}^{N_0}$ and $v_{\ell} \in \mathbb{R}^{T_0}$ are the left and right singular vectors, respectively, $s_{\ell} \in \mathbb{R}$ are the ordered singular values, and $R = \text{rank}(Y_0) \leq \min\{N_0, T_0\}$. Moreover, $U \in \mathbb{R}^{N_0 \times R}$ and $V \in \mathbb{R}^{T_0 \times R}$ denote the matrices formed by the left and right singular vectors, respectively, and $S \in \mathbb{R}^{R \times R}$ is the diagonal matrix of singular values. With this notation, the Moore-Penrose pseudoinverse of $Y_0$ is $Y_0^\dagger = \sum_{\ell=1}^{R} (1/s_{\ell}) v_{\ell} u_{\ell}' = VS^{-1}U'$.

### 3.1 Notable Regression Formulations

We present the most commonly studied regression formulations in the HZ and VT literatures. Critically, we do not place any assumptions on the relative magnitudes of $N$ and $T$. 

---

Figure 3: Panel data format with rows and columns indexed by units and time, respectively.
3.1.1 Description of Estimation Strategies

**Least squares.** A large class of least squares formulations are expressed as follows:

(a) **HZ Regression:** for \( \lambda_1, \lambda_2 \geq 0 \),
\[
\hat{\alpha} = \arg\min_{\alpha} \| y_T - Y_0 \alpha \|_2^2 + \lambda_1 \| \alpha \|_1 + \lambda_2 \| \alpha \|_2^2 \tag{1}
\]
\[
\hat{Y}^\text{hz}_{NT}(0) = (y_N, \hat{\alpha}).
\]

(b) **VT Regression:** for \( \lambda_1, \lambda_2 \geq 0 \),
\[
\hat{\beta} = \arg\min_{\beta} \| y_N - Y_0^T \beta \|_2^2 + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_2^2 \tag{2}
\]
\[
\hat{Y}^\text{vt}_{NT}(0) = (y_T, \hat{\beta}).
\]

Below, we overview common choices for \((\lambda_1, \lambda_2)\) and describe the corresponding estimation strategy.

**I:** Ordinary least squares (OLS). Arguably, the mother of all regressions is OLS, where \( \lambda_1 = \lambda_2 = 0 \). OLS is an unconstrained problem with possibly infinitely many solutions. Within the panel data literature, OLS has been analyzed in numerous works, including Hsiao et al. (2012), Li and Bell (2017), and Li (2020).

**II:** Principal component regression (PCR). Next, we consider PCR (Jolliffe, 1982). To formalize PCR, let \( Y_0^{(k)} = \sum_{\ell=1}^{k} s_{\ell} u_{\ell} v_{\ell}^T \) denote the rank \( k < R \) approximation of \( Y_0 \) that retains the top \( k \) principal components. HZ and VT PCR corresponds to replacing \( Y_0 \) with \( Y_0^{(k)} \) within (1) and (2), respectively, with \( \lambda_1 = \lambda_2 = 0 \). In words, PCR first finds a \( k \) dimensional representation of the covariate matrix via principal component analysis (PCA); then, PCR performs OLS with the compressed \( k \) dimensional covariates. Within the synthetic controls literature, Amjad et al. (2018, 2019) and Agarwal et al. (2021b) utilize PCR when \( Y_0 \) is low rank.

**III:** Ridge regression. Consider ridge regression, where \( \lambda_1 = 0 \) and \( \lambda_2 > 0 \) (Hoerl and Kennard, 1970). When \( Y_0 \) is rank deficient, the gram matrix, i.e., \( Y_0^T Y_0 \) for HZ regression and \( Y_0 Y_0^T \) for VT regression, is ill-conditioned. This often discourages the usage of OLS. In these settings, ridge provides a remedy by adding a *ridge* on the diagonal of the gram matrix (e.g., \( Y_0^T Y_0 + \lambda_2 I \)), which increases all eigenvalues by \( \lambda_2 \), thus removing the singularity problem. Put another way, when \( \lambda_2 > 0 \), the criterion in (1) and (2) are strictly convex so the solutions are unique. Within the synthetic controls literature, Ben-Michael et al. (2021) explores the properties of a doubly robust estimator that utilizes HZ ridge regression.

**IV:** Lasso regression. Consider lasso regression, where \( \lambda_1 > 0 \) and \( \lambda_2 = 0 \) (Tibshirani, 1996; Chen et al., 1998). Unlike ridge, lasso has become a popular tool for estimating sparse linear coefficients in high-dimensional regimes. Unlike ridge, the lasso criterion is convex but not strictly convex. As such, there is either a unique solution or infinitely many solutions. In our analysis of lasso only, we make the mild assumption that the entries of \( Y_0 \) are drawn from a continuous distribution. As established in Tibshirani (2013), this guarantees that the lasso solution is unique. Several notable works in the synthetic controls literature, e.g., Li and Bell (2017), Carvalho et al. (2018), and Chernozhukov et al. (2021), analyze the lasso.
V: Elastic net regression. Mixing both $\ell_1$ and $\ell_2$-penalties, i.e., $\lambda_1, \lambda_2 > 0$, is known as elastic net regression (Zou and Hastie, 2005). As noted in Hastie (2020), elastic net selects variables similar to the lasso, but deals with correlated variables more gracefully as with ridge. When $\lambda_2 > 0$, the criterion is strictly convex, which yields a unique solution. Doudchenko and Imbens (2016) propose an elastic net synthetic controls variant.

VI: Simplex regression. The next formulation constrains the regression weights to lie within the simplex:

(a) HZ Regression: for $\lambda \geq 0$,

$$\hat{\alpha} = \text{argmin}_{\alpha} \| y_T - Y_0\alpha \|_2^2 + \lambda \| \alpha \|_2^2$$
subject to $\alpha'1 = 1, \alpha \succeq 0$ \hspace{1cm} (4)

$$\hat{y}_{hz}^{NT}(0) = \langle y_N, \hat{\alpha} \rangle.$$

(b) VT Regression: for $\lambda \geq 0$,

$$\hat{\beta} = \text{argmin}_{\beta} \| y_N - Y_0\beta \|_2^2 + \lambda \| \beta \|_2^2$$
subject to $\beta'1 = 1, \beta \succeq 0$ \hspace{1cm} (5)

$$\hat{y}_{vt}^{NT}(0) = \langle y_T, \hat{\beta} \rangle.$$

While standard formulations set $\lambda = 0$, we consider a vanishing $\ell_2$ penalty since $\lambda = 0$ can induce multiple minima (Doudchenko and Imbens, 2016; Abadie and L’Hour, 2021; Arkhangelsky et al., 2021). When $\lambda > 0$, the criterion becomes strictly convex and the solution is unique. Simplex regression is studied extensively within the synthetic controls literature (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015; Abadie, 2021; Ferman and Pinto, 2021) and recent unconfoundedness literature (Arkhangelsky et al., 2021).

3.1.2 Formal Results

To answer Q1, we classify the regression formulations into two camps: (i) a symmetric class, where HZ and VT estimates agree, and (ii) an asymmetric class, where HZ and VT estimates disagree. We use the shorthand HZ $\neq$ VT if the two approaches produce numerically identical point estimates and HZ $= \neq$ VT otherwise.

I: Symmetric regressions. We begin by stating the regression formulations that uphold symmetry.

Theorem 1. HZ $= \neq$ VT for (i) OLS with $\hat{\alpha}$ and $\hat{\beta}$ as the minimum $\ell_2$-norm solutions, i.e.,

$$\hat{y}_{hz}^{NT}(0) = \hat{y}_{vt}^{NT}(0) = \langle y_N, Y_0^T y_T \rangle = \sum_{\ell=1}^{R} (1/s_{\ell}) \langle y_N, v_\ell \rangle \langle u_\ell, y_T \rangle;$$

(ii) PCR with $\hat{\alpha}$ and $\hat{\beta}$ as the minimum $\ell_2$-norm solutions, i.e.,

$$\hat{y}_{hz}^{NT}(0) = \hat{y}_{vt}^{NT}(0) = \langle y_N, (Y_0^{(k)})^T y_T \rangle = \sum_{\ell=1}^{k} (1/s_{\ell}) \langle y_N, v_\ell \rangle \langle u_\ell, y_T \rangle;$$

and (iii) ridge regression, i.e.,

$$\hat{y}_{hz}^{NT}(0) = \hat{y}_{vt}^{NT}(0) = \langle y_N, (Y_0'Y_0 + \lambda_2 I)^{-1} Y_0' y_T \rangle = \sum_{\ell=1}^{R} \frac{s_{\ell}}{s_{\ell}^2 + \lambda_2} \langle y_N, v_\ell \rangle \langle u_\ell, y_T \rangle.$$
Theorem 1 establishes that HZ and VT OLS models yield numerically identical point estimates regardless of the relative sizes of $N$ and $T$. The same holds true for PCR and ridge, though the symmetry breaks if the hyperparameters differ between the two models. Theorem 1 is likely to be an unfamiliar conception since it contradicts the general consensus that the two approaches are fundamentally different. In particular, it debunks the notion that HZ OLS and VT OLS are invalid when $T > N$ and $N > T$, respectively. The source of this misconception likely comes from the fact that infinitely many solutions exist when $Y_0$ is rank deficient. However, among these solutions is the one with minimum $\ell_2$-norm, which possesses several attractive properties. First, the solution is unique and entirely embedded within the rowspace of the “covariate” matrix, i.e., $Y_0$ for HZ regression and $Y_0'$ for VT regression. Arguably, the minimum $\ell_2$-norm solution is sufficient for valid inference since any element in the nullspace of the covariate matrix is mapped to zero. The minimum $\ell_2$-norm solution also emerges when an unregularized, undetermined least squares problem is optimized using gradient descent, a ubiquitous optimizer in practice. This phenomena is an example of “implicit regularization”, where the optimization algorithm is biased towards a particular global minima even though the bias is not explicit in the objective function (Neyshabur et al., 2015; Gunasekar et al., 2017).

II: Asymmetric regressions. Next, we define the class of formulations that fracture the symmetry.

Theorem 2. HZ $\neq$ VT for (i) lasso regression, (ii) elastic net regression, and (iii) simplex regression.

Theorem 2 establishes that the lasso, elastic net, and simplex regularization penalties each impose structural assumptions that induce HZ and VT asymmetry. The simplex result can be seen through an extreme scenario in which HZ and VT simplex regressions select a single most predictive pretreatment period and control unit, respectively. In this setting, the treated unit’s outcome during the most predictive pretreatment period may not be congruent with the treated period outcome of the most predictive control unit. For instance, consider $N = T$ with $Y_0 = I$, $y_N = e_i$, and $y_T = 2e_i$, where $e_i$ is the canonical basis whose $i$th coordinate is 1. Here, $\tilde{\alpha} = \tilde{\beta} = e_i$, yet $\tilde{y}_{N_T}(0) = 1$ and $\tilde{y}_{N_T}(0) = 2$. To explain the phenomena for lasso and elastic net, we recall that they, like simplex regression, promote sparsity (Abadie, 2021). Therefore, while sparsity is often deemed to be a salient feature of a regression model, it may also be the source of HZ and VT estimation asymmetry.

3.2 Including Intercept Terms

Intercepts can be included in regression models by a straightforward modification of the HZ and VT objective functions. Specifically, for HZ regression, the $\ell_2$-errors in (1) and (4) are updated as $\|y_T - Y_0\alpha - \alpha_01\|^2_2$; for VT regression, the $\ell_2$-errors in (2) and (5) are updated as $\|y_N - Y_0\beta - \beta_01\|^2_2$. To date, there is still a lack of general consensus on the inclusion of intercepts within the VT methodological literature. While the original works by Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2015) omit intercepts, several influential subsequent works advocate for its usage (Doudchenko and Imbens, 2016; Arkhangelsky et al., 2021; Ben-Michael et al., 2021; Ferman and Pinto, 2021). We attempt to shed light on the role of intercepts.

Proposition 1. HZ $\neq$ VT for (i) OLS, (ii) PCR, and (iii) ridge regression with intercepts.

We develop an intuition for Proposition 1 by interpreting intercepts in panel studies. A nonzero time intercept, $\alpha_0$, imposes a permanent constant difference between the treated period and pretreatment periods; a nonzero unit intercept, $\beta_0$, imposes a permanent constant difference between the treated unit and control units. From this perspective, intercepts may appear innocuous but they impose strong systematic structures on the data generating process (DGP). In turn, this creates a natural asymmetry between HZ and VT regressions.
4 Large-Sample Inference

Q2: “When the estimates agree, why are the HZ and VT uncertainty levels different?”

Towards answering Q2, we study the asymptotic properties of HZ and VT approaches. In this pursuit, we require probabilistic generative modeling assumptions. This section subscribes to concepts from the standard regression framework, which we believe is a natural starting point to extract valuable insights. For ease of exposition, we focus on OLS and its minimum $\ell_2$-norm solutions, $\hat{\alpha} = Y_0'y_T$ and $\hat{\beta} = (Y_0')^\dagger y_N$. Notably, our results immediately extend to PCR by replacing $Y_0$ with $Y_0^{(k)}$, as defined in (3), for any $k < R$. We state some of our results informally to improve readability and provide the precise forms in Appendix A.

4.1 HZ and VT Asymptotic Properties under Distinct Generative Models

We first study the asymptotic behaviors of HZ and VT regressions under their respective generative models.

4.1.1 HZ and VT Generative Models

Throughout, we denote $\varepsilon_T = [\varepsilon_{iT} : i \leq N_0]$ as the errors associated with the control units during treated period and $\varepsilon_N = [\varepsilon_{Nt} : t \leq T_0]$ as the errors associated with the treated unit during the pretreatment period.

Assumption 1 (HZ generative model). Let the following hold for HZ regression:

(i) Outcome model: (a) $y_T = Y_0'\alpha^* + \varepsilon_T$ and (b) $E[Y_{NT}(0)|y_N] = \langle y_N, \alpha^* \rangle$.

(ii) Strict exogeneity: $E[\varepsilon_T|Y_0] = 0$.

(iii) Independence: outcomes are sampled independently across units.

(iv) Subspace inclusion: (a) $\alpha^* \in \text{rowspan}(Y_0)$ or (b) $y_N \in \text{rowspan}(Y_0)$.

Assumption 2 (VT generative model). Let the following hold for VT regression:

(i) Outcome model: (a) $y_N = Y_0'\beta^* + \varepsilon_N$ and (b) $E[Y_{NT}(0)|y_T] = \langle y_T, \beta^* \rangle$.

(ii) Strict exogeneity: $E[\varepsilon_N|Y_0] = 0$.

(iii) Independence: outcomes are sampled independently across time.

(iv) Subspace inclusion: (a) $\beta^* \in \text{colsparse}(Y_0)$ or (b) $y_T \in \text{colsparse}(Y_0)$.

Outcome model. Assumption 1 (i) explicitly models time series patterns as a source of randomness in the HZ framework. Assumption 2 (i) explicitly models cross-sectional patterns as a source of randomness in the VT framework. Both DGPs posit that the correlation pattern obeys a linear relationship.

Strict exogeneity. Condition (ii) asserts that the errors are conditionally mean zero. This implies that the control units’ pretreatment outcomes (covariates) and errors are uncorrelated.
Independence. Assumption 1 (iii) allows time correlated errors in the HZ framework while Assumption 2 (iii) allows unit correlated errors in the VT framework.

Subspace inclusion. Though cosmetically different, condition (iv) is implicitly assumed in the standard OLS framework. Consider HZ regression. Typically, it is assumed that \( R = \text{rank}(Y_0) = T_0 \), i.e., \( \text{rowspan}(Y_0) = \mathbb{R}^{T_0} \). Accordingly, \( \alpha^* \in \text{rowspan}(Y_0) \) and \( y_N \in \text{rowspan}(Y_0) \), so (iv) immediately holds. From this perspective, Assumption 1 (iv) is a generalization of the canonical full column rank assumption to handle scenarios where \( Y_0 \) is rank deficient. This is our focus as we do not assume the sizes of \( N \) and \( T \).

To build further intuition, we zoom in on (iv.a) and consider the extreme scenario where \( \alpha^* \in \text{nullspace}(Y_0) \). There, \( \alpha^* \in \text{nullspace}(Y_0) \) becomes a meaningless predictive model. In contrast, if (iv.a) holds, then \( \alpha \) can recover \( \alpha^* \) and generalize to the \((N,T)\)th point. We apply a similar thought experiment to (iv.b) by considering \( y_N \in \text{nullspace}(Y_0) \). Again, \( \alpha \in \text{rowspan}(Y_0) \) is unlikely to predict well on \( y_N \) since \( \alpha \) is trained on in-sample data that are highly dissimilar (orthogonal) to \( y_N \). In contrast, if (iv.b) holds, then \( y_N \) resembles our in-sample data and can generalize to the \((N,T)\)th point, provided the in-sample error, \( Y_T - Y_0 \alpha \), is small. (iv.b) can also be rewritten as the existence of a vector \( b \) such that \( y_N = Y_T' b \). This is precisely Assumption 2 (i.a) if \( \varepsilon_N = 0 \). Hence, Assumption 1 (iv.b) effectively assumes the VT outcome model, but prohibits randomness. Since the same arguments apply for VT regression, this suggests HZ regression implicitly relies on cross-sectional correlations and VT regression implicitly relies on time series correlations.

Summary. Assumptions 1 and 2 describe fundamentally different DGPs that cannot simultaneously hold for the same data. From this perspective, it is arduous to compare the two approaches beyond point estimation.

4.1.2 Asymptotic Properties

Equipped with our assumptions, we are ready to answer Q2. We denote the HZ and VT causal estimands as \( \mu_{ht} = \mathbb{E}[Y_{NT}(0)|y_N] \) and \( \mu_{vt} = \mathbb{E}[Y_{NT}(0)|y_T] \), respectively. We denote the error covariance matrices as

\[
\Sigma_T = \text{Cov}(\varepsilon_T|Y_0) = \text{diag}(\sigma_{11}^2,\ldots,\sigma_{N1}^2)
\]

\[
\Sigma_N = \text{Cov}(\varepsilon_N|Y_0) = \text{diag}(\sigma_{N1}^2,\ldots,\sigma_{N_N}^2)
\]

**Theorem 3.** Let Assumption 1 hold. Conditioned on \((y_N,Y_0)\) and under suitable moment conditions,

\[
\frac{Y_{NT}(0) - \mu_{ht}^0}{\sqrt{\upsilon_{ht}^0}} \overset{d}{\rightarrow} \mathcal{N}(0,1),
\]

where \( \upsilon_{ht}^0 = \hat{\beta}' \Sigma_T \hat{\beta} \). Let Assumption 2 hold. Conditioned on \((y_T,Y_0)\) and under suitable moment conditions,

\[
\frac{Y_{NT}(0) - \mu_{vt}^0}{\sqrt{\upsilon_{vt}^0}} \overset{d}{\rightarrow} \mathcal{N}(0,1),
\]

where \( \upsilon_{vt}^0 = \hat{\alpha}' \Sigma_N \hat{\alpha} \).

Theorem 3 states that HZ and VT OLS converge to different asymptotic distributions. In particular, the statistical uncertainties of the HZ and VT predictions are purely governed by time series and cross-sectional patterns, respectively. This not only explains the curiosity behind Figure 2, but also the contrast in their uncertainty levels more broadly. We study the consequences of this asymmetry below.
4.1.3 Confidence Intervals

Theorem 3 motivates separate HZ and VT confidence intervals: for \( \theta \in (0, 1) \),

\[
\begin{align*}
\text{(HZ confidence interval)} & \quad \mu_0^{hz} \in \left[ \hat{Y}_{NT}^{hz}(0) \pm \frac{z_{\frac{\theta}{2}}}{\sqrt{v_0^{hz}}} \right], \\
\text{(VT confidence interval)} & \quad \mu_0^{vt} \in \left[ \hat{Y}_{NT}^{vt}(0) \pm \frac{z_{\frac{\theta}{2}}}{\sqrt{v_0^{vt}}} \right],
\end{align*}
\]

where \( z_{\frac{\theta}{2}} \) is the upper \( \theta/2 \) quantile of \( \mathcal{N}(0, 1) \), and \((\hat{v}_0^{hz}, \hat{v}_0^{vt})\) are the estimators of \((v_0^{hz}, v_0^{vt})\). We define

\[
\hat{v}_0^{hz} = \beta^T \hat{\Sigma}_T \hat{\beta} \quad \text{and} \quad \hat{v}_0^{vt} = \hat{\alpha}^T \hat{\Sigma}_N \hat{\alpha},
\]

where \((\hat{\Sigma}_T, \hat{\Sigma}_N)\) are the estimators of \((\Sigma_T, \Sigma_N)\). We precisely define these estimators under homoskedastic and heteroskedastic errors below. To reduce ambiguity, we index \(\hat{v}_0^{hz} \) and \(\hat{v}_0^{vt}\) by the covariance estimator. Finally, we denote \(H^{(u)} = UU^T, H^{(u)}_1 = I - H^{(u)}, H^{(v)} = VV^T\), and \(H^{(v)}_1 = I - H^{(v)}\). Thus, the HZ and VT in-sample prediction errors can be written as \(H^{(u)}_1y_T = y_T - Y_0\hat{\alpha} \) and \(H^{(v)}_1y_N = y_N - Y_0\hat{\beta}\), respectively.

**Homoskedastic errors.** Consider homoskedastic errors with \(\Sigma_T = \sigma_T^2 I\) and \(\Sigma_N = \sigma_N^2 I\). We design

\[
\begin{align*}
\hat{\Sigma}_T^{homo} & = \hat{\sigma}_T^2 I \quad \text{with} \quad \hat{\sigma}_T^2 = (N_0 - R)^{-1}\|H^{(u)}_1y_T\|_2^2, \\
\hat{\Sigma}_N^{homo} & = \hat{\sigma}_N^2 I \quad \text{with} \quad \hat{\sigma}_N^2 = (T_0 - R)^{-1}\|H^{(v)}_1y_N\|_2^2,
\end{align*}
\]

where \(R\) represents the rank of \(Y_0\), which can be computed as \(R = \text{tr}(H^{(u)}) = \text{tr}(H^{(v)})\).

**Lemma 1.** Consider homoskedastic errors. Let Assumption 1 hold. Then, \(\mathbb{E}[\hat{\Sigma}_T^{homo}|Y_0] = \Sigma_T; \) thus, \(\mathbb{E}[v_0^{hz,homo}|y_N, Y_0] = v_0^{hz}.\) Let Assumption 2 hold. Then, \(\mathbb{E}[\hat{\Sigma}_N^{homo}|Y_0] = \Sigma_N; \) thus, \(\mathbb{E}[v_0^{vt,homo}|y_T, Y_0] = v_0^{vt}.\)

Lemma 1 is a well known result within the OLS literature, albeit the standard result is formalized under the stricter full column (or row) rank assumption as opposed to Assumption 1 (iv) (or Assumption 2 (iv)).

**Heteroskedastic errors.** Under the heteroskedastic setting, we adopt two estimation strategies.

I. Jackknife. The first estimator is based on the jackknife. Traditionally, the jackknife estimates the covariance of the regression models \((\hat{\alpha}, \hat{\beta})\). By analyzing said estimates, we can then derive the following:

\[
\begin{align*}
\hat{\Sigma}_T^{jack} & = \text{diag} \left( \left[ H^{(u)}_1 \circ H^{(u)}_1 \circ I \right] \left[ H^{(u)}_1 y_T \circ H^{(u)}_1 y_T \right] \right), \\
\hat{\Sigma}_N^{jack} & = \text{diag} \left( \left[ H^{(v)}_1 \circ H^{(v)}_1 \circ I \right] \left[ H^{(v)}_1 y_N \circ H^{(v)}_1 y_N \right] \right).
\end{align*}
\]

**Lemma 2.** Consider heteroskedastic errors. Let Assumption 1 hold. If \((H^{(u)}_1 \circ H^{(u)}_1 \circ I)\) is nonsingular, then \(\mathbb{E}[\hat{\Sigma}_T^{jack}|Y_0] \geq \Sigma_T; \) thus, \(\mathbb{E}[v_0^{hz,jack}|y_N, Y_0] \geq v_0^{hz}.\) Let Assumption 2 hold. If \((H^{(v)}_1 \circ H^{(v)}_1 \circ I)\) is nonsingular, then \(\mathbb{E}[\hat{\Sigma}_N^{jack}|Y_0] \geq \Sigma_N; \) thus, \(\mathbb{E}[v_0^{vt,jack}|y_T, Y_0] \geq v_0^{vt}.\)

Lemma 2 establishes that the jackknife is conservative, provided \((H^{(u)}_1 \circ H^{(u)}_1 \circ I)\) and \((H^{(v)}_1 \circ H^{(v)}_1 \circ I)\) are nonsingular. Strictly speaking, the jackknife is well defined if these quantities are singular, as seen through the pseudoinverse in (8) and (9). Lemma 2 considers the nonsingular case for simplicity. We remark that \(\max_i H_{ii}^{(v)} < 1\) and \(\max_i H_{ii}^{(v)} < 1\) are sufficient conditions for invertibility.
If: HRK-estimator. Next, we consider the covariance estimator proposed by Hartley et al. (1969). We index this estimator by the authors, Hartley-Rao-Kiefer.

\[
\Sigma_{HRK}^T = \text{diag}
\left(
\frac{H^T v \perp \circ H^T v \perp y^T \circ H^T v \perp y^T}{H^T u \perp \circ H^T u \perp}
\right)
\]

\[
\Sigma_{HRK}^N = \text{diag}
\left(
\frac{H^T v \perp \circ H^T v \perp y^T \circ H^T v \perp y^T}{H^T u \perp \circ H^T u \perp}
\right)
\]

Lemma 3. Consider heteroskedastic errors. Let Assumption 1 hold. If \((H^T u \perp \circ H^T u \perp)\) is nonsingular, then
\[
E[\Sigma_{HRK}^T | Y_0] = \Sigma_T \quad \text{and} \quad E[v_{hz, HRK} | y_N, Y_0] = v_{hz}^0.
\]

Let Assumption 2 hold. If \((H^T v \perp \circ H^T v \perp)\) is nonsingular, then
\[
E[\Sigma_{HRK}^N | Y_0] = \Sigma_N \quad \text{and} \quad E[v_{vt, HRK} | y_T, Y_0] = v_{vt}^0.
\]

Lemma 3 establishes that the HRK estimator is unbiased, provided \((H^T u \perp \circ H^T u \perp)\) and \((H^T v \perp \circ H^T v \perp)\) are invertible. For the former quantity, \(\max \ell H^T u \perp \ell < 1/2\) is a sufficient condition for invertibility. Since \(\text{tr}(H^T u \perp) = R\), this restricts \(R < N_0/2\). A similar conclusion is drawn for VT regression.

Remark 1. A benefit of PCR, in contrast to OLS, is that the nonsingularity conditions required for the jackknife and HRK covariance estimators can be controlled by the number of chosen principal components \(k\).

4.1.4 Which Confidence Interval Should I Trust—HZ or VT?

Given distinct HZ and VT confidence intervals, how do we decide which to trust? Towards providing a response, we first underscore that the purpose of a confidence interval is to quantify uncertainties in the posited DGP. While some assumptions require oracle information for verification, others can be checked from observed data. In our context, the verifiable conditions correspond to Assumptions 1 and 2 (i.a) and (iv.b).

With this in mind, we inspect the intervals of Section 4.1.3. Notably, each interval only depends on its particular in-sample error. The HZ error quantifies the uncertainty in the HZ outcome model (presence of time series patterns) in Assumption 1 (i.a). Critically, HZ intervals overlook the VT error. While the VT error quantifies the uncertainty in the VT outcome model in Assumption 2 (i.a), it also checks Assumption 1 (iv.b). Hence, the VT error is an informative measure for the validity of the HZ DGP, yet is completely neglected by the HZ intervals. The same arguments apply to VT regression. Conceptually, this reinforces the intuition that both regressions rely on time series and cross-sectional patterns for valid inference. Computationally, this suggests that both intervals should incorporate HZ and VT errors. For these reasons, we conclude that neither is suitable in isolation as each may lead to undercoverage. This motivates the upcoming discussion.

4.2 HZ and VT Asymptotic Properties under a Mixed Generative Model

We look to develop a confidence interval that properly accounts for uncertainties across time and space.

4.2.1 Mixed Generative Model

We present a DGP that mixes the HZ and VT DGPs. More formally, we model randomness along both dimensions of the data but place additional constraints on the stochastic properties of the errors.
Assumption 3 (Mixed generative model). Let the following hold:

(i) Outcome model: (a) \( y_t = Y_0\alpha^* + \varepsilon_T \), (b) \( y_N = Y_0'\beta^* + \varepsilon_N \), and (c) \( E[Y_{NT}(0)|Y_0] = (\beta^*, Y_0\alpha^*) \).

(ii) Strict exogeneity: (a) \( E[\varepsilon_T|Y_0] = 0 \) and (b) \( E[\varepsilon_N|Y_0] = 0 \).

(iii) Independence: outcomes are sampled independently across units and time.

Outcome model. Roughly speaking, the mixed model makes weaker outcome modeling assumptions than the HZ and VT models. Assumption 3 (i.a) and (i.b) correspond to Assumption 1 (i.a) and Assumption 2 (i.a), respectively. This explicitly models time series and cross-sectional patterns as two distinct sources of randomness. From the HZ perspective, Assumption 3 (i.b) replaces Assumption 1 (iv), which implicitly disallows cross-sectional uncertainties; from the VT perspective, Assumption 3 (i.a) replaces Assumption 2 (iv), which implicitly disallows time series uncertainties. Assumption 3 (i.a) and (i.b) also justify Assumption 3 (i.c). To see this, consider the HZ DGP with Assumption 3 (i.a) in place of Assumption 1 (iv). Then,

\[
E[Y_{NT}(0)|Y_0] = E[E[Y_{NT}(0)|y_N, Y_0]|Y_0] = E[E[Y_{NT}(0)|y_N]|Y_0] = E[(y_N, \alpha^*)|Y_0] = (\beta^*, Y_0\alpha^*) \]

Now, consider the VT DGP with Assumption 3 (i.a) in place of Assumption 2 (iv). Then,

\[
E[Y_{NT}(0)|Y_0] = E[E[Y_{NT}(0)|y_T, Y_0]|Y_0] = E[E[Y_{NT}(0)|y_T]|Y_0] = E[(y_T, \beta^*)|Y_0] = (\beta^*, Y_0\alpha^*)
\]

The same conclusion is drawn from both perspectives.

Strict exogeneity. Assumption 3 (ii) combines Assumptions 1 and 2 (ii). In words, it states that \( Y_0 \) contains all measured confounders. In some sense, the HZ DGP implies Assumption 3 (ii.b) since \( y_N \) is treated deterministically; a similar logic applies to Assumption 3 (ii.a) and the VT DGP.

Independence. Assumption 3 (iii) combines Assumptions 1 and 2 (iii). In words, it states that the errors are independent across time and space, which is more restrictive than Assumptions 1 and 2 (iii) in isolation. We leave a proper analysis under dependent errors as future work, possibly by leveraging Li (2020).

Summary. Assumption 3 suggests the following DGP: (i) sample \( Y_0 \); (ii) sample \( (\varepsilon_T, \varepsilon_N) \) independent of all other variables in the system while satisfying \( E[\varepsilon_T|Y_0] = 0 \) and \( E[\varepsilon_N|Y_0] = 0 \); (iii) define \( y_T = Y_0\alpha^* + \varepsilon_T \), \( y_N = Y_0'\beta^* + \varepsilon_N \), and \( E[Y_{NT}(0)|Y_0] = (\beta^*, Y_0\alpha^*) \).

4.2.2 Asymptotic Properties

We establish the HZ and VT asymptotic properties under the mixed DGP. To this end, let \( \mu_0 = E[Y_{NT}(0)|Y_0] \).

Theorem 4. Let Assumption 3 hold. Conditioned on \( Y_0 \) and under suitable moment conditions,

\[
\frac{\hat{Y}_{NT}(0) - \mu_0}{\sqrt{v_0}} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{and} \quad \frac{\hat{Y}_{NT}(0) - \mu_0}{\sqrt{v_0}} \xrightarrow{d} \mathcal{N}(0, 1),
\]

where \( v_0 = (H^{(u)}\beta^*)'\Sigma_T(H^{(u)}\beta^*) + (H^{(e)}\alpha^*)'\Sigma_N(H^{(e)}\alpha^*) + \text{tr}(Y_0'\Sigma_T(Y_0')\Sigma_N) \).

Theorem 4 establishes that under the mixed DGP, both regressions converge to the same normal distribution with mean \( \mu_0 \) and asymptotic variance \( v_0 \). The first two terms in \( v_0 \) measure the variability of time series
and cross-sectional patterns while the third term measures their interaction. The third term naturally arises due to the quadratic form of the \((N,T)\)th prediction involving \((y_N, y_T)\), which are jointly random.

**Core intuition.** Theorem 4 affirms the statistical uncertainty of the \((N,T)\)th prediction is governed by time series and cross-sectional patterns. To internalize this, take the HZ perspective, which exploits time series correlations. If the treated unit is starkly different from the control units, then the regression model \(\hat{\alpha}\) learned on the control units should not extrapolate to the treated unit even if the treated period is similar to the pretreatment periods. Now, take the VT perspective, which exploits cross-sectional correlations. If the treated period is starkly different from the pretreatment periods, then the regression model \(\hat{\beta}\) learned during the pretreatment periods should not extrapolate to the treated period, even if the treated unit is similar to the control units. In other words, the HZ approach inevitably relies on cross-sectional patterns and the VT approach inevitably relies on time series patterns—neither pattern is individually sufficient for inference. Accordingly, we argue that the prediction uncertainty should account for both sources of randomness.

**Remark 2.** This intuition serves as the bedrock of this article and guides the upcoming discussions. Crucially, it is built upon the methodological framework of HZ and VT regressions, not our specific modeling assumptions. Therefore, our primary takeaways to follow are likely applicable for other inferential settings as well.

### 4.2.3 Confidence Interval

Building on our newfound insight, we construct a mixed confidence interval as follows: for \(\theta \in (0,1)\),

\[
\text{(mixed confidence interval)} \quad \mu_0 \in \left[ \hat{Y}_{NT}(0) \pm z_{\theta} \sqrt{\hat{v}_0} \right].
\]

Here, \(\hat{Y}_{NT}(0) = \hat{Y}^{hz}_{NT}(0) = \hat{Y}^{vt}_{NT}(0)\) (as per Theorem 1) and \(\hat{v}_0\) is the estimator of \(v_0\). The HZ and VT regressions now share the same point prediction and confidence interval, as reflective of Theorem 4. Motivated by Lemmas 1–3, we design our novel variance estimator as

\[
\hat{v}_0 = \hat{v}^{hz}_0 + \hat{v}^{vt}_0 - \text{tr}(Y_0' \hat{\Sigma}_T(Y_0')' \hat{\Sigma}_N),
\]

where \((\hat{v}^{hz}_0, \hat{v}^{vt}_0)\) are given by (6). We define \((\hat{\Sigma}_T, \hat{\Sigma}_N)\) using the estimators in Section 4.1.3. We index \(\hat{v}_0\) by said estimators, e.g., \(\hat{v}_0^{\text{jack}}\) uses \((\hat{\Sigma}^{\text{jack}}_T, \hat{\Sigma}^{\text{jack}}_N)\). In all cases, \(\hat{v}_0\) depends on both HZ and VT in-sample errors.

We consider \(\hat{v}_0\) with respect to \(v_0\). By construction, \(\hat{\alpha} = H^{(v)} \hat{\alpha}\) and \(\hat{\beta} = H^{(u)} \hat{\beta}\). To justify the negative trace in \(\hat{v}_0\), recall \(\hat{v}_0^{hza}\) is a quadratic involving \((y_N, y_T)\). Since both quantities are random, the expectation of \(\hat{v}_0^{hza}\) induces an additional term that precisely corresponds to the interactive trace in \(v_0\). The same property holds for \(\hat{v}_0^{vt}\). Thus, \(\hat{v}_0\) corrects for this bias via the negative trace. We formalize this notion below.

**Corollary 1.** Let Assumption 3 hold. (i) Under homoskedastic errors, \(E[\hat{v}^{\text{homo}}_0 | Y_0] = v_0\). (ii) Consider heteroskedastic errors. If \((H^{(u)}_0 \circ H^{(u)}_0 \circ I)\) and \((H^{(v)}_0 \circ H^{(v)}_0 \circ \Sigma_N)\) are nonsingular, then \(E[\hat{v}_0^{\text{jack}} | Y_0] \geq v_0\). (iii) Consider heteroskedastic errors. If \((H^{(u)}_0 \circ H^{(u)}_0)\) and \((H^{(v)}_0 \circ H^{(v)}_0)\) are nonsingular, then \(E[\hat{v}_0^{\text{HRK}} | Y_0] = v_0\).

Corollary 1 states that the homoskedastic and HRK based estimators are unbiased while the jackknife based estimator is conservative. Crucially, these results only hold in expectation. For any realization, \(\hat{v}_0\) may exhibit undesirable characteristics. Indeed, if \(\text{tr}(Y_0' \hat{\Sigma}_T(Y_0')' \Sigma_N) > \max\{\hat{v}^{hz}_0, \hat{v}^{vt}_0\}\), then \(\hat{v}_0 < \min\{\hat{v}^{hz}_0, \hat{v}^{vt}_0\}\); hence, the mixed coverage will be smaller than the HZ and VT coverages. In such a scenario, one naïve solution is to modify \(\hat{v}_0\) as \(\hat{v}^{\text{mod}}_0 = \hat{v}^{hz}_0 + \hat{v}^{vt}_0\), which is conservative as per Corollary 1.
4.3 The Advantage of a Mixed Confidence Interval

Armed with our mixed interval, we explore its performance relative to HZ and VT intervals. We consider two canonical synthetic controls studies: terrorism in Basque Country and the reunification of West Germany (Abadie et al., 2015). The latter study examines the economic impact of the 1990 reunification in West Germany. Our panel data contains per capita GDP of \( N = 17 \) countries across \( T = 44 \) years. There are \( T_0 = 30 \) pretreatment observations and \( N_0 = 16 \) control units. Similar to the Basque study, our interest here is to estimate West Germany’s GDP in the absence of reunification. We provide an overview of the results and primary messages below, and relegate details to Appendix B.

4.3.1 Data-Inspired Simulation Studies

To assess the performance of the uncertainty estimators in a realistic environment, we calibrate our simulations to the two case studies of interest. For a proper analysis, we perform model checks for Assumption 3.

Data generating process. In each study, we hold out a portion of the observed control data as a “test” set. Specifically, we only consider the pretreatment outcomes across all units, e.g., the GDP of all Spanish regions prior to the outset of terrorism in Basque Country. These observations constitute our ground truth of expected potential outcomes. In congruence with the article, we consider the single treated unit and single treated period setting. We retain the actual treated unit, e.g., Basque Country, but artificially designate the final pretreatment period, e.g., one year prior to the outset of terrorism, as the pseudo treated period. We define our causal estimand as \( \mu_0 = Y_{N_{T_0}} = Y_{N_{T_0}}(0) \), e.g., Basque Country’s GDP one year prior to the outset of terrorism. We fix \( \mu_0 \) and \( Y_0 \), which we temporarily redefine as \( Y_0 = Y_{it} : i \leq N_0, t < T_0 \); note that \( Y_0 \) excludes outcomes in the actual final pretreatment period. For stability analysis, we randomly perturb the treated unit’s pretreatment outcomes and the control units’ pseudo treated period outcomes. Specifically, in each simulation repeat, we independently sample \((y_N, y_{T_0})\) from their respective Gaussian distributions whose means are given by \([Y_{Nt} : t < T_0]\) and \([Y_{iT} : i \leq N_0]\), respectively. This design choice respects Assumption 3 (ii)–(iii). We also enforce homoskedasticity and designate \( \sigma_N^2 \) and \( \sigma_T^2 \) as the observed standard errors of \([Y_{Nt} : t < T_0]\) and \([Y_{iT} : i \leq N_0]\), respectively. By construction, our simulated DGPs reflect the underlying DGPs of the case studies since (i) \((\mu_0, Y_0)\) and (ii) the expected values of \((y_N, y_{T_0})\) are all observed quantities. Therefore, our conclusions to follow may be transferrable to other practical settings.

Simulation results. We conduct 500 replications of the above DGP for each case study.

1: Prediction errors. Table 1 summarizes the OLS and PCR prediction errors across four metrics: (i-ii) HZ and VT in-sample errors, (iii) bias, and (iv) root-mean-squared-error (RMSE). Our findings suggest that linear predictors adequately fit the underlying outcomes, i.e., Assumption 3 (i) is reasonable. The notable exceptions are the high bias and RMSE for OLS in the Basque study, which is unsurprising given Figure 1.

| Case study  | OLS               | PCR               |
|-------------|-------------------|-------------------|
|             | HZ error | VT error | Bias  | RMSE | HZ error | VT error | Bias  | RMSE |
| Basque      | 0.08     | 0.00     | -3.30  | 261   | 0.22     | 0.16     | 0.01   | 0.15  |
| W. Germany  | 0.00     | 0.01     | -0.01  | 0.01  | 0.01     | 0.01     | -0.01  | 0.01  |
II: Coverage. After checking our modeling assumptions, we are ready to compare the performances of the HZ, VT, mixed, and modified mixed intervals. Table 2 shows the coverage probabilities (CP) and average lengths (AL) for nominal 95% confidence intervals of OLS and PCR. To facilitate our comparisons, we normalize the coverage length by the magnitude of the corresponding prediction value.

For each OLS study, either the HZ or VT interval is degenerate. Subsequently, the mixed interval achieves the same coverage as the non-degenerate interval. The PCR HZ and VT intervals are non-degenerate, though their CPs can be relatively low compared to the mixed intervals. The exception is the standard mixed jackknife interval for the Basque study. In summary, our simulations provide evidence that the mixed interval coverage is generally closer to the nominal coverage compared to HZ and VT intervals, which often undercover.

Table 2: Coverage results for OLS and PCR nominal 95% confidence intervals across 500 simulation replications.

| Case study | OLS Homoskedastic | OLS Jackknife | PCR Homoskedastic | PCR Jackknife |
|------------|-------------------|---------------|-------------------|---------------|
|            | HZ    | VT    | Mixed | HZ    | VT    | Mixed | Mod. | HZ    | VT    | Mixed | Mod. |
| Basque (CP)| 0.81  | 0.00  | 0.81  | 0.99  | 0.00  | 0.99  | 0.80  | 0.84  | 0.92  | 0.95  | 0.82  | 0.85  | 0.80  | 0.97  |
| Basque (AL)| 14.2  | 0.00  | 14.2  | 211   | 0.00  | 211   | 0.39  | 0.45  | 0.54  | 0.56  | 0.57  | 0.54  | 0.62  | 0.75  |
| W. Ger. (CP)| 0.00  | 0.90  | 0.90  | 0.00  | 0.98  | 0.98  | 0.65  | 0.94  | 0.97  | 0.97  | 0.74  | 0.86  | 0.96  | 0.96  |
| W. Ger. (AL)| 0.00  | 0.04  | 0.04  | 0.00  | 0.08  | 0.08  | 0.02  | 0.03  | 0.03  | 0.03  | 0.02  | 0.03  | 0.04  | 0.04  |

4.3.2 Empirical Illustrations

Next, we apply our uncertainty estimators to analyze the two case studies. We construct an illustrative narrative around Figures 4 and 5 with the help of three researchers: one HZ, one VT, and one “agnostic”. We note that the following discussion only concerns the mixed interval \( \hat{\delta}_0 \) as defined in (10).

Our three researchers begin with the Basque study and use OLS. Despite their different beliefs of the underlying DGP, they independently arrive at the same counterfactual trajectory that suggests terrorism is net positive for the Basque economy. Although the OLS predictions are clearly implausible, the VT intervals shown in Figures 4b and 4e reveal zero uncertainty. In contrast, the HZ intervals shown in Figures 4a and 4d reflect massive levels of uncertainty, which resonates with intuition. Finally, the agnostic researcher who values both of their colleagues’ perspectives, employs the mixed confidence intervals displayed in Figures 4c and 4f. Like the HZ intervals, the mixed intervals provide quantitative evidence against the OLS model.

The three researchers then transition to the West Germany study. This time, the HZ intervals shown in Figures 4g and 4j reveal zero uncertainty. The VT intervals shown in Figures 4h and 4k, on the other hand, expand over time. Since the HZ intervals are degenerate, the mixed intervals in Figures 4i and 4l match the VT intervals. Our researchers observe a similar storyline when they use PCR, as shown in Figure 5.

Through our researchers’ experiences, we observe that the mixed interval can relieve the burden on the researcher of deliberating between which interval—the HZ or the VT—to trust. Put differently, the mixed interval can safeguard a researcher from drawing causal conclusions from the “less informative” interval. The robustness feature of the mixed interval is a direct consequence of accounting for both sources of uncertainty in the underlying DGP. Combined with our simulation studies and discussion in Section 4.1.4, we argue that the mixed interval is generally advantageous over HZ and VT intervals.
Figure 4: OLS predictions and confidence intervals. The top two rows of figures correspond to the Basque study while the bottom two rows of figures correspond to the West Germany study. We index rows by the covariance estimator, e.g., jackknife, and the columns by the type of confidence interval, e.g., HZ.
Figure 5: PCR predictions and confidence intervals. The top two rows of figures correspond to the Basque study while the bottom two rows of figures correspond to the West Germany study. We index rows by the covariance estimator, e.g., jackknife, and the columns by the type of confidence interval, e.g., HZ.
4.4 Extension to Ridge Regression

Asymptotic results for ridge can be established under the generative models stated in Assumptions 1–3. Like OLS and PCR, HZ and VT ridge converge to different distributions under their respective DGPs, but converge to the same distribution under the mixed DGP. However, since ridge inflates bias for reduced variance, inference is generally more challenging (Dempster et al., 1977; van Wieringen, 2015).

5 The Researcher’s Guide to Panel Data

Q3: “In general, how do we decide between HZ and VT regressions?”

5.1 Current Approaches

**Data configuration principle.** A widely accepted rule-of-thumb is to choose the regression based on the configuration of the data. Often, there is limited information in one form or another, which lends the problem to a particular approach. Concretely, HZ and VT are cautioned against when \( T \gg N \) and \( N \gg T \), respectively (Abadie et al., 2015; Doudchenko and Imbens, 2016; Li and Bell, 2017; Athey et al., 2021; Ferman and Pinto, 2021). However, Athey et al. (2021) argues that regularization allows the two approaches to be simultaneously applicable for the same setting, which then enables a more systematic comparison. Nevertheless, VT methods continue to be the leading approach for panel data studies across many literatures.

**Doubly robust estimation.** A second popular response is to remove the burden of deciding altogether. This is an old idea that dates back to the difference-in-differences (DID) estimator (Ashenfelter, 1978; Bertrand et al., 2004; Angrist and Pischke, 2009; Bechtel and Hainmueller, 2011), which is still widely used in applied economic research today. At a high level, DID posits an additive outcome model with unit- and time-specific fixed effects, known more colloquially as the “parallel trends” assumption. The recent work of Arkhangelsky et al. (2021) anchors on the DID principle and brings in concepts from the unconfoundedness and synthetic controls literatures to derive a doubly robust estimator called synthetic difference-in-differences (SDID). In our setting, the SDID prediction for the missing \((N,T)\)th entry can be written as

\[
\hat{Y}_{NT}^{\text{sdid}}(0) = \sum_{i \leq N_0} \hat{\beta}_i Y_{iT} + \sum_{t \leq T_0} \hat{\alpha}_t Y_{Nt} - \sum_{i \leq N_0} \sum_{t \leq T_0} \hat{\beta}_i \hat{\alpha}_t Y_{it} = \langle \mathbf{y}_T, \hat{\beta} \rangle + \langle \mathbf{y}_N, \hat{\alpha} \rangle - \langle \mathbf{y}_0, \hat{\alpha} \rangle,
\]

where \(\hat{\alpha}\) and \(\hat{\beta}\) represent general HZ and VT models, respectively. The authors note that \(\hat{\alpha} = (1/T_0)\mathbf{1}\) and \(\hat{\beta} = (1/N_0)\mathbf{1}\) recovers DID. Moving beyond simple DID to performing a weighted two-way bias removal, the authors propose to learn \(\hat{\alpha}\) via simplex regression and \(\hat{\beta}\) via simplex regression with an \(\ell_2\)-penalty.

Another notable work is that of Ben-Michael et al. (2021). The authors introduce the augmented synthetic control (ASC) estimator, which uses an outcome model to correct the bias induced by the classical synthetic controls estimator. Concretely, the ASC estimator predicts the \((N,T)\)th missing potential outcome as

\[
\hat{Y}_{NT}^{\text{asc}}(0) = \hat{M}_{NT}(0) + \sum_{i \leq N_0} \hat{\beta}_i (Y_{iT} - \hat{M}_{iT}(0)),
\]
where $\hat{M}_{IT}(0)$ is the estimator for the $(i, T)$th entry. The authors instantiate $\hat{M}_{IT}(0)$ as
\[
\hat{M}_{IT}(0) = \sum_{t \leq T_0} \alpha_t Y_{it} = \langle y_i, \alpha \rangle,
\] (13)

where $y_i = [Y_{it} : t \leq T_0]$ for any $i \leq N$. Plugging the HZ outcome model in (13) into (12) then gives
\[
\hat{Y}_{NIT}(0) = \langle y_T, \hat{\beta} \rangle + \langle y_N, \hat{\alpha} \rangle - \langle \hat{\beta}, Y_0 \hat{\alpha} \rangle.
\]

We consider this particular variant of ASC, which takes the same form as SDID, as seen in (11). In contrast to Arkhangelsky et al. (2021), Ben-Michael et al. (2021) learns $\hat{\alpha}$ via ridge regression and $\hat{\beta}$ via simplex regression. Given their rising popularities, we look to gain a deeper understanding of both approaches by studying their estimation characteristics when $(\hat{\alpha}, \hat{\beta})$ are learned via OLS and PCR.

**Corollary 2.** SDID = ASC = HZ = VT for (i) $\hat{\alpha}$ and $\hat{\beta}$ as the OLS minimum $\ell_2$-norm solutions and (ii) $\hat{\alpha}$ and $\hat{\beta}$ as the PCR minimum $\ell_2$-norm solutions.

Somewhat surprisingly, Corollary 2 states that a researcher who uses the unregularized variants of SDID and ASC inevitably arrive at the same point estimate as their colleague who simply uses HZ or VT OLS. The same phenomena occurs for PCR. From this, we observe that SDID and ASC lose their weighted double-differencing effects under certain formulations. However, Corollary 2 is not to discredit either approach. The fact remains that both methods allow researchers to naturally and simultaneously encode their knowledge of time and unit specific structures into the estimator. SDID and ASC warrant further studies and careful consideration.

### 5.2 Our Approach—Panel Data Agnostic (PANDA) Regression

To answer Q3, we summarize two insights derived in this article: (i) HZ and VT regressions yield numerically identical estimates under several formulations; (ii) time series and cross-sectional correlations play critical roles for valid inference. In this view, both HZ and VT perspectives deserve our appreciation and attention. Neglecting either side can lead to misinformed conclusions, as showcased in Section 4.3. Thus, this article challenges the one-approach-fits-all mindset that is trending in several fields. It also discourages biases towards asymmetric formulations without justifiable cause as they implicitly establish a hierarchy between the two approaches. With the hope of changing course, this article offers an alternative perspective:

*if the DGP is unknown, then the researcher should be agnostic.*

The core philosophy of PANDA regression is to advocate for the coexistence of both regressions when there is little to no evidence that suggests one correlation pattern is more meaningful than the other. We translate philosophy into practice through a (i) symmetric estimator and (ii) mixed confidence interval. PANDA regression relieves the burden of deliberation placed on the researcher—the same counterfactual estimate is derived from both perspectives and the uncertainty associated with each perspective is accounted for. Through the lens of model checking, we discuss facets of PANDA regression to construct principles for best practice.

**I: Symmetric estimator.** In many applications, it is not known a priori which estimator to use. Systematic comparisons such as cross-validation and stability analyses are effective instruments for such cases (Yu and Kumbier, 2020; Athey et al., 2021). If these evaluations do not suggest a particular formulation, then this article recommends PCR. To arrive at this viewpoint, we compare the estimators within the symmetric class.
OLS is a natural first candidate since it excludes hyperparameters that can induce more degrees of freedom and thus subjectivity. However, applying OLS in limited data regimes can lead to specious outcomes, as illustrated in Figure 1. PCR and ridge can overcome this drawback through regularization. Both methods are also connected to one another. Ridge can be viewed as a “smooth” variant of PCR: in ridge, higher order singular values are largely unaffected by the penalty while the lower order singular values are heavily penalized; in PCR, the higher order singular values are untouched while the lower order singular values are removed entirely. Since it is arguably easier to conduct inference for PCR than ridge, we elevate PCR as the benchmark. We refer the interested reader to Agarwal et al. (2021c,a) for detailed analyses of PCR.

II: Mixed confidence interval. As argued in Section 4.3, our mixed confidence interval is generally advantageous over HZ and VT specific intervals. The aim of this discussion, however, is not to promote our specific inference procedure. In fact, \( \hat{\nu}_0 \) is unlikely to be optimal given that it is an artifact of our assumed DGP and choice of analysis. There are numerous other approaches with differing philosophies on the sources of randomness, e.g., design-based permutation inference (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015; Doudchenko and Imbens, 2016; Ferman and Pinto, 2017; Bottmer et al., 2021), time series-based permutation inference (Chernozhukov et al., 2021), cross-sectional-based permutation inference (Shaikh and Toulis, 2019), prediction intervals (Cattaneo et al., 2021), and large-sample approximations under correct specification (Li, 2020; Agarwal et al., 2021b). Every approach possesses their unique merits. Extensive model checking, therefore, is critical towards identifying the most suitable approach for a given problem. All in all, the purpose of this discussion is to motivate future methods to account for both sources of randomness rather than just one. We hope this article serves as a stepping stone in this new direction.

Remark 3. Model checks cannot validate an analysis, but they can reveal problems. Thus, the checks may suggest that linear predictors are ill-suited and none of the HZ, VT, or mixed intervals are valid.

6 Beyond Panel Data

The purpose of this section is to contextualize our findings in settings beyond panel data.

6.1 Estimation and Inference with Covariates

We depart from panel data to a setting with both outcomes and covariates. For cohesiveness, we continue with the Basque study but now consider a different dataset, the details of which are provided in Abadie and Gardeazabal (2003) and Abadie et al. (2011). Specifically, for each region \( i \), we observe (pretreatment) covariates \( x_i \in \mathbb{R}^{T_0} \), which includes per-capita GDP (the outcome variable), as well as population density, sectoral production, investment, and human capital (the auxiliary predictor variables) all prior to the onset of terrorism. Focusing on the single treated period setting, we denote \( Y_i \) as the observed GDP for the \( i \)th region one year after the onset of terrorism. Let \( X = [x_i : i \leq N_0] \) and \( y = [Y_i : i \leq N_0] \) denote the matrix of covariates and vector of outcomes for the control regions, respectively. In line with the rest of this article, our interest is to predict Basque Country’s GDP in the absence of terrorism, \( Y_N(0) \).

Two general solutions. This is a well studied problem in the causal inference literature. Among the numerous existing methodologies, we discuss two prominent approaches.
I: Covariate adjustment. One classical approach is known as covariate adjustment. At a high level, covariate adjustment posits a (linear) relationship between outcome variables and covariates that is stable across units. Methodologically, this is carried out by regressing the control regions’ treated period GDPs, $y$, on the control regions’ covariates, $X$; the learned regression coefficients are then applied to Basque Country’s covariates, $x_N$, to predict Basque Country’s GDP in the absence of terrorism.

II: Synthetic controls. Synthetic controls has also emerged as a popular tool. Philosophically, synthetic controls posits a relationship between treated and control units that is stable across covariates and outcome variables. Methodologically, this is carried out by regressing Basque Country’s covariates, $x_N$, on the control regions’ covariates, $X’$; the learned regression coefficients are then applied to control regions’ treated period GDPs, $y$, to predict Basque Country’s GDP in the absence of terrorism.

An “obvious” duality. Recalling our earlier notation and mapping $x_N$ to $y_N$, $X$ to $Y_0$, and $y$ to $y_T$, we identify covariate adjustment and synthetic controls as simply $HZ$ and $VT$ regression, respectively. Therefore, Theorem 1 asserts that if the regression weights are learned via OLS, PCR, or ridge, then the two methods produce numerically identical estimates for $Y_N(0)$. Theorem 4 then states that valid inference relies on cross-sectional patterns and a linear relationship between outcomes and covariates. Conceptually, this suggests that synthetic controls and covariate adjustment are two sides of the same coin. As a result, we can interpret covariate adjustment through the lens of synthetic controls and vice versa.

6.2 General Linear Regression

We step out of the causal inference frame of reference to look at the problem of linear regression in generality, whereby panel data and learning with covariates are special instances. Theorem 4 states that the statistical uncertainty of the out-of-sample prediction is governed by two distinct sources of randomness: (i) the construction of the regression model from the in-sample data, and (ii) the similarity between the in- and out-of-sample data. Standard OLS uncertainty estimators, which correspond to $HZ$ and $VT$ specific intervals in panel studies, account for the first source of randomness. They are rigorous when the design matrix is full column rank. By contrast, the mixed OLS interval is derived without the canonical full column rank assumption or restrictions on the data configuration. Accordingly, it incorporates both sources of randomness and thus, generalizes the standard OLS interval. In this view, our results in Section 4 build upon the broader OLS literature, particularly in the high-dimensional regime, which may be of independent interest.

7 Conclusion

We show that unconfoundedness and synthetic controls yield numerically identical estimates under several popular settings. We then consider a generalized regression framework and establish that valid inference relies on time series and cross-sectional patterns. Accordingly, we construct a mixed confidence interval that accounts for both sources of uncertainty, and illustrate its advantages over inference procedures that consider one source. When the DGP is unknown, we advocate for panel data agnostic (PANDA) regression, which is rooted in model checking and based on symmetric estimators and mixed confidence intervals. As a byproduct, our findings (i) imply that synthetic controls and covariate adjustment are symmetric perspectives and (ii) build upon the general OLS literature.
References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

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).

Abadie, A., Diamond, A., and Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, 42(13):1–17.

Abadie, A., Diamond, A., and Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, 42(13):1–17.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review*, 93(1):113–132.

Abadie, A. and L’Hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817–1834.

Agarwal, A., Shah, D., and Shen, D. (2021a). On principal component regression in a high-dimensional error-in-variables setting. *arXiv preprint arXiv:2010.14449*.

Agarwal, A., Shah, D., and Shen, D. (2021b). Synthetic interventions. *arXiv preprint arXiv:2006.07691*.

Agarwal, A., Shah, D., Shen, D., and Song, D. (2021c). On robustness of principal component regression. *Journal of the American Statistical Association*, 116(536):1731–1745.

Amjad, M., Misra, V., Shah, D., and Shen, D. (2019). Mrsc: Multi-dimensional robust synthetic control. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(2).

Amjad, M., Shah, D., and Shen, D. (2018). Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51.

Angrist, J. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion*.

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.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 60(1):47–57.

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. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Bechtel, M. M. and Hainmueller, J. (2011). How lasting is voter gratitude? an analysis of the short- and long-term electoral returns to beneficial policy. *American Journal of Political Science*, 55(4):852–868.
Ben-Michael, E., Feller, A., and Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates?* The Quarterly Journal of Economics, 119(1):249–275.

Bottmer, L., Imbens, G., Spiess, J., and Warnick, M. (2021). A design-based perspective on synthetic control methods.

Carvalho, C., Masini, R., and Medeiros, M. C. (2018). Arco: An artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics*, 207(2):352–380.

Cattaneo, M. D., Feng, Y., and Titiunik, R. (2021). Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864.

Cline, R. E. (1965). Representations for the generalized inverse of sums of matrices. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, 2(1):99–114.

Dempster, A. P., Schatzoff, M., and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, 72(357):77–91.

Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Working Paper 22791, National Bureau of Economic Research.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, volume 38. Siam.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Statistics, pages 59–82. University of California Press, Berkeley, CA.

Ferman, B. and Pinto, C. (2017). Placebo Tests for Synthetic Controls. MPRA Paper 78079, University Library of Munich, Germany.

Ferman, B. and Pinto, C. (2021). Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12(4):1197–1221.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017). Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*.

Hartley, H. O., Rao, J. N. K., and Kiefer, G. (1969). Variance estimation with one unit per stratum. *Journal of the American Statistical Association*, 64(327):841–851.

Hastie, T. (2020). Ridge regularization: An essential concept in data science. *Technometrics*, 62(4):426–433.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.
Hoff, P. D. (2017). Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115(C):186–198.

Hsiao, C., Steve Ching, H., and Ki Wan, S. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with mainland China. *Journal of Applied Econometrics*, 27(5):705–740.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Statistics, pages 221–233. University of California Press, Berkeley, CA.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, USA.

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

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society*, 31(3):300–303.

King, G. and Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, 14(2):131–159.

Lehmann, E. (2000). Elements of large-sample theory.

Li, K. T. (2020). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 115(532):2068–2083.

Li, K. T. and Bell, D. R. (2017). Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics*, 197(1):65–75.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

Meyer, C. D. (1973). Generalized inversion of modified matrices. *SIAM Journal on Applied Mathematics*, 24(3):315–323.

Miller, R. G. (1974). An unbalanced jackknife. *The Annals of Statistics*, 2(5):880–891.

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51. in Polish.

Neyshabur, B., Tomioka, R., and Srebro, N. (2015). In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations*.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press.

Shaikh, A. and Toulis, P. (2019). Randomization tests in observational studies with staggered adoption of treatment.
Styan, G. P. (1973). Hadamard products and multivariate statistical analysis. *Linear Algebra and its Applications*, 6:217–240.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456 – 1490.

van Wieringen, W. N. (2015). Lecture notes on ridge regression.

Varga, R. S. (1962). *Matrix Iterative Analysis*. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, volume 1 of *MIT Press Books*. The MIT Press.

Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14(4):1261 – 1295.

Yu, B. and Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.
A Large-Sample Inference: Part II

A.1 Formal Results

We provide the formal statements, coupled with technical discussions, of the informal statements in Section 4.

A.1.1 HZ and VT Regressions under Distinct Generative Models

First, we state the precise form of Theorem 3.

**Theorem 5.** Let Assumption 1 hold. Conditioning on \((y_N, Y_0)\), if

\[
\left( \sum_{i \leq N_0} E \left[ |\beta_i \epsilon_{iT}|^3 \right] y_N, Y_0 \right)^2 = o \left( \left( \sum_{i \leq N_0} \beta_i^2 \sigma_{iT}^2 \right)^3 \right),
\]

then

\[
\frac{\hat{Y}_{hT}(0) - \mu_{h}}{\sqrt{v_{h0}}} \to d N(0, 1),
\]

where \(v_{h0} = \bar{\beta}' \Sigma_T \bar{\beta} \). Let Assumption 2 hold. Conditioning on \((y_T, Y_0)\), if

\[
\left( \sum_{t \leq T_0} E \left[ |\alpha_t \epsilon_{NT}|^3 \right] y_T, Y_0 \right)^2 = o \left( \left( \sum_{t \leq T_0} \alpha_t^2 \sigma_{NT}^2 \right)^3 \right),
\]

then

\[
\frac{\hat{Y}_{vT}(0) - \mu_{v}}{\sqrt{v_{v0}}} \to d N(0, 1),
\]

where \(v_{v0} = \bar{\alpha}' \Sigma_N \bar{\alpha} \).

If \((\epsilon_{iT}, \sigma_{iT}^2)\) are bounded for all \(i\), then (14) translates to \(\sum_{i \leq N_0} |\beta_i|^3 = o(\|\bar{\beta}\|_2^3)\), which rules out outlier coefficients; a similar interpretation can be derived for (15). We note that (14) and (15) are known as Lyapunov’s condition, and we refer the interested reader to Lehmann (2000) for details.

A.1.2 HZ and VT Regressions under a Mixed Generative Model

We now state the precise form of Theorem 4.
Theorem 6. Let Assumption 3 hold. Conditioning on $Y_0$, if

$$
\left( \sum_{i \leq N_0} \sum_{t \leq T_0} \mathbb{E} \left[ \left[ (Y^i_0)_{it} \left\{ \mathbb{E}[Y_{iT}|Y_0] \varepsilon_{Nt} + \mathbb{E}[Y_{NT}|Y_0] \varepsilon_{NT} + \varepsilon_{iT} \varepsilon_{NT} \right\} \right] \bigg| Y_0 \right] \right)^2
$$

$$
= o \left( \left( \sum_{i \leq N_0} \sum_{t \leq T_0} (Y^i_0)_{it}^2 \left\{ \mathbb{E}[Y_{iT}|Y_0]^2 \sigma^2_{Nt} + \mathbb{E}[Y_{NT}|Y_0]^2 \sigma^2_{NT} + \sigma^2_{iT} \sigma^2_{NT} \right\} \right)^3 \right),
$$

(16)

then

$$
\frac{\hat{Y}^h_{N}(0) - \mu_0}{\sqrt{\nu_0}} \overset{d}{\rightarrow} \mathcal{N}(0, 1) \quad \text{and} \quad \frac{\hat{Y}^v_{N}(0) - \mu_0}{\sqrt{\nu_0}} \overset{d}{\rightarrow} \mathcal{N}(0, 1),
$$

where $v_0 = (H^{(v)} \beta^*)' \Sigma_T (H^{(v)} \beta^*) + (H^{(v)} \alpha^*)' \Sigma_N (H^{(v)} \alpha^*) + \text{tr} (Y^i_0 \Sigma_T (Y^i_0)' \Sigma_N)$.

Finally, similar to our interpretation of (14) and (15), observe that if $(\varepsilon_{iT}, \varepsilon_{NT})$ and $(\sigma^2_{iT}, \sigma^2_{NT})$ are bounded for all $(i, t)$, then (16) loosely translates to

$$
\sum_{i \leq N_0} |\hat{\beta}_i|^2 + \sum_{t \leq T_0} |\hat{\alpha}_t|^2 + \sum_{i \leq N_0} \sum_{t \leq T_0} |(Y^i_0)_{it}|^2 = o \left( \|\hat{\beta}\|^2_2 + \|\hat{\alpha}\|^2_2 + \|Y^i_0\|^2_2 \right),
$$

which effectively bounds the magnitudes of the $H^2$ and $VT$ OLS coefficients and pseudoinverse matrix entries.

Next, we provide a bound on the trace term in $v_0$. Beginning with the upper bound, observe that

$$
\text{tr} (Y^i_0 \Sigma_T (Y^i_0)' \Sigma_N) \leq \max_{i \leq N_0} \sigma^2_{iT} \max_{t \leq T_0} \sigma^2_{NT} \text{tr} (Y^i_0 (Y^i_0)').
$$

By the cyclic property of the trace operator,

$$
\text{tr} (Y^i_0 (Y^i_0)') = \text{tr} (V S^{-2} V') = \text{tr} (S^{-2} V' V) = \text{tr} (S^{-2}).
$$

Putting everything together, we obtain

$$
\text{tr} (Y^i_0 \Sigma_T (Y^i_0)' \Sigma_N) \leq \max_{i \leq N_0} \sigma^2_{iT} \max_{t \leq T_0} \sigma^2_{NT} \text{tr} (S^{-2}).
$$

The same arguments can be applied to derive the lower bound, which yields

$$
\text{tr} (Y^i_0 \Sigma_T (Y^i_0)' \Sigma_N) \in \text{tr} (S^{-2}) \left[ \min_{i \leq N_0} \sigma^2_{iT}, \min_{t \leq T_0} \sigma^2_{NT}, \max_{i \leq N_0} \sigma^2_{iT}, \max_{t \leq T_0} \sigma^2_{NT} \right].
$$

(17)

A.1.3 Variance Estimators

Next, we discuss technical aspects of the covariance estimators presented in Section 4.1.3.

Homoskedastic errors. We take note of the recent work of Agarwal et al. (2021b) in the synthetic controls literature. Agarwal et al. (2021b) propose a VT PCR estimator under the homoskedastic setting and provide a similar confidence interval to that of (7) via large-sample approximations. In particular, they propose $\hat{\beta}' \hat{\Sigma}^\text{homo}_N \hat{\beta}$ in place of $\hat{\alpha}' \hat{\Sigma}^\text{homo}_N \hat{\alpha}$. While the point estimate of Agarwal et al. (2021b) also takes the form $(\hat{y}_T, \hat{\beta})$, their variance estimator only depends on $(y_N, Y_0)$; in comparison, ours depends on $(y_N, y_T, Y_0)$. 

Therefore, the confidence interval as per Agarwal et al. (2021b) is numerically identical for every post-treatment estimate while ours can vary across the post-treatment periods, which may be favorable.

**Heteroskedastic errors.** Consider the heteroskedastic setting.

**I: Jackknife.** To contextualize the jackknife, we highlight that the jackknife holds close relation to the Eicker-Huber-White (EHW) heteroskedastic robust covariance matrix (Eicker, 1967; Huber, 1967; White, 1980). The EHW estimator, which only includes the in-sample errors, has been noted to exhibit poor performance given finite samples. As such, there have been numerous modifications to the EHW estimator, including one motivated by the jackknife. Among the EHW variants, Long and Ervin (2000) recommends the jackknife correction based on their simulation studies. Its properties are studied in Miller (1974), Efron (1982), and Wu (1986). In the following lemma, we quantify the bias in Lemma 2.

**Lemma 4.** Consider heteroskedastic errors. Let Assumption 1 hold. If \((H^{(u)}_\perp \circ H^{(u)}_\perp \circ I)\) is nonsingular, then \(\mathbb{E}[^{\text{jack}}]\Sigma_T | \bar{Y}_0 = \Sigma_T + \Delta\), where

\[
\Delta_{\ell\ell} = \sum_{j \neq \ell} \frac{(H^{(u)}_{ij})^2}{(1 - H^{(u)}_{ij})^2}\sigma_{jT}^2; \tag{18}
\]

thus, \(\mathbb{E}[^{\text{haz,jack}}] [y_N, \bar{y}_0] = v_0^{hz} + \alpha' \Delta \alpha\). Let Assumption 2 hold. If \((H^{(u)}_\perp \circ H^{(u)}_\perp \circ I)\) is nonsingular, then \(\mathbb{E}[^{\text{jack}}]\Sigma_N | \bar{Y}_0 = \Sigma_N + \Gamma\), where

\[
\Gamma_{\ell\ell} = \sum_{j \neq \ell} \frac{(H^{(u)}_{ij})^2}{(1 - H^{(u)}_{ij})^2}\sigma_{Nj}^2; \tag{19}
\]

thus, \(\mathbb{E}[^{\text{vt,jack}}] [y_T, \bar{y}_0] = v_0^{vt} + \beta' \Gamma \beta\).

We consider the magnitudes of \((\Delta_{\ell\ell}, \Gamma_{\ell\ell})\). Towards bounding the former quantity, notice that \(H^{(u)}\) is an orthogonal projector and is thus idempotent, i.e., \((H^{(u)}_u)^2 = H^{(u)}_u\), and symmetric. Therefore,

\[
H^{(u)}_{\ell\ell} = (H^{(u)}_\ell)^2 + \sum_{j \neq \ell} (H^{(u)}_{ij})^2 = \sum_{j \neq \ell} (H^{(u)}_{ij})^2 = H^{(u)}_\ell (1 - H^{(u)}_\ell). \tag{20}
\]

Plugging (20) into (18) gives \(\Delta_{\ell\ell} \in H^{(u)}_\ell (1 - H^{(u)}_\ell)^{-1}[\min_{j \neq \ell} \sigma_{jT}^2, \max_{j \neq \ell} \sigma_{jT}^2]\). Since \(\sum_{j \neq \ell} (H^{(u)}_{ij})^2 \geq 0\), (20) implies \(H^{(u)}_\ell \in [0, 1]\). Thus, \(\Delta_{\ell\ell} = 0\) if \(H^{(u)}_\ell = 0\) and diverges if \(H^{(u)}_\ell = 1\). If \(H^{(u)}_\ell\) takes the average value \(N_0^{-1} \sum H^{(u)}_{\ell\ell} = RN_0^{-1}\), then \(\Delta_{\ell\ell} \in R(N_0 - R)^{-1}[\min_{j \neq \ell} \sigma_{jT}^2, \max_{j \neq \ell} \sigma_{jT}^2]\). A similar result is derived for \(\Gamma_{\ell\ell}\).

**II: HRK-estimator.** Recall Lemma 3. Consider the HZ estimator and the invertibility of \((H^{(u)}_u \circ H^{(u)}_u)\). A sufficient condition is strict diagonal dominance (Varga, 1962): \((1 - H^{(u)}_\ell)^2 > \sum_{j \neq \ell} (H^{(u)}_{ij})^2\). Using (20), we simplify this condition as \((1 - H^{(u)}_\ell)^2 > H^{(u)}_\ell - (H^{(u)}_\ell)^2\). Thus, \(\max_i H^{(u)}_{\ell\ell} < 1/2\) is a sufficient condition for invertibility. The same arguments apply for \(\Sigma_T\) regression.

**Mixed variance estimator.** Let us first bound \(\hat{v}_0\). From (17), it follows that \(\hat{v}_0 \in [\hat{v}_0^{\text{min}}, \hat{v}_0^{\text{max}}]\), where

\[
\hat{v}_0^{\text{min}} = \min_{i \leq N_0} \hat{\alpha}_{iT}^2 \|\hat{\beta}\|^2_2 + \min_{t \leq T_0} \hat{\sigma}_{Nt}^2 \|\hat{\alpha}\|^2_2 - \max_{i \leq N_0} \hat{\sigma}_{iT}^2 \max_{t \leq T_0} \hat{\sigma}_{Nt}^2 \text{tr}(S^{-2}),
\]

\[
\hat{v}_0^{\text{max}} = \max_{i \leq N_0} \hat{\sigma}_{iT}^2 \|\hat{\beta}\|^2_2 + \max_{t \leq T_0} \hat{\sigma}_{Nt}^2 \|\hat{\alpha}\|^2_2 - \min_{i \leq N_0} \hat{\sigma}_{iT}^2 \min_{t \leq T_0} \hat{\sigma}_{Nt}^2 \text{tr}(S^{-2}).
\]

Observe that \(\hat{v}_0^{\text{min}} = \hat{v}_0^{\text{max}}\) for homoskedastic errors.
Next, we formalize Corollary 1. In particular, we quantify the bias of the jackknife by leveraging Lemma 4.

**Corollary 3.** Let Assumption 3 hold. (i) Consider homoskedastic errors. Then, \( E[\hat{r}_0^{\text{home}}|Y_0] = v_0 \). (ii) Consider heteroskedastic errors. If \((H_+^u \circ H_+^u \circ I)\) and \((H_+^v \circ H_+^v \circ I)\) are nonsingular, then

\[
E[\hat{r}_0^{\text{jack}}(Y_0)|Y_0] = v_0 + \left(H_+^u \beta^* \Delta(H_+^u \beta^*) + (H_+^v \alpha^*) \Gamma(H_+^v \alpha^*) + \text{tr}(Y_0^1 \Delta(Y_0^1) \Gamma)\right),
\]

where \( \Delta \) and \( \Gamma \) are defined as in (18) and (19), respectively. (iii) Consider heteroskedastic errors. If \((H_+^u \circ H_+^u)\) and \((H_+^v \circ H_+^v)\) are nonsingular, then \( E[\hat{r}_0^{\text{HTK}}|Y_0] = v_0 \).

### A.2 Learning from the Mistakes of Others

As highlighted in Section 4.1.4, the confidence intervals derived for each approach only depends on its particular in-sample error, though the other approach’s error remains informative for the chosen approach. This motivates an alternative natural proposition to the mixed confidence interval of Section 4.2.3: for a given regression direction, use the other direction’s in-sample error as a diagnostic to examine the validity of Assumptions 1 and 2 (iv.b). More formally, we define the following statistics:

\[
\tau^{hx}(y_N, Y_0) = \frac{\|H_+^v y_N\|_2^2}{\|y_N\|_2^2} \quad \text{and} \quad \tau^{vt}(y_T, Y_0) = \frac{\|H_+^u y_T\|_2^2}{\|y_T\|_2^2}. \tag{21}
\]

The numerators of \( \tau^{hx}(y_N, Y_0) \) and \( \tau^{vt}(y_T, Y_0) \) are the VT and HZ in-sample errors, respectively. We use \( \tau^{hx}(y_N, Y_0) \) to determine if \((y_N, Y_0)\) obeys Assumption 1 (iv.b), and use \( \tau^{vt}(y_T, Y_0) \) to determine if \((y_T, Y_0)\) obeys Assumption 2 (iv.b). From this perspective, \( \tau^{hx} \) and \( \tau^{vt} \) can serve as a first line of defense to identify out-of-sample inputs, e.g., \( y_N \) for HZ regression, that will be difficult for \( \hat{\alpha} \) and \( \hat{\beta} \) to predict.

Given (21), it is natural to ask how we determine which inputs are flagged? By construction, \( \tau^{hx} \) and \( \tau^{vt} \) take values within the interval \([0, 1]\). If \( y_N \in \text{rowspan}(Y_0) \), then \( \tau^{hx}(y_N, Y_0) = 0 \), otherwise \( \tau^{hx}(y_N, Y_0) > 0 \). Similarly, if \( y_T \in \text{colsan}(Y_0) \), then \( \tau^{vt}(y_T, Y_0) = 0 \), otherwise \( \tau^{vt}(y_T, Y_0) > 0 \). However, we may not wish to be so strict on using zero as a hard threshold. Instead, given that \( 1 - \tau^{hx} \) and \( 1 - \tau^{vt} \) are similar to the standard \( R^2 \) metric used in regression analyses, we can transfer the same principles used to determine “acceptable” \( R^2 \) scores to our setting of interest.

Crucially, we underscore that \( \tau^{hx} \approx 0 \) and \( \tau^{vt} \approx 0 \) do not suggest that the \((N, T)\)th predictions will necessarily be accurate. These values only reduce our uncertainty in conditions (i.a) and (iv.b)—no more no less. Indeed, there still remains other potential sources of uncertainty, i.e., other conditions may be violated without our knowledge. To fully determine the suitability of HZ or VT regression towards a given dataset, we would ideally examine the validity of each aspect of their posited DGP. Currently, only a subset of the conditions in Assumptions 1 and 2 are commonly checked. As discussed, condition (i.a) is traditionally evaluated through the chosen direction’s in-sample error. Conditions (ii) and (iii) are commonly investigated through residual analyses, though they cannot ever be fully verified. Conditions (i.b) and (iv.a) involve unobservable quantities that make it arduous if not impossible to furnish data driven metrics. In light of this, we advocate treating \( \tau^{hx} \) and \( \tau^{vt} \) as one-sided tests: if \( \tau^{hx} \approx 1 \) or \( \tau^{vt} \approx 1 \), then we should be cautious about our out-of-sample prediction as our uncertainty in the overall composition of assumptions has only increased.

**Comparison with the literature.** To the best of our knowledge, similar metrics have remained largely
absent in the literature, likely because traditional frameworks have not explored condition (iv) in great depth. Below, we discuss connections with the few related works in the causal inference literature.

I: **Agarwal et al. (2021b)**. Our proposed statistics presented in (21) are motivated by the subspace inclusion statistic of Agarwal et al. (2021b). More specifically, Agarwal et al. (2021b) introduces $\tau_{vt}(y_T, Y_0)$ to check if Assumption 2 (iv.b) holds in practice. Notably, Agarwal et al. (2021b) focuses purely on VT regression. As such, Agarwal et al. (2021b) do not consider $\tau_{hz}(y_N,Y_0)$. A key aspect of this work is the observation that HZ regression implicitly relies on cross-sectional correlations and VT regression implicitly relies on time series correlations. Therefore, just as Agarwal et al. (2021b) introduces $\tau_{vt}(y_T, Y_0)$ as a useful metric for VT regression, this work introduces $\tau_{hz}(y_N,Y_0)$ as a useful metric for HZ regression.

II: **King and Zeng (2006)**. To facilitate this comparison, consider HZ regression. Similar to the focus of this section, King and Zeng (2006) set out to provide a useful measure to guide researchers in identifying causal estimands that can be answered. In our context, the authors argue that $\alpha$ can be safely applied to $y_N$ if $y_N$ is embedded within the convex hull of $Y_0$, i.e., there exists weights $w$ satisfying $w \succeq 0, \mathbf{1}'w = 1$, and $y_N = Y_0w$. Despite their different derivations, the convex hull metric and (21) are based on the same concept: they both relate the prediction quality of $\hat{Y}_{NT}^{hz}(0) = \langle y_N, \hat{\alpha} \rangle$ to the distance between $y_N$ and $Y_0$. The key difference is the notion of distance, specifically interpolation versus extrapolation. King and Zeng (2006) defines datapoints lying within the convex hull as interpolation and those lying outside as extrapolation. The authors claim that extrapolation is generally more hazardous than interpolation, albeit without formal arguments. In contrast, we assert that extrapolation is a theoretically sound condition, provided it is considered in conjunction with the other conditions laid out in Assumptions 1 and 2 and not in a vacuum. We further argue that (21) is more robust than the convex hull metric. To see this, note that if $y_N$ lies within the convex hull of $Y_0$, then for every $i \leq N_0$, $y_i = [Y_{it} : t \leq T_0]$ (the $i$th row of $Y_0$) necessarily falls outside the convex hull of $\{y_N, y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_{N_0}\}$. Put another way, if the $i$th and $N$th units swapped roles (i.e., unit $i$ receives treatment and unit $N$ remains under control), then the convex hull test would pass in one case and fail in the other, even though the same data is considered in both tests. This highlights the convex hull metric’s sensitivity to the permutation of in-and out-of-sample datapoints.

**Comparison with the mixed confidence interval.** We consider $(\tau_{hz}, \tau_{vt})$ relative to $\tilde{v}_0$. Although $(\tau_{hz}, \tau_{vt})$ can serve as useful metrics for regression analyses, we argue that $\tilde{v}_0$ is generally more favorable for two primary reasons: (i) while $\tilde{v}_0$ couples the HZ and VT errors into one uncertainty estimator, $(\tau_{hz}, \tau_{vt})$ decouples the two errors into two separate stages thereby maintaining the separation of HZ and VT regressions; (ii) the utility of $(\tau_{hz}, \tau_{vt})$ relies on a subjective threshold while the parameter associated with the nominal coverage for $\tilde{v}_0$ is well-defined. For these reasons, we advocate for the usage of $\tilde{v}_0$ over $(\tau_{hz}, \tau_{vt})$.  

31
B Simulations & Empirical Illustrations: Part II

In this section, we provide details on the Basque Country and West Germany studies that were absent in the main article. We also revisit another canonical case study from the synthetic controls literature: California’s Proposition 99 (Abadie et al., 2010).

B.1 Case Study Backgrounds

We provide detailed backgrounds for our three case studies of interest.

**Terrorism in Basque Country.** The purpose of this study is to examine the impact of terrorism on Basque Country’s economy. This dataset consists of $N = 17$ Spanish regions, including Basque Country, whose per capita GDPs are collected from 1955–1998 yielding $T = 43$ years. Basque Country is our single treated region, which we will label as the $N$th region; the remaining regions act as our control regions. The first wave of terrorist activity in 1970 represents the treatment. Thus, the pre- and post-intervention durations are 15 and 28 years, respectively. Our outcome variable of interest is the per capita GDP. In our attempt to isolate the economic cost of terrorism on Basque Country, we set out to estimate its counterfactual GDP growth in the absence of terrorism, i.e., $Y_{Nt}(0)$ for $t \in [1970, 1998]$. Notice that our causal estimand is no longer a single point, but rather a time series of 28 years. We will apply the estimators provided in Section 3 to estimate the counterfactual GDP in each year with the following inputs at our disposal: Basque Country’s pre-intervention GDP trajectory, $y_N \in \mathbb{R}^{15}$; the control regions’ pre-intervention GDP trajectories, $Y_0 \in \mathbb{R}^{16 \times 15}$; and the control regions’ GDPs during the post-intervention year $t$, $y_t \in \mathbb{R}^{16}$ for $t \in [1970, 1998]$.

**West Germany Reunification.** We examine the economic impact of the 1990 reunification in West Germany. This dataset comprises of $N = 17$ countries and their recorded per capita GDPs from 1960–2003, i.e., $T = 44$ years. West Germany represents our single treated country $N$ and the remaining 16 countries serve as our control countries. The reunification represents the treatment, which set the pre- and post-intervention lengths as 30 and 14 years, respectively. The per capita GDP is the primary outcome variable and our goal is to estimate West Germany’s counterfactual GDP trajectory in the absence of reunification, i.e., $Y_{Nt}(0)$ for $t \in [1990, 2003]$. We apply the estimators of Section 3 using the following inputs: West Germany’s pre-intervention GDP trajectory, $y_N \in \mathbb{R}^{30}$; the control countries’ pre-intervention GDP trajectories, $Y_0 \in \mathbb{R}^{16 \times 30}$; and the control countries’ GDPs during the post-intervention year $t$, $y_t \in \mathbb{R}^{16}$ for $t \in [1990, 2003]$.

**CA Proposition 99.** In this study, we examine the effect of California’s Proposition 99, the first modern-time large scale anti-tobacco legislation in the U.S., on its tobacco consumption. Per capita cigarette sales across $N = 39$ U.S. states from 1970–2000, i.e., $T = 31$ years, is collected in this dataset. We consider California as our single treated state, which we label as the $N$th state. The remaining 38 states act as our control states as they neither instituted a tobacco control program nor raised cigarette sales taxes by 50 cents or more. The introduction of Proposition 99 in 1988 represents the treatment. Thus, the pre- and post-treatment lengths are 18 and 13 years, respectively. Our primary outcome variable is the per capita cigarette sales. Our object of interest is the counterfactual time series of sales in the absence of Proposition 99, i.e., $Y_{Nt}(0)$ for $t \in [1988, 2000]$. Below, we apply the estimators of Section 3 using the following inputs: California’s pretreatment sales trajectory, $y_N \in \mathbb{R}^{18}$; the control states’ pretreatment sales trajectories, $Y_0 \in \mathbb{R}^{38 \times 18}$; and the control states’ sales during the post-treatment year $t$, $y_t \in \mathbb{R}^{38}$ for $t \in [1988, 2000]$.
B.2 Implementation Details

We discuss the implementation details of our studies. For ridge, lasso, and elastic net regressions, we use the default scikit-learn hyperparameters ($\lambda_1, \lambda_2$). For PCR, we choose the number of principal components $k$ via a visual heuristic. Specifically, we plot the singular value spectra of $Y_0$, as shown in Figure 6, and define $k$ as the number of singular values whose magnitudes (represented by cross hashes) are above the horizontal red line. This yields $k = 2$ for the Basque study and $k = 4$ for the West Germany and California studies. Finally, we implement simplex regression using the code made available at https://matheusfacure.github.io/python-causality-handbook/15-Synthetic-Control.html.

B.3 Data-Inspired Simulation Studies

The purpose of this section is to continue assessing the relative performances between the mixed intervals and the HZ and VT specific intervals. Below, we follow the outline described in Section 4.3.1.

Data generating process. We follow the DGP described in Section 4.3.1.

Simulation results. We conduct 500 replications of the DGP for each of the studies.

I: Prediction errors. We consider four prediction errors: (i) HZ in-sample error, (ii) VT in-sample error, (iii) bias, and (iv) root-mean-squared-error (RMSE). Formally, we define the HZ and VT in-sample errors as

$$\text{HZ error} = \frac{1}{m} \sum_{\ell=1}^{m} \frac{\|H^{(w)} y_{T_0}^{(\ell)}\|_2}{\|y_{T_0}^{(\ell)}\|_2} \quad \text{and} \quad \text{VT error} = \frac{1}{m} \sum_{\ell=1}^{m} \frac{\|H^{(v)} y_N^{(\ell)}\|_2}{\|y_N^{(\ell)}\|_2},$$

where $(y_{T_0}^{(\ell)}, y_N^{(\ell)})$ are the $\ell$th independent draws and $m$ denotes the number of replications. We define the out-of-sample errors, i.e., average bias and RMSE, as

$$\text{bias} = \frac{1}{m} \sum_{\ell=1}^{m} \frac{\text{prediction}_{N_{T_0}}^{(\ell)}(0) - \mu_0}{|\mu_0|} \quad \text{and} \quad \text{RMSE} = \left[ \frac{\sum_{\ell=1}^{m} (\text{prediction}_{N_{T_0}}^{(\ell)}(0) - \mu_0)^2}{\sum_{\ell=1}^{m} \mu_0^2} \right]^{1/2},$$

where $\text{prediction}_{N_{T_0}}^{(\ell)}(0)$ is the prediction for the $\ell$th replication. Figure 7 visualizes the distribution of biases for OLS and PCR. For ease of comparison, we summarize the average prediction errors in Table 3 across all studies. Similar to the conclusion drawn from Table 1, we find that the true outcome model is approximately linear.
Figure 7: Distribution of biases across the three case studies. The blue and green distributions correspond to OLS and PCR, respectively. The distributions are visually masked in 7a due to the large bias gap between OLS and PCR.

Table 3: Average prediction errors for OLS and PCR across 500 replications.

| Case study | OLS | PCR |
|------------|-----|-----|
|        | HZ error | VT error | Bias | RMSE | HZ error | VT error | Bias | RMSE |
| Basque | 0.08 | 0.00 | −3.30 | 261 | 0.22 | 0.16 | 0.01 | 0.15 |
| W. Germany | 0.00 | 0.01 | −0.01 | 0.01 | 0.01 | 0.01 | −0.01 | 0.01 |
| California | 0.04 | 0.00 | 0.00 | 0.01 | 0.05 | 0.03 | 0.01 | 0.03 |

II: Coverage. We compute 95% confidence intervals for the various uncertainty estimators of interest. Again, for ease of comparison, we present the coverage probabilities (CP) and average coverage lengths (AL) across studies in Table 4. Formally, we define AL as

$$AL = \frac{1}{m} \sum_{t=1}^{m} \left(2 \cdot 1.96 \sqrt{\frac{\hat{\nu}(t)}{\|Y(\nu(t)|N(0)|}} \right),$$

where $\hat{\nu}(t)$ is the variance estimate for the $t$th simulation repeat. This is an overloaded notation as $\hat{\nu}(t)$ depends on the regression type, e.g., HZ, and covariance estimator, e.g., jackknife. We do not report HRK coverage results as the estimators are not well-defined as per the nonsingularity conditions discussed in Section 4.1.3. The coverage results for the California study are in line with our observations for the other two studies. Therefore, we conclude that the HZ and VT intervals generally undercover while the mixed coverages are closer to the nominal coverage.

Table 4: Coverage results for OLS and PCR nominal 95% confidence intervals across 500 simulation replications.

| Case study | OLS Homoskedastic | OLS Jackknife | PCR Homoskedastic | PCR Jackknife |
|------------|-------------------|---------------|-------------------|---------------|
|          | HZ VT Mixed | HZ VT Mixed | HZ VT Mixed | Mod. | HZ VT Mixed | Mod. | HZ VT Mixed | Mod. |
| Basque (CP) | 0.81 0.00 0.81 | 0.99 0.00 0.99 | 0.80 0.84 0.92 | 0.95 | 0.82 0.85 0.80 | 0.97 | 0.82 0.85 0.80 | 0.97 |
| Basque (AL) | 14.2 0.00 14.2 | 211 0.00 211 | 0.39 0.45 0.54 | 0.56 | 0.57 0.54 0.62 | 0.75 | 0.57 0.54 0.62 | 0.75 |
| W. Ger. (CP) | 0.00 0.90 0.90 | 0.00 0.98 0.98 | 0.65 0.94 0.97 | 0.97 | 0.74 0.86 0.96 | 0.96 | 0.74 0.86 0.96 | 0.96 |
| W. Ger. (AL) | 0.00 0.04 0.04 | 0.00 0.08 0.08 | 0.02 0.03 0.03 | 0.03 | 0.02 0.03 0.04 | 0.04 | 0.02 0.03 0.04 | 0.04 |
| California (CP) | 0.90 0.00 0.90 | 0.99 0.00 0.99 | 0.70 0.94 0.96 | 0.96 | 0.76 0.90 0.96 | 0.96 | 0.76 0.90 0.96 | 0.96 |
| California (AL) | 0.28 0.00 0.28 | 0.54 0.00 0.54 | 0.06 0.12 0.13 | 0.13 | 0.07 0.12 0.14 | 0.14 | 0.07 0.12 0.14 | 0.14 |
B.4 Empirical Illustrations

We present results for the Basque Country and West Germany studies that are absent in the body of the article. We also analyze the California Proposition 99 study.

**Terrorism in Basque Country.**

I: Estimation. We display the counterfactual predictions in Figure 8. We find that the symmetric regressions arrive at the same prediction values from both HZ and VT approaches. The asymmetric regressions, on the other hand, produce different HZ and VT counterfactual trajectories. The OLS estimate is wildly different from the other estimates while the HZ simplex regression puts all mass on the latest pre-intervention time period, which reduces it to the last observation carried forward (LOCF) estimator.

![Figure 8: The onset of terrorism is given by the dotted vertical gray line. Basque Country’s observed GDP are shown in black. The HZ and VT counterfactual trajectories are given by solid and dashed-dotted lines, respectively.](image)

II: Inference. The HZ and VT residuals are provided in Figure 9. We find that \( Y_0 \) has full column rank, i.e., \( R = \text{tr}(Y_0 Y_0^\dagger) = 15 \). It is no surprise then that the VT OLS errors are zero, while the HZ OLS errors are nonzero though still small. The HZ and VT PCR in-sample errors are both nonzero. We present the OLS and PCR predictions, along with their corresponding homoskedastic and jackknife confidence intervals, in Figure 4. We do not apply the HRK estimator as the non-invertibility conditions failed to hold.

![Figure 9: (Left) HZ OLS and HZ PCR residuals. The vertical line partitions the control regions and Basque. The residuals to the left of the vertical line represent the in-sample errors across control regions and post-treatment periods. The residuals to the right of the vertical line represent the treatment effects for Basque, marked by the red label, over the post-treatment period. (Bottom) VT OLS and VT PCR residuals. The vertical line partitions the pre- and post-treatment periods with the latter labeled in red. The residuals to the left of the vertical line represent Basque’s in-sample errors over the pretreatment period. The residuals to the right display the same treatment effects over time for Basque as shown in the left figure.](image)
West Germany reunification.

I: Estimation. We present the counterfactual estimates in Figure 10. Our empirical findings reinforce our results in Section 3 with the symmetric regressions producing the same HZ and VT predictions and the asymmetric regressions producing different HZ and VT predictions, even with the same penalty values. Although the hyperparameters are not tuned, the results across all regressions are qualitatively similar with the exception of HZ simplex regression, which again reduces to LOCF. In fact, the OLS and ridge estimates appear to overlap, as with the lasso and elastic net estimates.

![Symmetric regressions.](image)

![Asymmetric regressions.](image)

Figure 10: The reunification is given by the dotted vertical gray line. West Germany’s observed GDP is shown in black. The HZ and VT counterfactual trajectories are given by solid and dashed-dotted lines, respectively.

II: Inference. We display the OLS and PCR HZ and VT residuals in Figure 11. We find that $Y_0$ has full row rank with $R = \text{tr}(Y_0^*Y_0^\dagger) = 16$. Thus, the HZ OLS in-sample errors are zero while the VT OLS in-sample errors are small but nonzero. The HZ PCR errors are roughly distributed between $[-4000, 4000]$ and the VT PCR errors are much smaller in magnitude. For OLS, we apply the HZ, VT, and mixed homoskedastic and jackknife variance estimators since the conditions for the HRK variance estimator are not satisfied. The resulting intervals are shown in Figures 4g–4l. The OLS HZ intervals are degenerate. The OLS VT and mixed intervals, on the other hand, expand over time. In general, the jackknife intervals engulf the homoskedastic intervals. For PCR, we observe that the HZ confidence intervals are tightly centered around the prediction values but the VT confidence intervals also increase over time.

![HZ residuals.](image)

![VT residuals.](image)

Figure 11: (Left) HZ OLS and HZ PCR residuals. The vertical line partitions the control countries and West Germany. The residuals to the left of the vertical line represent the in-sample errors across control units and post-treatment periods. The residuals to the right of the vertical line represent the treatment effects for West Germany, marked by the red label, over the post-treatment period. (Bottom) VT OLS and VT PCR residuals. The vertical line partitions the pre- and post-treatment periods with the latter labeled in red. The residuals to the left of the vertical line represent West Germany’s in-sample errors over the pretreatment period. The residuals to the right display the same treatment effects over time for West Germany as shown in the left figure.
CA Proposition 99.

I: Estimation. Our counterfactual predictions are shown in Figure 12. Our results provide further evidence of the formal statements in Section 3. Like the West Germany study, the results across all regressions are qualitatively similar with the exception of HZ simplex regression, which once more reduces to a LOCF estimator. In particular, we observe that OLS and ridge overlap again, similar to lasso and elastic net.

![Figure 12](image)

**Figure 12:** The passage of Prop 99 is given by the dotted vertical gray line. California’s observed sales are shown in black. The HZ and VT counterfactual trajectories are given by solid and dashed-dotted lines, respectively.

II: Inference. We display the HZ and VT residuals in Figure 13. We find that $Y_0$ has rank $R = \text{tr}(Y_0Y_0^\dagger) = 17$. The VT OLS in-sample errors are zero while the HZ OLS in-sample errors are distributed between $[-20, 20]$. The HZ and VT PCR in-sample errors are distributed between $[-35, 35]$ and $[-5, 5]$, respectively. For OLS and PCR, we apply the homoskedastic and jackknife intervals since the conditions for the HRK interval are not satisfied. The resulting intervals are shown in Figure 14 and 15. As expected, the VT OLS confidence intervals are degenerate. In contrast, the HZ and mixed confidence intervals expand over time. The jackknife intervals are generally larger than the homoskedastic intervals, which is expected given that the jackknife is conservative. Further, the VT PCR confidence intervals are tightly centered around the prediction values but are nontrivial. In all other aspects, the PCR confidence intervals are similar to that of OLS.

![Figure 13](image)

**Figure 13:** (Left) HZ OLS and HZ PCR residuals. The vertical line partitions the control states and California. The residuals to the left of the vertical line represent the in-sample errors across control states and post-treatment periods. The residuals to the right of the vertical line represent the treatment effects for California, marked by the red label, over the post-treatment period. (Bottom) VT OLS and VT PCR residuals. The vertical line partitions the pre- and post-treatment periods with the latter labeled in red. The residuals to the left of the vertical line represent California’s in-sample errors over the pretreatment period. The residuals to the right display the same treatment effects over time for California as shown in the left figure.
Figure 14: OLS predictions and confidence intervals. We index rows by the covariance estimator, e.g., jackknife, and the columns by the type of confidence interval, e.g., HZ.

Figure 15: PCR predictions and confidence intervals. We index rows by the covariance estimator, e.g., jackknife, and the columns by the type of confidence interval, e.g., HZ.
Common themes. We discuss a few common themes from our empirical illustrations. First, with respect to estimation, our findings support our classification results in Section 3. Notably, we observe that the simplex constraint induces the greatest asymmetry between the $HZ$ and $VT$ predictions. In every study, including the Basque study presented in the main body, $HZ$ simplex regression reduces to the LOCF predictor. This suggests that the simplex constraint imposes a strong systematic structure on the underlying DGP. Finally, with respect to inference, our results provide further evidence in support of the mixed interval by showcasing its robustness relative to the $HZ$ and $VT$ intervals.

C Proofs for Finite-Sample Point Estimation

Helper lemmas. To establish Theorems 1 and 2, we first state the following useful lemmas for the collection of regression formulations presented in Section 3. We provide their proofs in Appendix C.3.

Lemma 5 (OLS). $HZ = VT$ for OLS with $\hat{\alpha}$ and $\hat{\beta}$ as the minimum $\ell_2$-norm solutions, i.e.,

$$\hat{Y}_{NT}^{hz}(0) = \hat{Y}_{NT}^{vt}(0) = \langle y_N, Y_0^\dagger y_T \rangle = \sum_{\ell=1}^{R} (1/s_\ell) \langle y_N, v_\ell \rangle \langle u_\ell, y_T \rangle.$$

Lemma 6 (PCR). $HZ = VT$ for PCR with $\hat{\alpha}$ and $\hat{\beta}$ as the minimum $\ell_2$-norm solutions, i.e.,

$$\hat{Y}_{NT}^{hz}(0) = \hat{Y}_{NT}^{vt}(0) = \langle y_N, (Y_0^{(k)})^\dagger y_T \rangle = \sum_{\ell=1}^{k} (1/s_\ell) \langle y_N, v_\ell \rangle \langle u_\ell, y_T \rangle.$$

Lemma 7 (Ridge). $HZ = VT$ for ridge regression, i.e.,

$$\hat{Y}_{NT}^{hz}(0) = \hat{Y}_{NT}^{vt}(0) = \langle y_N, (Y_0^\dagger Y_0 + \lambda_2 I)^{-1} Y_0^\dagger y_T \rangle = \sum_{\ell=1}^{R} \frac{s_\ell}{s_\ell^2 + \lambda_2} \langle y_N, v_\ell \rangle \langle u_\ell, y_T \rangle.$$

Lemma 8 (Lasso). $HZ \neq VT$ for lasso regression.

Lemma 9 (Elastic net). $HZ \neq VT$ for elastic net regression.

Lemma 10 (Simplex regression). $HZ \neq VT$ for simplex regression.

C.1 Proof of Theorem 1

Proof. The proof is immediate from Lemmas 5–7.

C.2 Proof of Theorem 2

Proof. The proof is immediate from Lemmas 8–10.
C.3 Proofs of Finite-Sample Estimation Lemmas

C.3.1 Proof of Lemma 5: OLS

Proof. Consider HZ regression. By the optimality conditions,
\[ \nabla_\alpha \| y_T - Y_0 \alpha \|^2 = 0. \]
Solving for \( \alpha \), we derive the well-known “normal equations”
\[ Y_0' Y_0 \alpha = Y_0' y_T. \]
Using the pseudoinverse, we obtain
\[ \hat{\alpha} = (Y_0' Y_0)^+ Y_0' y_T = Y_0^+ y_T. \]
Observe that this corresponds to the unique minimum \( \ell_2 \)-norm solution that lies within the rowspace of \( Y_0 \). Therefore, the HZ prediction is given by
\[ \hat{Y}_{NT}^{hz}(0) = \langle y_N, \hat{\alpha} \rangle = \langle y_N, Y_0^+ y_T \rangle. \]
Following the arguments above for VT regression, it follows that
\[ \hat{\beta} = (Y_0' Y_0')^+ Y_0 y_N = (Y_0')^T y_N, \]
which corresponds to the unique minimum \( \ell_2 \)-norm solution that lies within the columnspace of \( Y_0 \). Therefore,
\[ \hat{Y}_{NT}^{vt}(0) = \langle y_T, \hat{\beta} \rangle = \langle y_T, (Y_0')^T y_N \rangle. \]
Given that \( (Y_0')^T = (Y_0')' \), we conclude
\[ \hat{Y}_{NT}^{hz}(0) = \langle y_N, Y_0^+ y_T \rangle = \langle y_T, (Y_0')^T y_N \rangle = \hat{Y}_{NT}^{vt}(0). \]

\[ \square \]

C.3.2 Proof of Lemma 6: PCR

Proof. Consider HZ regression with any \( k < R \). Let \( U_k \in \mathbb{R}^{N_0 \times k} \) and \( V_k \in \mathbb{R}^{T_0 \times k} \) denote the matrices formed by the top \( k \) left and right singular vectors, respectively, and \( S_k \in \mathbb{R}^{k \times k} \) denote the matrix of top \( k \) singular values. Observe that
\[ \left( (Y_0^{(k)})' Y_0^{(k)} \right)^+ (Y_0^{(k)})' = V_k S_k^{-2} V_k' V_k S_k U_k' = V_k S_k^{-1} U_k' = (Y_0^{(k)})'. \]
Therefore, \( \hat{\alpha} = (Y_0^{(k)})^\dagger y_T \), which corresponds to the unique minimum \( \ell_2 \)-norm solution that lies within the rowspace of \( Y_0^{(k)} \). Following the proof of Lemma 5, we conclude that

\[
\hat{Y}_{NT}^{hz}(0) = \langle y_N, \hat{\alpha} \rangle = \langle y_N, (Y_0^{(k)})^\dagger y_T \rangle.
\]

Similarly, for VT regression, we note that

\[
\left( Y_0^{(k)} (Y_0^{(k)})^\dagger \right)^\dagger Y_0^{(k)} = (U_k S_k^{-2} U_k^T) U_k S_k V_k^T = U_k S_k^{-1} V_k = ((Y_0^{(k)})^\dagger)^\dagger.
\]

In turn, we have \( \hat{\beta} = ((Y_0^{(k)})^\dagger)^\dagger y_N \), which corresponds to the unique minimum \( \ell_2 \)-norm solution that lies within the columnspace of \( Y_0^{(k)} \). Moreover,

\[
\hat{Y}_{NT}^{vt}(0) = \langle y_T, \hat{\beta} \rangle = \langle y_T, ((Y_0^{(k)})^\dagger)^\dagger y_N \rangle.
\]

We finish by establishing

\[
\hat{Y}_{NT}^{hz}(0) = \langle y_N, (Y_0^{(k)})^\dagger y_T \rangle = \langle y_T, ((Y_0^{(k)})^\dagger)^\dagger y_N \rangle = \hat{Y}_{NT}^{vt}(0).
\]

\(\square\)

**Helper lemma.** To establish Lemmas 7–9, we first establish a general result in Lemma 11 for \( \ell_p \)-penalties, where \( p = 2/K \) and \( K \) is an integer \( \geq 1 \), based on the contributions of Hoff (2017). More formally, consider

(a) HZ regression: for \( K \geq 1 \) and \( \lambda > 0 \),

\[
\hat{\alpha} = \arg\min_{\alpha} \|y_T - Y_0 \alpha\|_2^2 + \lambda \|\alpha\|_p^p
\]

\[
\hat{Y}_{NT}^{hz}(0) = \langle y_N, \hat{\alpha} \rangle.
\]

(b) VT regression: for \( K \geq 1 \) and \( \lambda > 0 \),

\[
\hat{\beta} = \arg\min_{\beta} \|y_N - Y_0^\dagger \beta\|_2^2 + \lambda \|\beta\|_p^p
\]

\[
\hat{Y}_{NT}^{vt}(0) = \langle y_T, \hat{\beta} \rangle.
\]

We remark that \( K = 1 \) and \( K = 2 \) yield ridge and lasso regression, respectively, while \( K > 2 \) yields non-convex penalties. We relegate the proof of Lemma 11 to Appendix C.3.7.

**Lemma 11.** For any \( K \geq 1 \) and \( \lambda > 0 \), a HZ and VT regression solution is given by

\[
\hat{Y}_{NT}^{hz}(0) = \langle y_N, \hat{\alpha}_1 \circ \ldots \circ \hat{\alpha}_K \rangle
\]

\[
\hat{Y}_{NT}^{vt}(0) = \langle y_T, \hat{\beta}_1 \circ \ldots \circ \hat{\beta}_K \rangle,
\]
where for every $k \leq K$,

\[
\hat{\alpha}_k = \left( D(\widehat{\alpha}_{\sim k})Y_0'Y_0D(\widehat{\alpha}_{\sim k}) + \frac{\lambda}{K} I \right)^{-1} D(\widehat{\alpha}_{\sim k})Y_0'y_T,
\]

\[
\hat{\beta}_k = \left( D(\widehat{\beta}_{\sim k})Y_0'Y_0D(\widehat{\beta}_{\sim k}) + \frac{\lambda}{K} I \right)^{-1} D(\widehat{\beta}_{\sim k})Y_0'y_N,
\]

$\hat{\alpha}_{\sim k} = \hat{\alpha}_1 \circ \ldots \circ \hat{\alpha}_{k-1} \circ \hat{\alpha}_{k+1} \circ \ldots \circ \hat{\alpha}_K$, $\hat{\beta}_{\sim k} = \hat{\beta}_1 \circ \ldots \circ \hat{\beta}_{k-1} \circ \hat{\beta}_{k+1} \circ \ldots \circ \hat{\beta}_K$, and $D(\hat{\alpha}_{\sim k})$ and $D(\hat{\beta}_{\sim k})$ are diagonal matrices formed from $\hat{\alpha}_{\sim k}$ and $\hat{\beta}_{\sim k}$, respectively.

C.3.3 Proof of Lemma 7: Ridge Regression

*Proof.* By Lemma 11 for $K = 1$ and $\lambda = \lambda_2 > 0$, the HZ regression solution is given by

\[
\hat{Y}_{NT}^{hz}(0) = \langle y_N, (Y_0'Y_0 + \lambda_2 I)^{-1}Y_0'y_T \rangle.
\]

Similarly, the VT regression solution is given by

\[
\hat{Y}_{NT}^{vt}(0) = \langle y_T, (Y_0'Y_0 + \lambda_2 I)^{-1}Y_0'y_N \rangle.
\]

Since $(Y_0'Y_0 + \lambda I)^{-1}Y_0' = Y_0'(Y_0'Y_0' + \lambda I)^{-1}$, it follows that $\hat{Y}_{NT}^{hz}(0) = \hat{Y}_{NT}^{vt}(0)$. \qed

C.3.4 Proof of Lemma 8: Lasso Regression

*Proof.* By Lemma 11 for $K = 2$ and $\lambda = \lambda_1 > 0$, a HZ regression solution solution is given by

\[
\hat{Y}_{NT}^{hz}(0) = \langle y_N, \hat{\alpha}_1 \circ \hat{\alpha}_2 \rangle,
\]

(24)

where

\[
\hat{\alpha}_{1+k} = \left( D(\widehat{\alpha}_{2-k})Y_0'Y_0D(\widehat{\alpha}_{2-k}) + \frac{\lambda_1}{2} I \right)^{-1} D(\widehat{\alpha}_{2-k})Y_0'y_T
\]

for $k = \{0, 1\}$. Similarly, a VT regression solution is given by

\[
\hat{Y}_{NT}^{vt}(0) = \langle y_T, \hat{\beta}_1 \circ \hat{\beta}_2 \rangle,
\]

(25)

where

\[
\hat{\beta}_{1+k} = \left( D(\widehat{\beta}_{2-k})Y_0'Y_0D(\widehat{\beta}_{2-k}) + \frac{\lambda_1}{2} I \right)^{-1} D(\widehat{\beta}_{2-k})Y_0'y_N
\]

for $k = \{0, 1\}$. Leveraging (24) and (25), we find that the HZ regression solution can be linear in $y$ and at least quadratic in $q$. On the other hand, the VT regression solution can be linear in $q$ and at least quadratic in $y$. Since the lasso solution is unique under the assumption the entries of $Y_0$ are drawn from a continuous distribution, this implies that HZ regression and VT regression do not yield matching solutions in general. \qed
C.3.5 Proof of Lemma 9: Elastic Net Regression

Proof. Consider HZ regression. We rewrite (1) with $\lambda_1, \lambda_2 > 0$ in a lasso formulation:

$$\hat{\alpha}^* = \arg\min_{\alpha} \|y_T - Y_0^* \alpha^*\|_2^2 + \lambda^* \|\alpha^*\|_1,$$

(26)

where

$$q^* = \begin{pmatrix} y_T \\ 0 \end{pmatrix}, \quad Y_0^* = \frac{1}{\sqrt{1 + \lambda_2}} \begin{pmatrix} Y_0 \\ \sqrt{\lambda_2} I \end{pmatrix}, \quad \lambda^* = \frac{\lambda_1}{\sqrt{1 + \lambda_2}}, \quad \alpha^* = (\sqrt{1 + \lambda_2}) \alpha.$$

We apply Lemma 11 to (26) with $K = 2$ and $\lambda = \lambda^* > 0$ to obtain

$$\hat{\gamma}_{NT}^{hz}(0) = \langle y_N^T, \hat{\alpha}_1^* \circ \hat{\alpha}_2^* \rangle \sqrt{1 + \lambda_2},$$

(27)

where

$$\hat{\alpha}_{1+k}^* = \left( \frac{1}{\sqrt{1 + \lambda_2}} D(\hat{\alpha}_{2-k}^*) (Y_0' Y_0 + \lambda_2 I) D(\hat{\alpha}_{2-k}^*) + \frac{\lambda_1}{2} I \right)^{-1} D(\hat{\alpha}_{2-k}^*) Y_0^T y_T$$

for $k \in \{0, 1\}$. Similarly, for VT regression, we proceed as above to obtain

$$\hat{\gamma}_{NT}^{vt}(0) = \langle y_N^T, \hat{\beta}_1^* \circ \hat{\beta}_2^* \rangle \sqrt{1 + \lambda_2},$$

(28)

where

$$\hat{\beta}_{1+k}^* = \left( \frac{1}{\sqrt{1 + \lambda_2}} D(\hat{\beta}_{2-k}^*) (Y_0 Y_0' + \lambda_2 I) D(\hat{\beta}_{2-k}^*) + \frac{\lambda_1}{2} I \right)^{-1} D(\hat{\beta}_{2-k}^*) Y_0' y_N$$

for $k \in \{0, 1\}$. Leveraging (27) and (28), we find that the HZ regression solution can be linear in $y$ and at least quadratic in $q$. On the other hand, the VT regression solution can be linear in $q$ and at least quadratic in $y$. Since the elastic net regression solution is unique, provided $\lambda_2 > 0$, this implies that HZ regression and VT regression do not yield matching solutions in general. \hfill \square

C.3.6 Proof of Lemma 10: Simplex Regression

Proof. Consider HZ regression. We write the Lagrangian of (4) as

$$\hat{\alpha} = \arg\min_{\alpha} \|y_T - Y_0 \alpha\|_2^2 + \lambda\|\alpha\|_2^2 - (\theta^{hz})' \alpha + \nu^{hz}(1' \alpha - 1),$$

for $k \in \{0, 1\}$. Leveraging (27) and (28), we find that the HZ regression solution can be linear in $y$ and at least quadratic in $q$. On the other hand, the VT regression solution can be linear in $q$ and at least quadratic in $y$. Since the elastic net regression solution is unique, provided $\lambda_2 > 0$, this implies that HZ regression and VT regression do not yield matching solutions in general. \hfill \square
where \( \theta^{hz} \in \mathbb{R}^{T_0} \) and \( \nu^{hz} \in \mathbb{R} \). By the Karush-Kuhn-Tucker (KKT) conditions, optimality is achieved if the following are satisfied:

\[
\begin{align*}
\hat{\alpha} & \succeq 0, \quad I' \hat{\alpha} = 1, \\
\hat{\theta}^{hz} & \succeq 0, \\
\hat{\theta}^{hz}_i \hat{\alpha}_i & = 0 \text{ for } i = 1, \ldots, T_0, \\
\hat{\alpha} & = (Y_0'Y_0 + \lambda I)^{-1} \left( Y_0'y_T + \frac{1}{2} \hat{\theta}^{hz} - \frac{\hat{\nu}^{hz}}{2}1 \right).
\end{align*}
\]

Therefore, given primal and dual feasible variables \((\hat{\alpha}, \hat{\theta}^{hz}, \hat{\nu}^{hz})\), we can write the final HZ prediction as

\[
\hat{Y}_{NT}^{hz}(0) = \hat{Y}_{NT}^{hz, ols}(0) + (1/2)y_N'(Y_0'Y_0 + \lambda I)^{-1}(\hat{\theta}^{vt} - \hat{\nu}^{vt}1),
\]

where \( \hat{Y}_{NT}^{hz, ols}(0) = y_N'(Y_0'Y_0 + \lambda I)^{-1}Y_0'\nu_T \) converges to the prediction corresponding to the OLS solution with minimum \( \ell_2 \)-norm as \( \lambda \to 0^+ \). Similarly, for VT regression, the KKT conditions are

\[
\begin{align*}
\hat{\beta} & \succeq 0, \quad I' \hat{\beta} = 1, \\
\hat{\nu}^{vt} & \succeq 0, \\
\hat{\nu}^{vt}_i \hat{\beta}_i & = 0 \text{ for } i = 1, \ldots, N_0, \\
\hat{\beta} & = (Y_0'Y_0' + \lambda I)^{-1} \left( Y_0'y_N + \frac{1}{2} \hat{\nu}^{vt} - \frac{\hat{\theta}^{vt}}{2}1 \right).
\end{align*}
\]

For primal and dual feasible variables \((\hat{\beta}, \hat{\nu}^{vt}, \hat{\theta}^{vt})\), this yields

\[
\hat{Y}_{NT}^{vt}(0) = \hat{Y}_{NT}^{vt, ols}(0) + (1/2)y_T'(Y_0'Y_0' + \lambda I)^{-1}(\hat{\theta}^{vt} - \hat{\nu}^{vt}1),
\]

where \( \hat{Y}_{NT}^{vt, ols}(0) = y_T'(Y_0'Y_0' + \lambda I)^{-1}Y_0'\nu_T \) converges to the prediction corresponding to the OLS solution with minimum \( \ell_2 \)-norm as \( \lambda \to 0^+ \). Notably, as per Theorem 1, \( \hat{Y}_{NT}^{hz, ols}(0) = \hat{Y}_{NT}^{vt, ols}(0) = \hat{Y}_{NT}^{ols}(0) \) for any \( \lambda \geq 0 \). As a result,

\[
\begin{align*}
\hat{Y}_{NT}^{hz}(0) & = \hat{Y}_{NT}^{ols}(0) + (1/2)y_N'(Y_0'Y_0 + \lambda I)^{-1}(\hat{\theta}^{hz} - \hat{\nu}^{hz}1) \quad (29) \\
\hat{Y}_{NT}^{vt}(0) & = \hat{Y}_{NT}^{ols}(0) + (1/2)y_T'(Y_0'Y_0' + \lambda I)^{-1}(\hat{\theta}^{vt} - \hat{\nu}^{vt}1). \quad (30)
\end{align*}
\]

As seen from (29) and (30), the leading terms in the HZ and VT simplex regression predictions are identical. The remaining terms, however, can differ from one another. As an example, consider \( N = T \) with

\[
Y_0 = I, \quad y_N = 0, \quad y_T = (1 + \lambda)(\hat{\theta}^{vt} - \hat{\nu}^{vt}1). \quad (31)
\]

By construction, observe that

\[
\hat{\beta} = \frac{1}{2(1 + \lambda)}(\hat{\theta}^{vt} - \hat{\nu}^{vt}1).
\]

Recall from the KKT conditions for VT regression that \( \hat{\beta} \succeq 0 \) and \( I' \hat{\beta} = 1 \). Therefore, at least one entry of \( (\hat{\theta}^{vt} - \hat{\nu}^{vt}1) \) must be strictly positive. This yields

\[
(1 + \lambda)^{-1}y_T'(\hat{\theta}^{vt} - \hat{\nu}^{vt}1) = (\hat{\theta}^{vt} - \hat{\nu}^{vt}1)'(\hat{\theta}^{vt} - \hat{\nu}^{vt}1) > 0. \quad (32)
\]
Plugging (31) and (32) into (29) and (30), we obtain
\[ \hat{\nu}_{NT}(0) = 0 \quad \text{and} \quad \hat{\nu}_{NT}(0) > 0, \]
which concludes our proof.

C.3.7 Proof of Lemma 11: \( \ell_p \)-penalties

Proof. We recall the Hadamard product parametrization (HPP): for any vector \( z \) and integer \( K \geq 1 \),
\[ \|z\|_p = \min_{z_1 \circ \ldots \circ z_K = z} \frac{1}{K} \sum_{k=1}^{K} \|z_k\|_2^2, \]
where \( \circ \) denotes the Hadamard (componentwise) product. We rewrite our subclass of \( \ell_p \)-penalties, i.e., (22) and (23), as sums of \( \ell_2 \)-penalties via the HPP technique:
\[
(\hat{\alpha}_1, \ldots, \hat{\alpha}_K) = \arg\min_{\alpha_1, \ldots, \alpha_K} \|y_T - Y_0(\alpha_1 \circ \ldots \circ \alpha_K)\|_2^2 + \frac{\lambda}{K} \sum_{k=1}^{K} \|\alpha_k\|_2^2 \tag{33}
\]
\[
(\hat{\beta}_1, \ldots, \hat{\beta}_K) = \arg\min_{\beta_1, \ldots, \beta_K} \|y_N - Y_0'(\beta_1 \circ \ldots \circ \beta_K)\|_2^2 + \frac{\lambda}{K} \sum_{k=1}^{K} \|\beta_k\|_2^2, \tag{34}
\]
where \( \hat{\alpha} = \hat{\alpha}_1 \circ \ldots \circ \hat{\alpha}_K \) and \( \hat{\beta} = \hat{\beta}_1 \circ \ldots \circ \hat{\beta}_K \). Below, we leverage the results of Hoff (2017), which provides an alternating ridge regression algorithm to solve for (33) and (34).

Consider HZ regression. Let us solve for \( \alpha_k \) for \( k \in [K] \) by fixing \( \alpha_{k'} \) for \( k' \neq k \). By the optimality conditions,
\[
\nabla_{\alpha_k} \left\{ (\alpha_1 \circ \ldots \circ \alpha_K)'Y_0'Y_0(\alpha_1 \circ \ldots \circ \alpha_K) - 2(\alpha_1 \circ \ldots \circ \alpha_K)'Y_0'y_T + \frac{\lambda}{K} \alpha_k' \alpha_k \right\} = 0. \tag{35}
\]
In order to solve for (35), observe that
\[
(\alpha_1 \circ \ldots \circ \alpha_K)'Y_0'Y_0(\alpha_1 \circ \ldots \circ \alpha_K) = \alpha_k'(Y_0'Y_0 \circ \alpha_{\sim k} \alpha_{\sim k}) \alpha_k
\]
\[
(\alpha_1 \circ \ldots \circ \alpha_K)'Y_0'y_T = \alpha_k'(\alpha_{\sim k} \circ Y_0'y_T),
\]
where \( \alpha_{\sim k} = \alpha_1 \circ \ldots \circ \alpha_{k-1} \circ \alpha_{k+1} \circ \ldots \circ \alpha_K \). In turn, this allows us to rewrite (35) as
\[
\nabla_{\alpha_k} \left\{ \alpha_k' \left( Y_0'Y_0 \circ \alpha_{\sim k} \alpha_{\sim k} + \frac{\lambda}{K} I \right) \alpha_k - 2\alpha_k'(\alpha_{\sim k} \circ Y_0'y_T) \right\} = 0.
\]
This is quadratic in \( \alpha_k \) for fixed \( \alpha_{\sim k} \). Therefore, the unique minimizer at convergence is given by
\[
\hat{\alpha}_k = \left( Y_0'Y_0 \circ \hat{\alpha}_{\sim k} \hat{\alpha}_{\sim k} + \frac{\lambda}{K} I \right)^{-1} (\hat{\alpha}_{\sim k} \circ Y_0'y_T), \tag{36}
\]
where \( \hat{\alpha}_{\sim k} = \hat{\alpha}_1 \circ \ldots \circ \hat{\alpha}_{k-1} \circ \hat{\alpha}_{k+1} \circ \ldots \circ \hat{\alpha}_K \). Leveraging properties of the Hadamard product noted in
Styan (1973), we rewrite

\[ Y'_0 Y_0 \circ \tilde{\alpha}_{\sim k} \tilde{\alpha}'_{\sim k} = D(\tilde{\alpha}_{\sim k}) Y'_0 Y_0 D(\tilde{\alpha}_{\sim k}) \]

\[ Y'_0 y_T \circ \tilde{\alpha}_{\sim k} = D(\tilde{\alpha}_{\sim k}) Y'_0 y_T, \]

where \( D(\tilde{\alpha}_{\sim k}) \) is the diagonal matrix formed from \( \tilde{\alpha}_{\sim k} \). Leveraging these equalities, we simplify (36) as

\[ \tilde{\alpha}_k = \left( D(\tilde{\alpha}_{\sim k}) Y'_0 Y_0 D(\tilde{\alpha}_{\sim k}) + \frac{\lambda}{K} I \right)^{-1} D(\tilde{\alpha}_{\sim k}) Y'_0 y_T. \]

We now turn to VT regression. Following the arguments above, we obtain for every \( k \in [K], \)

\[ \beta_k = \left( D(\tilde{\beta}_{\sim k}) Y_0 Y'_0 D(\tilde{\beta}_{\sim k}) + \frac{\lambda}{K} I \right)^{-1} D(\tilde{\beta}_{\sim k}) Y_0 y_N, \]

where \( \tilde{\beta}_{\sim k} = \tilde{\beta}_1 \circ \ldots \circ \tilde{\beta}_{k-1} \circ \tilde{\beta}_{k+1} \circ \ldots \circ \tilde{\beta}_K \) and \( D(\tilde{\beta}_{\sim k}) \) is the diagonal matrix formed from \( \tilde{\beta}_{\sim k} \). This completes the proof.

\[ \square \]

### D Proofs for Large-sample Inference

We state a useful lemma that will be used throughout our derivations.

**Lemma 12.** Consider a random vector \( x \) and random matrix \( A \). Let \( \mathbb{E}[x|A] = 0 \) and \( \text{Cov}(x|A) = \Sigma \). Then \( \mathbb{E}[x'Ax|A] = \text{tr}(A\Sigma) \).

**D.1 Proof of Theorem 3**

To establish Theorem 3, we state Lyapunov’s central limit theorem (CLT) in Lemma 13.

**Lemma 13 (Theorem 2.7.1 of Lehmann (2000)).** Let \( X_i \) for \( i = 1, \ldots, n \) be independently distributed with means \( \mathbb{E}[X_i] = \zeta_i \) and variances \( \sigma_i^2 \), and with finite third moments. Let \( \bar{X} = (1/n) \sum_{i=1}^n X_i \). Then

\[ \bar{X} - \mathbb{E}[\bar{X}] \quad \xrightarrow{(\text{Var}(\bar{X}))^{1/2}} \mathcal{N}(0,1), \]

provided

\[ \left( \sum_{i=1}^n \mathbb{E}[[X_i - \zeta_i]^3] \right)^2 = o \left( \sum_{i=1}^n \sigma_i^2 \right)^3. \]

**Proof.** Consider HZ regression. Throughout, we condition on \( (y_N, Y_0) \). Given (14), Lemma 13 establishes

\[ \frac{\hat{Y}_{NT}^{hz}(0) - \mathbb{E}[\hat{Y}_{NT}^{hz}(0)|y_N, Y_0]}{\text{Var}(\hat{Y}_{NT}^{hz}(0)|y_N, Y_0)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1). \]
To evaluate \( E[\hat{Y}_{NT}^h(0)|y_N, Y_0] \), we first observe that

\[
E[\hat{Y}_{NT}^h(0)|y_N, Y_0] = E[(y_N, \hat{\alpha})|y_N, Y_0] = E[(y_N, Y_0^\dagger y_T)|y_N, Y_0] = y_N Y_0^\dagger E[y_T|y_N, Y_0].
\]  

By Assumption 1 (i)–(iii),

\[
E[y_T|y_N, Y_0] = E[y_T|Y_0] = Y_0 \alpha^* + E[\varepsilon_T|Y_0] = Y_0 \alpha^*. 
\]

Combining (37) and (38) gives

\[
E[\hat{Y}_{NT}^h(0)|y_N, Y_0] = y_N^\dagger \alpha^* = y_N^\dagger H^{(v)} \alpha^*. 
\]  

Next, we note that \( E[Y_{NT}(0)|y_N] = E[Y_{NT}(0)|y_N, Y_0] \) by Assumption 1 (iii). Coupled with Assumption 1 (i.b), it follows that

\[
E[Y_{NT}(0)|y_N, Y_0] = \langle y_N, \alpha^* \rangle = \langle y_N, H^{(v)} \alpha^* \rangle + \langle y_N, H^{(v)} \alpha^* \rangle.
\]

Notice that \( \langle y_N, H^{(v)} \alpha^* \rangle = 0 \) if \( y_N \in \text{rowspan}(Y_0) \) or \( \alpha^* \in \text{rowspan}(Y_0) \). Thus, under Assumption 1 (iv)

\[
E[\hat{Y}_{NT}^h(0)|y_N, Y_0] = E[Y_{NT}(0)|y_N].
\]

Moving to the variance term, we note that

\[
\text{Var}(\hat{Y}_{NT}^h(0)|y_N, Y_0) = y_N^\dagger \text{Cov}(\hat{\alpha}|y_N, Y_0)y_N.
\]

Towards evaluating the above, we note that

\[
\text{Cov}(\hat{\alpha}|y_N, Y_0) = \text{Cov}(Y_0^\dagger y_T|y_N, Y_0) = \text{Cov}(Y_0^\dagger y_T|Y_0) = Y_0^\dagger \text{Cov}(y_T|Y_0)(Y_0')^\dagger = Y_0^\dagger \Sigma_T (Y_0')^\dagger.
\]

Plugging (41) into (40), we obtain

\[
\text{Var}(\hat{Y}_{NT}^h(0)|y_N, Y_0) = y_N^\dagger \Sigma_T (Y_0')^\dagger y_N = \hat{\beta}^\dagger \Sigma_T \hat{\beta},
\]

where we recall that \( \hat{\beta} = (Y_0')^\dagger y_N \). Putting it all together, we conclude

\[
\frac{\hat{Y}_{NT}^h(0) - E[Y_{NT}(0)|y_N]}{(\hat{\beta}^\dagger \Sigma_T \hat{\beta})^{1/2}} \overset{d}{\rightarrow} \mathcal{N}(0, 1).
\]

Switching gears to VT regression, we now condition on \((y_T, Y_0)\) and operate under the VT regression as-
sumptions. Following the same arguments above, we obtain
\[
\mathbb{E}[\hat{Y}_{NT}^v(0)| y_T, Y_0] = \mathbb{E}[Y_{NT}(0)| y_T]
\]
\[
\text{Var}(\hat{Y}_{NT}^v(0)| y_T, Y_0) = \alpha' \Sigma_N \alpha.
\]
As with HZ regression, (15) with Lemma 13 establishes
\[
\frac{\hat{Y}_{NT}^v(0) - \mathbb{E}[Y_{NT}(0)| y_T]}{(\alpha' \Sigma_N \alpha)^{1/2}} \overset{d}{\to} \mathcal{N}(0, 1).
\]
The proof is complete.

### D.2 Proof of Lemma 1

**Proof.** Consider HZ regression. Taking note that \( H_\perp (u) \parallel Y_0 = 0 \),
\[
\|H_\perp (u) y_T\|_2^2 = y_T' H_\perp (u) y_T
\]
\[
= (Y_0 \alpha + \varepsilon_T)' H_\perp (u) (Y_0 \alpha + \varepsilon_T)
\]
\[
= \varepsilon_T' H_\perp (u) \varepsilon_T.
\]

Applying Lemma 12 then gives
\[
\mathbb{E}[\varepsilon_T' H_\perp (u) \varepsilon_T| Y_0] = \sigma_T^2 \text{tr}(H_\perp (u)) = (N_0 - R)\sigma_T^2,
\]
where the final equality follows from the fact that the trace of a projection matrix equals its rank. Putting everything together, we have \( \mathbb{E}[\sigma_T^2| Y_0] = \sigma_T^2 \) and \( \mathbb{E}[\hat{\Sigma}_{T, \text{homo}}| Y_0] = \Sigma_T \). Therefore,
\[
\mathbb{E}[\hat{\sigma}_{h_0, \text{homo}}| y_N, Y_0] = \hat{\beta}' \mathbb{E}[\hat{\Sigma}_{T, \text{homo}}| y_N, Y_0] \hat{\beta}
\]
\[
= \hat{\beta}' \mathbb{E}[\hat{\Sigma}_{T, \text{homo}}| Y_0] \hat{\beta}
\]
\[
= \nu_{h_0}^v.
\]
Similarly for VT regression, we conclude \( \mathbb{E}[\hat{\Sigma}_{N, \text{homo}}| Y_0] = \Sigma_N \) and \( \mathbb{E}[\nu_{v_0, \text{homo}}^v| y_T, Y_0] = \nu_{v_0}^v \).

### D.3 Proof of Lemma 2

**Proof.** Before we establish the bias of \( \hat{\Sigma}_{T, \text{jack}} \) and \( \hat{\Sigma}_{N, \text{jack}} \), we first justify their forms. Consider HZ regression. Recall from (42) that \( \text{Var}(\hat{Y}_{NT}^h(0)| y_N, Y_0) = y_N' \text{Cov}(\hat{\alpha}| Y_0) y_N \). As noted in Section 4.1.3, jackknife is a popular methodology to estimate \( \text{Cov}(\hat{\alpha}| Y_0) \). Below, we follow the standard techniques to derive the jackknife estimate of \( \text{Cov}(\hat{\alpha}| Y_0) \), which will then be used to derive \( \hat{\Sigma}_{T, \text{jack}} \). However, while the standard derivation considers \( Y_0 \) with full column rank, we consider a general matrix \( Y_0 \) that may be rank deficient. This difference is subtle so the following proof is by no means novel. We provide it simply for completeness.

To describe the jackknife, we define \( \widehat{\alpha}_{-i} \) as the minimum Euclidean norm solution to (1) (where \( \lambda_1 = \lambda_2 = 0 \)
without the $i$th observation, i.e.,

$$\hat{\alpha}_{\sim i} = (Y_{0,\sim i} Y_{0,\sim i})^T Y_{0,\sim i} y_{T,\sim i}, \quad (43)$$

where $Y_{0,\sim i}$ and $y_{T,\sim i}$ correspond to $Y_0$ and $y_T$ without the $i$th observation. We define the pseudo-estimator as $\hat{\alpha}_i = T_0 \hat{\alpha} - (T_0 - 1)\hat{\alpha}_{\sim i}$. With these quantities defined, we write the jackknife variance estimator as

$$\hat{\Sigma} = \frac{1}{(T_0 - 1)^2} \sum_{i \leq N_0} (\hat{\alpha}_i - \hat{\alpha})(\hat{\alpha}_i - \hat{\alpha})' \quad (44)$$

To evaluate this quantity, we will rewrite $\hat{\alpha}_{\sim i}$ in a more convenient form. In particular, we first note that

$$\begin{align*}
Y'_{0,\sim i} Y_{0,\sim i} &= Y_0' Y_0' - y_i y_i', \\
Y'_{0,\sim i} y_{T,\sim i} &= Y_0' y_T - y_i y_{iT},
\end{align*}$$

where $y_i = [Y_{it} : t \leq T_0]$ is the $i$th row of $Y_0$. We do not assume that $Y_0' Y_0$ is nonsingular. As such, we use a generalized form of the Sherman-Morrison formula (Cline, 1965; Meyer, 1973) to obtain

$$(Y'_0 Y_0)^T = (Y'_0 Y_0)^T + (1 - H_i^{(u)})^{-1}(Y'_0 Y_0)^T y_i y_i'(Y'_0 Y_0)^T.$$  

Recall $\hat{\alpha} = (Y'_0 Y_0)^T Y'_0 y_T$ and note $y_{iT} - y_i \hat{\alpha}$ is the $i$th element of $\hat{\varepsilon}_T = H^{(u)}_i y_T$. Using these facts, we plug (45) into (43) to yield

$$\begin{align*}
\hat{\alpha}_{\sim i} &= \left[ (Y'_0 Y_0)^T + (1 - H_i^{(u)})^{-1}(Y'_0 Y_0)^T y_i y_i'(Y'_0 Y_0)^T \right] (Y'_0 y_T - y_i y_{iT}) \\
&= \hat{\alpha} - (Y'_0 Y_0)^T y_i y_{iT} + (1 - H_i^{(u)})^{-1}(Y'_0 Y_0)^T y_i y_i'(\hat{\alpha} - H_i^{(u)})(1 - H_i^{(u)})^{-1}(Y'_0 Y_0)^T y_i y_{iT} \\
&= \hat{\alpha} - (1 - H_i^{(u)})^{-1}(Y'_0 Y_0)^T y_i \hat{\varepsilon}_{iT}. \quad (46)
\end{align*}$$

Inserting (46) into our pseudo-estimate, we have

$$\begin{align*}
\hat{\alpha}_i &= T_0 \hat{\alpha} - (T_0 - 1) \left[ \hat{\alpha} - (1 - H_i^{(u)})^{-1}(Y'_0 Y_0)^T y_i \hat{\varepsilon}_{iT} \right] \\
&= \hat{\alpha} + (T_0 - 1)(1 - H_i^{(u)})^{-1}(Y'_0 Y_0)^T y_i \hat{\varepsilon}_{iT}. \quad (47)
\end{align*}$$

Inserting (47) into (44), we have

$$\hat{\Sigma} = (Y'_0 Y_0)^T \left( \sum_{i \leq N_0} \frac{\hat{\varepsilon}_{iT}^2}{1 - H_i^{(u)^2}} y_i y_i' \right) (Y'_0 Y_0)^T$$

$$= (Y'_0 Y_0)^T Y_0^T \Omega Y_0 (Y'_0 Y_0)^T,$$

where $\Omega$ is a diagonal matrix with $\Omega_{ii} = \hat{\varepsilon}_{iT}^2 (1 - H_i^{(u)})^{-2}$. Equivalently, $\Omega = \text{diag}([H^{(u)}_i \circ H^{(u)}_i \circ I]^T [\hat{\varepsilon}_T \circ \hat{\varepsilon}_T])$. It then follows that

$$y_N' \hat{\Sigma}^{\text{jack}} y_N = \hat{\beta}' \Omega \hat{\beta}.$$  

To arrive at (8), we define $\hat{\Sigma}_T^{\text{jack}} = \Omega$. This corresponds to the EHW estimator with the jackknife correction. We arrive at (9) for VT regression by applying the same arguments above.
From this, we conclude that

\[ \mathbb{E}[(\mathbf{H}_\perp (u) \circ \mathbf{H}_\perp (u) \circ \mathbf{I})^\dagger (\mathbf{e}_T \circ \mathbf{e}_T)] \mathbf{Y}_0 = (\mathbf{H}_\perp (u) \circ \mathbf{H}_\perp (u) \circ \mathbf{I})^\dagger \mathbb{E}[\mathbf{e}_T \circ \mathbf{e}_T] \mathbf{Y}_0. \]  

(48)

To evaluate (48), we follow the derivations of (38) and (41) to obtain

\[ \mathbb{E}[\mathbf{e}_T | \mathbf{Y}_0] = \mathbf{H}_\perp (u) \mathbf{Y}_0 \alpha^* = \mathbf{0} \]  

(49)

\[ \text{Cov}(\mathbf{e}_T | \mathbf{Y}_0) = \mathbf{H}_\perp (u) \Sigma_T \mathbf{H}_\perp (u). \]  

(50)

Recall that \( \mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2 \) for any random variable \( X \). Thus, combining (49) with (50) gives

\[ \mathbb{E}[\mathbf{e}_T \circ \mathbf{e}_T | \mathbf{Y}_0] = (\mathbf{H}_\perp (u) \Sigma_T \mathbf{H}_\perp (u) \circ \mathbf{I}) 1. \]  

(51)

Let \( \hat{\gamma} = \mathbb{E}[\mathbf{e}_T \circ \mathbf{e}_T | \mathbf{Y}_0] \). By (51), the \( i \)th entry of \( \hat{\gamma} \) can be written as

\[ \hat{\gamma}_i = \sum_{j \neq i} (\mathbf{H}_{ji}^{(u)})^2 \sigma_{Tj}^2 + (1 - H_{ii}^{(u)})^2 \sigma_{iT}^2, \]

where \( H_{ji}^{(u)} \) is the \((j, i)\)th entry of \( \mathbf{H}^{(u)} \). In turn, this allows us to rewrite (51) as

\[ \hat{\gamma} = (\mathbf{H}_\perp (u) \circ \mathbf{H}_\perp (u)) \Sigma_T 1. \]  

(52)

Next, let \( \hat{\zeta} = (\mathbf{H}_\perp (u) \circ \mathbf{H}_\perp (u) \circ \mathbf{I})^{-1} \hat{\gamma} \). Notice that the \( i \)th entry of \( \hat{\zeta} \) is given by

\[ \hat{\zeta}_i = \sigma_T^2 + \sum_{j \neq i} \frac{(H_{ij}^{(u)})^2}{(1 - H_{ii}^{(u)})^2} \sigma_{Tj}^2. \]

Therefore, \( \text{diag}(\hat{\zeta}) = \Sigma_T + \Delta \), where \( \Delta_{ii} = \sum_{j \neq i} \sigma_{Tj}^2 (H_{ij}^{(u)})^2 (1 - H_{ii}^{(u)})^{-2} \). Now, notice if \( \max_i H_{ii}^{(u)} < 1 \), then \( (\mathbf{H}_\perp (u) \circ \mathbf{H}_\perp (u) \circ \mathbf{I}) \) is nonsingular, i.e., the pseudo-inverse is precisely the inverse. In this situation, plugging the above into (48) gives the desired result:

\[ \mathbb{E}[^{\text{jack}} \Sigma_T | \mathbf{Y}_0] = \text{diag} \left( (\mathbf{H}_\perp (u) \circ \mathbf{H}_\perp (u) \circ \mathbf{I})^{-1} \mathbb{E}[\mathbf{e}_T \circ \mathbf{e}_T | \mathbf{Y}_0] \right) \]

\[ = \text{diag} \left( (\mathbf{H}_\perp (u) \circ \mathbf{H}_\perp (u) \circ \mathbf{I})^{-1} \hat{\gamma} \right) \]

\[ = \text{diag}(\hat{\zeta}) \]

\[ = \Sigma_T + \Delta. \]

From this, we conclude that

\[ \mathbb{E}[v_{0, \text{jack}}^\prime | \mathbf{y}_N, \mathbf{Y}_0] = \beta' \mathbb{E}[^{\text{jack}} \Sigma_T | \mathbf{y}_N, \mathbf{Y}_0] \beta \]

\[ = \beta' \mathbb{E}[^{\text{jack}} \Sigma_T | \mathbf{Y}_0] \beta \]

\[ = \beta'(\Sigma_T + \Delta) \beta \]

\[ = v_{0, \text{jack}}^\prime + \beta' \Delta \beta, \]
where we note that \( \hat{\beta}^T \Delta \hat{\beta} \geq 0 \). Following the same arguments for \( VT \) regression, we conclude
\[
E(\hat{\Sigma}_N^{\text{jack}} | Y_0) = \Sigma_N + \Gamma \\
E(\hat{\Sigma}_0^{\text{vt, jack}} | y_T, Y_0) = v_0^{vt} + \hat{\alpha}' T \hat{\alpha},
\]
where \( \hat{\alpha}' T \hat{\alpha} \geq 0 \). The proof is complete. \( \square \)

D.4 Proof of Lemma 3

Proof. Below, we adopt the strategy of Hartley et al. (1969) to prove our desired result. Consider \( HZ \) regression. As in the proof of Lemma 2, let \( \hat{e}_T = H^{(u)}_1 y_T \). Observe that
\[
E[(H^{(u)}_1 \circ H^{(u)}_1)^{-1}(\hat{e}_T \circ \hat{e}_T)|Y_0] = (H^{(u)}_1 \circ H^{(u)}_1)^{-1} E[\hat{e}_T \circ \hat{e}_T|Y_0]. \tag{53}
\]
To evaluate (53), we plug in (52) to obtain
\[
E[(H^{(u)}_1 \circ H^{(u)}_1)^{-1}(\hat{e}_T \circ \hat{e}_T)|Y_0] = (H^{(u)}_1 \circ H^{(u)}_1)^{-1} E[\hat{e}_T \circ \hat{e}_T|Y_0] = \Sigma_T. \tag{54}
\]
Plugging (54) into (53) yields
\[
E[\hat{\Sigma}_T^{\text{HRK}} | Y_0] = \text{diag} ((H^{(u)}_1 \circ H^{(u)}_1)^{-1} E[\hat{e}_T \circ \hat{e}_T|Y_0]) = \Sigma_T.
\]
It then follows that \( E[\hat{\epsilon}_0^{\text{hz, HRK}} | y_N, Y_0] = v_0^{hz} \). The same arguments apply towards \( VT \) regression, we conclude
\[
E[\hat{\Sigma}_N^{\text{HRK}} | Y_0] = \Sigma_N \\
E[\hat{\epsilon}_0^{\text{vt, HRK}} | y_T, Y_0] = v_0^{vt}.
\]
This completes the proof. \( \square \)

D.5 Proof of Theorem 4

Proof. Consider \( HZ \) regression. Throughout, we will only condition on \( Y_0 \). We write
\[
\hat{Y}_{N_T}(0) = y_N^T Y_0^T = \sum_{i<N} \sum_{t<T} (Y_0^T)_i Y_{it} Y_{Nt}. \tag{55}
\]
By our independence assumptions, (55) is a sum of independent random variables with
\[
E[Y_{it} Y_{Nt}|Y_0] = E[Y_{it}|Y_0]E[Y_{Nt}|Y_0] \\
\text{Var}(Y_{it} Y_{Nt}|Y_0) = E[Y_{it}|Y_0]^2 \sigma_{Nt}^2 + E[Y_{Nt}|Y_0]^2 \sigma_{it}^2 + \sigma_{it}^2 \sigma_{Nt}^2.
\]
Theorem 13 then establishes that
\[
\frac{\hat{Y}_{N_T}(0) - E[\hat{Y}_{N_T}(0)] | Y_0}{\sqrt{\text{Var}(\hat{Y}_{N_T}(0)|Y_0)^{1/2}}} \overset{d}{\rightarrow} N(0, 1).
\]
Our aim is to evaluate $E[\hat{Y}_{NT}^{hz}(0)|Y_0]$ and $\text{Var}(\hat{Y}_{NT}^{hz}(0)|Y_0)$. Towards the former, we apply the law of total expectation and recall (39) to obtain

$$E[\hat{Y}_{NT}^{hz}(0)|Y_0] = E[\hat{Y}_{NT}^{hz}(0)|y_N, Y_0|Y_0]$$
$$= E[y_N|Y_0]Y_0^\alpha^*$$
$$= \langle \beta^*, Y_0\alpha^* \rangle$$
$$= E[Y_{NT}(0)|Y_0].$$

By the law of total variance,

$$\text{Var}(\hat{Y}_{NT}^{hz}(0)|Y_0) = E[\text{Var}(\hat{Y}_{NT}^{hz}(0)|y_N, Y_0)|Y_0] + \text{Var}(E[\hat{Y}_{NT}^{hz}(0)|y_N, Y_0]|Y_0).$$

(56)

Following the derivation of (42), we further have

$$E[\text{Var}(\hat{Y}_{NT}^{hz}(0)|y_N, Y_0)|Y_0] = E[y_N Y_0^\dagger \Sigma_T Y_0^*|Y_0]$$
$$= (Y_0^*\beta^*)'A\Sigma_N(Y_0^*\beta^*) + E[e_N' A e_N|Y_0] + 2E[e_N' Y_0^*\beta^*|Y_0],$$

(57)

where $A = Y_0^\dagger \Sigma_T(Y_0^*)\dagger$. Note that $A$ is deterministic given $Y_0$. Observe $E[e_N' Y_0^*\beta^*|Y_0] = 0$ from Assumption 3 (ii.b). Further, Lemma 12 gives

$$E[e_N' A e_N|Y_0] = \text{tr}(A\Sigma_N),$$

(58)

Following the arguments that led to the derivation of (39),

$$\text{Var}(E[\hat{Y}_{NT}^{hz}(0)|y_N, Y_0]|Y_0) = \text{Var}(y_N H^{(v)}\alpha^*|Y_0) = (H^{(v)}\alpha^*)'\Sigma_N(H^{(v)}\alpha^*).$$

(59)

Plugging (57), (58), and (59) into (56), we arrive at the following:

$$\text{Var}(\hat{Y}_{NT}^{hz}(0)|Y_0) = (H^{(v)}\alpha^*)'\Sigma_N(H^{(v)}\alpha^*) + (H^{(u)}\beta^*)'\Sigma_T(H^{(u)}\beta^*) + \text{tr}(Y_0^\dagger \Sigma_T(Y_0^*)\dagger \Sigma_N).$$

Let us now switch to VT regression. By Theorem 1, $\hat{Y}_{NT}^{vt}(0) = \hat{Y}_{NT}^{hz}(0)$. Therefore, the same analysis for HZ regression also holds for VT regression. This concludes the proof.

D.6 Proof of Corollary 1

Proof. By linearity of expectations,

$$E[\tilde{v}_0|Y_0] = E[\tilde{v}_0^{hz}|Y_0] + E[\tilde{v}_0^{vt}|Y_0] - E[\text{tr}(Y_0^\dagger \Sigma_T(Y_0^*)\dagger \Sigma_N)]|Y_0].$$

(60)
Let us independently evaluate each term in (60). Under Assumption 3 (i.b), (ii.b), and (iii), Lemma 12 gives

\[
\mathbb{E}[\hat{\sigma}_0^{\text{homo}}|Y_0] = \mathbb{E}[\hat{\beta}' \hat{\Sigma}_T \hat{\beta}|Y_0]
\]
\[
= \mathbb{E}[\mathbb{E}[\hat{\beta}' \hat{\Sigma}_T \hat{\beta}|y_N, Y_0]|Y_0]
\]
\[
= \mathbb{E}\left[y_N' Y_0 | \hat{\Sigma}_T | Y_0 | (Y_0')^\dagger | y_N | Y_0 \right]
\]
\[
= \mathbb{E}\left[(Y_0' \beta^* + \varepsilon_N) Y_0 | \hat{\Sigma}_T | Y_0 | (Y_0')^\dagger | (Y_0' \beta^* + \varepsilon_N) | Y_0 \right]
\]
\[
= (H^{(u)} \beta^*)' \mathbb{E}[\hat{\Sigma}_T | Y_0 | (H^{(u)} \beta^*) + \mathbb{tr}(Y_0' \mathbb{E}[\hat{\Sigma}_T | Y_0 | (Y_0')^\dagger | \Sigma_N]). \quad (61)
\]

By a similar argument, we derive

\[
\mathbb{E}[\hat{\sigma}_0^{\text{jack}}|Y_0] = (H^{(v)} \alpha^*)' \mathbb{E}[\hat{\Sigma}_N | Y_0 | (H^{(v)} \alpha^*) + \mathbb{tr}(Y_0' \mathbb{E}[\hat{\Sigma}_T | Y_0 | (Y_0')^\dagger | \Sigma_N)]. \quad (62)
\]

Transitioning to the trace term,

\[
\mathbb{E}[\mathbb{tr}(Y_0' \hat{\Sigma}_T | Y_0' ) | \Sigma_N)]Y_0] = \mathbb{E}\left[\mathbb{tr}(Y_0' \mathbb{E}[\hat{\Sigma}_T | Y_0' ) | \Sigma_N)]Y_0 \right]
\]
\[
= \mathbb{tr}(Y_0' \mathbb{E}[\hat{\Sigma}_T | Y_0' ) | \Sigma_N)]Y_0]
\]
\[
= \mathbb{tr}(Y_0' \mathbb{E}[\hat{\Sigma}_T | Y_0' ) | \Sigma_N)]Y_0]. \quad (63)
\]

**I: Homoskedastic errors.** Plugging Lemma 1 into (61), (62), and (63), we obtain

\[
\mathbb{E}[\hat{\sigma}_0^{\text{homo}}|Y_0] = \sigma_0^2 (H^{(u)} \beta^*)' (H^{(u)} \beta^*) + \sigma_0^2 (H^{(v)} \alpha^*)' (H^{(v)} \alpha^*) + \sigma_0^2 \sigma_N^2 \mathbb{tr}(Y_0' (Y_0')^\dagger) = \nu_0.
\]

**II: Heteroskedastic errors–Jackknife.** Plugging Lemma 2 into (61), (62), and (63), we obtain

\[
\mathbb{E}[\hat{\sigma}_0^{\text{jack}}|Y_0] = \nu_0 + (H^{(u)} \beta^*)' \Delta (H^{(u)} \beta^*) + (H^{(v)} \alpha^*)' \Gamma (H^{(v)} \alpha^*) + \mathbb{tr}(Y_0' \Delta (Y_0')^\dagger \Gamma).
\]

**III: Heteroskedastic errors–HRK.** Plugging Lemma 3 into (61), (62), and (63), we obtain

\[
\mathbb{E}[\hat{\sigma}_0^{\text{HRK}}|Y_0] = (H^{(u)} \beta^*)' \Sigma_T (H^{(u)} \beta^*) + (H^{(v)} \alpha^*)' \Sigma_N (H^{(v)} \alpha^*) + \mathbb{tr}(Y_0' \Sigma_T (Y_0')^\dagger \Sigma_N) = \nu_0.
\]

The proof is complete. \(\square\)

### E Proofs for The Researcher’s Guide to Panel Data

#### E.1 Proof of Corollary 2

**Proof.** We begin with the OLS setting. Recall that \(\hat{Y}^{\text{ht}}_{N,T}(0) = \langle y_N, \hat{\alpha} \rangle \) and \(\hat{Y}^{\text{vt}}_{N,T}(0) = \langle y_T, \hat{\beta} \rangle\). By Theorem 1,

\[
\hat{Y}^{\text{ht}}_{N,T}(0) = \hat{Y}^{\text{vt}}_{N,T}(0) = \langle y_N, Y_0' y_T \rangle.
\]

53
thus $\hat{\alpha} = Y_0^+ y_T$ and $\hat{\beta} = (Y_0')^+ y_N$. Returning to (11), we obtain

$$\tilde{Y}_{NT}^\text{std}(0) = \tilde{Y}_{NT}^\text{asc}(0) = \langle y_T, \hat{\beta} \rangle + \langle y_N, \hat{\alpha} \rangle - \langle \hat{\alpha}, Y_0' \hat{\beta} \rangle = 2\tilde{Y}_{NT}^\text{hz}(0) - \langle \hat{\alpha}, Y_0' \hat{\beta} \rangle.$$

(64)

Recall $(Y_0')^+ = (Y_0')'$. Therefore,

$$\langle \hat{\alpha}, Y_0' \hat{\beta} \rangle = y_T (Y_0')^+ (Y_0')^\dagger y_N = y_T ^\dagger \hat{\beta}.$$ 

(65)

Plugging (65) into (64), we conclude

$$\tilde{Y}_{NT}^\text{std}(0) = \tilde{Y}_{NT}^\text{asc}(0) = 2\tilde{Y}_{NT}^\text{hz}(0) - \tilde{Y}_{NT}^\text{hz}(0) = \tilde{Y}_{NT}^\text{hz}(0) = \tilde{Y}_{NT}^\text{st}(0).$$

Now, observe that the same arguments above hold when $Y_0^{(k)}$ takes the place of $Y_0$ for any $k < R$. Therefore, the same reduction can be derived for PCR.