# Causal Inference with Difference-in-Difference Models under Parallel Trends Violations: Revisiting Rambachan and Roth (2023)

Master Thesis Presented to the

Department of Economics at the

Rheinische Friedrich-Wilhelms-Universität Bonn

In Partial Fulfillment of the Requirements for the Degree of

Master of Science (M.Sc.)

Supervisor: Prof. Dr. Dominik Liebl

Submitted in April 2024 by:
Zihan Yang
Matriculation Number: 3504373

# Contents

# 1 Introduction

Pre-post assessments of policy changes may yield misleading results if underlying time-dependent trends in outcomes exist before the intervention. The Difference-in-Differences (DiD) study design addresses this by utilizing a comparison group experiencing similar trends but without the policy change. This approach captures significant differences in outcomes between treatment and control groups during both pre- and post-treatment periods, providing a more accurate assessment of the policy's impact.

For a robust estimation in DiD analysis, a critical assumption is the parallel trends assumption. The parallel trends assumption states that the trends in outcomes between the treated and comparison groups are the same before the intervention. If true, it is reasonable to assume that these parallel trends would continue for both groups even if the program was not implemented.

Researchers often assess the plausibility of the parallel trends assumption by testing for "pre-trends". However, a handful of recent papers such as Bilinski & Hatfield (2018) have warned that the common approach of testing for pre-trends will be imperfect in finite samples, and may suffer from low power, e.g.,Kahn-Lang & Lang (2020).

This paper shows the potential deviations from the assumption of parallel trends and introduces a methodology developed by Rambachan & Roth (2023). The method is presented to facilitate robust inference and sensitivity analysis in empirical settings, particularly in scenarios where the parallel trend assumption may not hold. The application of this methodology addresses various violations, providing a comprehensive approach to handling different situations and enhancing the reliability of empirical results.

In Section 2, the method of DiD is introduced, laying the foundation for subsequent analyses. Moving forward to Section 3, I thoroughly show various violations of the parallel trends assumption and provide a detailed demonstration of Rambachen and Roth's method and the inference. Section 4 is dedicated to the simulation part, different from the Rambachan & Roth (2023), which uses estimates from 12 existing papers, I established a data generation process and then estimated $\hat{\sigma}$ with two-way fixed effect(TWFE). Furthermore, I conduct a comparative analysis of the effectiveness of different methods in addressing various parallel trends assumption violations. Finally, in Section 5, the HonestDiD package is used to conduct sensitivity analysis, wherein the methodology is applied to two distinct examples to ensure a comprehensive evaluation

An additional aspect should be mentioned: all code used in this thesis is available in the following public GitHub repository: https://github.com/Zihan0000/Honest-DiD.

# 2 Difference-in-Differences(DiD)

In this section, I will introduce two types of DiD methods. The first is the simple version, known as canonical DiD, and the second is the DiD with multiple periods.

## 2.1 Canonical DiD

In the canonical DiD, there are two time periods, and two groups: in the first period no one is treated, and in the second period some units are treated (the treated group), and some units are not (the comparison group). Methodological extensions of DiD methods often focus on this standard two periods, two groups setup; see, e.g.,Abadie (2005), Qin & Zhang (2008), Callaway *et al.* (2018).

To interpret a difference estimate as a policy effect there are three key conditions: the estimated regression is the correct conditional mean; the policy is exogenous, which means that it satisfies the conditional independence; there are no other relevant unincluded factors coincident with the policy change. The DiD estimation is valid if these assumptions are satisfied. In the example shown, the treatment group receives treatment between time 0 and time 1. The outcome in the treatment group is represented by the red line, and the outcome in the control group is represented by the blue line. The dashed blue line illustrates the potential values of the treatment group without treatment.

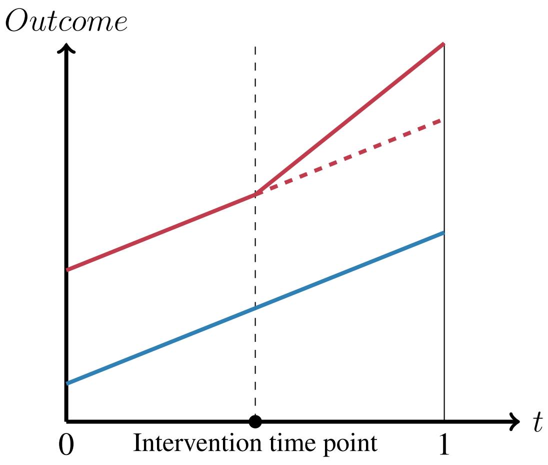


Figure 1: The graphical explanation of canonical DiD

As usual when important quantitative measures vary across observational units a canonical summary statistic is the average. The Average Treatment Effect (ATE) is the average of the individual treatment effects of the population under consideration. The Average Treatment

Effect on the Treated (ATT) is the average of the individual treatment effects of those treated (hence not the entire population). As in most of the DiD literature, we focus on the ATT.

Assume $T$ = 2, and there is a binary policy $d_i$ that is received between period $t$ = 1 and $t$ = 2. The potential outcome notation for $y_{it}$: $y_{it}(0)$ is the outcome of individual $i$ at period $t$ if he does not receive the treatment; $y_{it}(1)$ is the outcome of individual $i$ at period $t$ if he receives the treatment. Thus we can write:

$$y_{it} = d_i y_{it}(1) - (1 - d_i) y_{it}(0) \tag{1}$$

Then, the ATT in period 2 is:

$$E[y_{i,2}(1) - y_{i,2}(0) | d_i = 1] \tag{2}$$

To identify ATT, we need two important assumptions: parallel trends and no-anticipation assumptions. Parallel trends assumption suggests that in the absence of the treatment, the average outcome would have evolved in parallel:

$$E[y_{i,2}(0) - y_{i,1}(0) | d_i = 1] = E[y_{i,2}(0) - y_{i,1}(0) | d_i = 0]$$

No-anticipation assumption states that the policy does not have an effect before the treatment period. This can be expressed mathematically as $y_{i,1}(0) = y_{i,1}(1)$. The two assumptions play a significant role in calculating ATT. The specific process of calculating ATT is outlined in the Appendix A

An illustrative example is provided below: Burde & Linden (2013) who looked at the effect of building new village schools (as opposed to having children commuting) on students' academic outcomes. They estimated the following model: $Y_{ijk} = \beta_0 + \beta_1 D_k + e_{ijk}$, Where $Y_{ijk}$ is an academic outcome of child $i$ in household $j$ in a village $k$. Here the $D_k$ is a dummy that signifies whether the village got a school in the first year or not, and it can only take two values 1 or 0. In this setting, we will have 2 possible conditional outcomes:

$$Y_{ij1} = \beta_0 + \beta_1 \times 1 + e_{ij1}$$

$$Y_{ij0} = \beta_0 + \beta_1 \times 0 + e_{ij0}$$

The average treatment effect is: $ATE = E(Y_{ij1} - Y_{ij0})$. It is the difference between potential outcomes, which in this case is the academic achievement between children who got access to village schools and children who did not get access to village schools. When it comes to the average treatment effect on treated this is defined as $ATT = E(Y_{ij1} - Y_{ij0}|D_k = 1)$. So this is the difference in academic potential outcomes between children that got access to village schools, and children that did not get access to village schools conditionally on the fact that they both are assigned to village schools.

## 2.2 DiD with Multiple Periods

### 2.2.1 Setup

We consider the case with $T$ periods and denote a particular time period by $t$ where $t = 1, ...T$. Let $d_{it}$ be a binary variable equal to one if unit $i$ is treated in period $t$ and equal to zero otherwise. Here are the assumptions about the treatment process:

**Assumption 2.1 (Irreversibility of Treatment)** $D_1 = 0$ *almost surely. And for* $t = 2, ...T$, $D_{t-1} = 1$*, implies that* $D_t = 1$.

Assumption 2.1 states that no one is treated at time $t = 1$, and once a unit becomes treated, it will remain treated in the next period.

**Assumption 2.2 (Random Sampling)** $\{Y_{i,1}, Y_{i,2}, ...Y_{i,T}, X_i, d_{i,1}, d_{i,2}, ...d_{i,T}\}_{i=1}^{n}$ *is independent and identically distributed (iid).*

Assumption 2.2 implies access to panel data, which typically involves observing a set of individuals (units) over multiple time periods. Furthermore, this assumption states that there are no constraints or limitations regarding the relationship between potential outcomes and the allocation of treatment. Similarly, there are no restrictions on the time series dependence of the observed random variables. On the other hand, this assumption also introduces a constraint by stating that each unit (denoted as $i$) is randomly drawn from a large population of interest, this implies that the units are representative of a broader population, and the random drawing ensures that the sample is not biased.

### 2.2.2 The Group-Time Average Treatment Effect

Define $G$ as the time period when a unit is first being treated. $G_g$ is a binary variable that is equal to one if a unit is first treated in the period $g$ (i.e., $G_{i,g} = 1\{G_i = g\}$, and define $C$ as a binary variable that is equal to one for units that do not participate in the dataset (i.e.,

$C_i = 1\{G_i = \infty\} = 1 - d_{i,T}$). Let $\bar{g} = max_{i=1,...n} G_i$ be the maximum $G$ in the dataset. And define $p_g(X) = p_{g,T}(X) = P(G_g = 1|X, G_g + C = 1)$ as the probability of being first treated in period $g$ conditional on covariates and either being a member of group $g$ or not participating in the treatment in any time period. Let $\hat{g} = supp(G) \backslash \{\bar{g}\} \subseteq \{2, 3, ..., T\}$ denote the support of $G$ excluding $\bar{g}$.

Next, the potential outcomes framework is set up. Let $Y_{i,t}(0)$ represent the untreated potential outcome of unit $i$ at time $t$ if they remain untreated throughout the time period $T$. For $g = 2, ..., T$, let $Y_{i,t}(g)$ denote the potential outcome that unit $i$ would experience at time $t$ if they were treated in time period $g$. Then, the observed and potential outcomes for each unit $i$ are related through:

$$Y_{i,t} = Y_{i,t}(0) + \sum_{g=2}^{T} (Y_{i,t}(g) - Y_{i,t}(0)) \cdot G_{i,g} \tag{3}$$

Then we can construct the ATT with multiple treatment groups and multiple periods, similar to Callaway & Sant'Anna (2021), which computes the average treatment effect for units who are members of a particular group $g$ at a specific time period $t$. This is denoted by:

$$ATT(g, t) = E[Y_t(g) - Y_t(0)|G_g = 1] \tag{4}$$

The $ATT(g, t)$ does not impose constraints on the heterogeneity of treatment effects across groups or time. Consequently, directing attention to the set of $ATT(g, t)$'s enables a comprehensive analysis of how average treatment effects fluctuate across diverse dimensions. We introduce the following assumptions to establish the identification of $ATT(g, t)$.

**Assumption 2.3 (Limited Treatment Anticipation)** *There is a known $\delta \geq 0$, for all $g \in \hat{g}$, $t \in \{1, ..., T\}$ such that t <g$-\delta$.*

$$E[Y_t(g)|X, G_g = 1] = E[Y_t(0)|X, G_g = 1] \tag{5}$$

Assumption 2.3 imposes restrictions on the anticipation of treatment for all "eventually treated" groups. Specifically, when $\delta = 0$, it represents the "no-anticipation" assumption. Conversely, when $\delta \neq 0$, it permits some degree of anticipation behavior, as exemplified by Malani & Reif (2015).

**Assumption 2.4 (Conditional Parallel Trends Based on a "Never-Treated" Group)** *Here, we use the same assumption outlined in Callaway & Sant'Anna (2021).*

*Let $\delta$ be as defined in assumption 2.3. For each $g \in \hat{g}$ and $t \in \{2, ..., T\}$ such that $t \geq g - \delta$,*

$$E[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = E[Y_t(0) - Y_{t-1}(0)|X, C = 1] \tag{6}$$

**Assumption 2.5 (Conditional Parallel Trends Based on "Not-Yet-Treated" Groups)** *Let $\delta$ be as defined in assumption 2.3. For each $g \in \hat{g}$ and each $(s, t) \in \{2, ..., T\} \times \{2, ..., T\}$ such that $t \geq g - \delta$ and $t + \delta \leq s < \bar{g}$,*

$$E[Y_t(0) - Y_{t-1}(0)|X, G_g] = E[Y_t(0) - Y_{t-1}(0)|X, d_s = 1, G_g = 0] \tag{7}$$

Assumptions 2.4 and 2.5 extend the concept of parallel trends assumption from two periods (see, e.g.,Abadie (2005), and Sant'Anna & Zhao (2020)) to situations involving multiple time periods and treatment groups. Assumption 2.4 states that conditional on $X$, the average outcomes for the group first treated in time period $g$ and for the group never treated follow parallel trends in the absence of treatment. Assumption 2.5 states that conditional on $X$, the average outcomes for the group treated in $g$ and the groups not yet treated by time $t + \delta$ follow parallel trends.

**Assumption 2.6 (Overlap)** *For each $t \in \{2, ..., T\}, g \in \hat{g}$, there exist some $\epsilon > 0$ such that $P(G_g = 1) > \epsilon$ and $p_{g,t}(X) < 1 - \epsilon$.*

Assumption 2.6 states that, a portion of the population begins treatment in time period $g$, and for treatment $g$ and all periods $t$, the propensity score is constrained in such a way that it is not close to one, maintaining a level of variation in the likelihood of treatment across different time periods.

ATT in multiple time periods and treatment groups can be calculated by outcome regression, inverse probability weighting, or doubly robust, and the above assumptions play a great role when forming these different estimations. (see Appendix A of Callaway & Sant'Anna (2021)).

## 2.3 Pre-Trends Tests

The key assumption in DiD is the parallel trends assumption, which requires that when the treatment of interest does not occur, the mean of the outcome would have evolved in parallel between the treatment and control groups. A popular way to test the plausibility of the assumption is pre-trends tests, which is the differences in trends between the treatment and control group before the time period of treatment assignment($\delta_s$).

In the pre-trends tests, $H_0$ assumption is: $\forall s < 0, \delta_s = 0$. However, the pre-trends tests have some drawbacks, which are explained clearly by Roth (2022): fristly, conventional pre-trends tests for parallel trends often have low power even against substantial linear violations of parallel trends. This suggests that the possibility of failing to detect a meaningful violation of parallel trends is not merely a theoretical concern. Second, in many cases, the bias and coverage issues can be substantially different conditionally on surviving pre-trends tests, and in some instances, they can be exacerbated. Even though homoskedasticity typically does not hold in practice, the bias from pre-trends tests nonetheless amplifies the bias from a monotone trend in most cases and can be of a substantial magnitude. Third, parametric approaches to controlling for pre-existing trends may be sensitive to functional form assumptions, which is clearly explained by Lee & Solon (2011). So now instead of insisting on strict adherence to parallel trends, this paper involves placing constraints on the permissible divergence between post-treatment violations of parallel trends and pre-treatment differences in trends.

# 3 Rambachan and Roth (2023)

In this section, the different kinds of restrictions on how different the post-treatment violations of parallel trends can be from the pre-treatment differences in trends will be introduced. I will present two approaches from Rambachan & Roth (2023) that ensure uniformly valid inference under these restrictions, and show the derived results which indicate they have desirable power properties.

## 3.1 Model Setup

### 3.1.1 Event Study Coefficients

In a non-staggered DiD model, assume the $\underline{T}$ is the number of pre-treatment periods, the $\bar{T}$ is the number of post-treatment periods, and $t = \underline{T}, ..., \bar{T}$. Here, $D_i = 1$ indicates that the unit receives treatment beginning in period $t = 1$, while $D_i = 0$ indicates that the unit never receives treatment throughout the entire period. The parameter $\hat{\beta}_s$ compares the difference in the mean outcome between period 0 and period $s$ for treatment and control groups. In this setting, the estimates $\hat{\beta}_s$ are numerically equivalent to the coefficients obtained from ordinary least squares (OLS) regression analysis, where the regression model is:

$$Y_{it} = \lambda_i + \phi_t + \sum_{s \neq 0} \beta_s \times 1[t = s] \times D_i + \epsilon_{st} \tag{8}$$

The estimation of the $\beta$ yields event-study coefficients denoted as $\hat{\beta}$: $\hat{\beta} \in R^{\underline{T}+\bar{T}}$, $\hat{\beta} = (\hat{\beta}'_{pre}, \hat{\beta}'_{post})'$. Here, $\hat{\beta}_{pre}$ represents the estimated coefficients corresponding to the periods before treatment, while $\hat{\beta}_{post}$ represents the estimated coefficients corresponding to the periods after treatment.

We assume the parameter vector $\beta$ can be decomposed as:

$$\beta = \begin{pmatrix} \tau_{pre} \\ \tau_{post} \end{pmatrix} + \begin{pmatrix} \delta_{pre} \\ \delta_{post} \end{pmatrix} \tag{9}$$

$\tau$ is the causal parameter of interest, in which $\tau_{pre}$ is assumed to 0. $\delta$ is a bias from a differences in trend. Under the parallel trends assumption, $\delta_{pre} = 0$, so $\beta_{post} = \tau_{post}$.

**Example of Canonical DiD.** $\hat{\beta}_1$ equals the difference of sample means across treated and untreated between period 0 and 1: $\hat{\beta}_1 = (\bar{Y}_{1,1} - \bar{Y}_{1,0}) - (\bar{Y}_{0,1} - \bar{Y}_{0,0})$, where $\bar{Y}_{d,t}$ is the sample mean of $Y_{it}$ for the treatment group $d$. $\tau$ is the ATT for policy of interest, and $\delta$ is the difference in trends of untreated potential outcomes between the treated and comparison groups. Taking the expectations and re-arranging, we see that:

$$E[\bar{\beta}_1] = \tau_{ATT} + E[Y_{i,1}(0) - Y_{i,0}(0)|D_i = 1] - E[Y_{i,1}(0) - Y_{i,0}(0)|D_i = 0]$$

$$E[\bar{\beta}_{-1}] = E[Y_{i,-1}(0) - Y_{i,0}(0)|D_i = 1] - E[Y_{i,-1}(0) - Y_{i,0}(0)|D_i = 0]$$

$E[Y_{i,1}(0) - Y_{i,0}(0)|D_i = 1] - E[Y_{i,1}(0) - Y_{i,0}(0)|D_i = 0]$ is the $\delta_1$: Post-period differential trend. $E[Y_{i,-1}(0) - Y_{i,0}(0)|D_i = 1] - E[Y_{i,-1}(0) - Y_{i,0}(0)|D_i = 0]$ is the $\delta_{-1}$: pre-treatment periods differential trend.

**Example of staggered DiD**. Define $Y_{it}(g)$ as the potential outcome for unit $i$ in period $t$ if they are first treated at period $g$, and $Y_{it}(\infty)$ is the potential outcome for unit $i$ in period $t$ if they are never treated. $\beta_r = \tau_r + \delta_r$, where $\tau_r = \sum_g w_g ATT(g, g+r)$. $w_g$ are weights that sum to 1, and $ATT(g, g+r) = E[Y_{i,g+r}(g) - Y_{i,g+r}(\infty)|G_i = g]$: the ATT in period $g + r$ for unit $i$ fist treated at period $g$. $\delta_{g,g+r} = E[Y_{i,g+r}(\infty) - Y_{i,g-1}|G_i = g] - E[Y_{i,g+r}(\infty) - Y_{i,g-1}|G_i > g + r]$ is the difference in trends in never treated potential outcome. Under the no-anticipation assumption, $\tau_r = 0$ for $r < 0$. In the staggered DiD design, the comparison groups can be categorized as follows: i) the group never treated compared with the group treated at time point $g$; ii) the group never treated compared with the group treated at time point $s$, where $s > g$; iii) the group treated at time $g$ compared with the group treated at time $s$, with the comparison periods limited to $g < s$. However, there is a potential bias in detecting the treatment effect using TWFE because it considers all time periods for both treatment groups.

### 3.1.2 Target Parameter and Identification

Under the parallel trends, the $\delta$ should be 0, so the coefficient of interest is $\tau$. $\tau_{pre} = 0$, then we build our target parameter: $\theta{:}= l^{'}\tau_{post}$, $l^{'}$ is a vector, so $\theta$ here is a scale about the post-treatment causal effects. For example, if $l^{'} = (1, ..., 1)^{'}$, $\theta$ equals the sum of the causal effect.

Assume $\delta$ lies in a set: $\triangle \subseteq R^{\underline{T}+\bar{T}}$. Then we can assume that $\delta$ did not follow the parallel trends: $\delta \in \triangle \neq \{\delta : \delta_{post} = 0\}$.

Finally, we can use the above assumptions to get a set of $\theta$, which is based on the value of $\beta$ and the set of $\delta$:

$$S(\beta, \triangle) := \{\theta : \exists\delta \in \triangle, \tau_{post} \in R^{\bar{T}} s.t. l^{'}\tau_{post} = \theta, \beta = \delta + \begin{pmatrix} 0 \\ \tau_{post} \end{pmatrix} \tag{10}$$

**Lemma 3.1** *If $\triangle$ is closed and convex, then $S(\beta, \triangle)$ is an interval in $R$. $S(\beta, \triangle) = [\theta^{lb}(\beta, \triangle)$, $\theta^{ub}(\beta, \triangle)]$. $\beta = \delta + \tau$, and $\tau_{pre} = 0$, so $\beta_{pre} = \delta_{pre}$, then we can get the upper bound and the lower bound of the $S(\beta, \triangle)$:*

$$\theta^{lb}(\beta, \triangle) := l^{'}\beta_{post} - \left(\max_{\delta} l^{'}\delta_{post}, s.t.\delta \in \triangle, \delta_{pre} = \beta_{pre}\right) \tag{11a}$$

$$\theta^{ub}(\beta, \triangle) := l^{'}\beta_{post} - \left(\min_{\delta} l^{'}\delta_{post}, s.t.\delta \in \triangle, \delta_{pre} = \beta_{pre}\right) \tag{11b}$$

*And we define the $l^{'}\delta_{post}$ as $b$ constrained by $\triangle$ and $\beta_{pre}$:*

$$\left(\max_{\delta} l^{'}\delta_{post}, s.t.\delta \in \triangle, \delta_{pre} = \beta_{pre}\right) =: b^{max}(\beta_{pre}, \triangle) \tag{12a}$$

$$\left(\min_{\delta} l^{'}\delta_{post}, s.t.\delta \in \triangle, \delta_{pre} = \beta_{pre}\right) =: b^{min}(\beta_{pre}, \triangle) \tag{12b}$$

*When $\triangle = \cup_{k=1}^{K} \triangle_k$, which is a finite union of sets, then the identified set should be:*

$$S(\beta, \triangle) = \bigcup_{k=1}^{K} S(\beta, \triangle_k) \tag{13}$$

### 3.1.3 Inferential Goal

The coefficient $\hat{\beta}_n$ is the estimator of the $\beta$, and the $\hat{\beta}$ will satisfy: $\sqrt{n}(\hat{\beta}_n - \beta) \rightarrow_d \mathcal{N}(0, \Sigma^*)$, so in a finite sample:

$$\hat{\beta}_n \approx_d \mathcal{N}(\beta, \Sigma_n), \Sigma_n = \Sigma^*/n \tag{14}$$

Then the confidence sets $C_n(\hat{\beta}_n, \Sigma_n)$ can be constructed, which are uniformly valid for all parameter values $\theta$ in the identified set when the above approximation equation holds, this typically involves ensuring that the confidence sets provide coverage at the desired confidence level for any true parameter value. Thus the $C_n(\hat{\beta}_n, \Sigma_n)$ satisfy the following equation:

$$\underset{\delta\in\triangle,\tau}{inf}\ \underset{\theta\in S(\delta+\tau,\triangle)}{inf}\ P_{\hat{\beta}_n\approx_d\mathcal{N}(\beta,\Sigma_n)}\left(\theta\in C_n(\hat{\beta}_n,\Sigma_n)\right)\geq 1-\alpha. \tag{15}$$

Suppose $\triangle = \triangle_k$, and $\forall k = 1, 2, ..., K$, the confidence set $C_{n,k}$ satisfies the above equation, then according to equation (13), we can rewrite the coverage requirement as:

$$\underset{\delta\in\triangle,\tau}{inf}\ \underset{k}{inf}\ \underset{\theta\in S(\delta+\tau,\triangle)}{inf}\ P_{\hat{\beta}_n\approx_d\mathcal{N}(\beta,\Sigma_n)}\left(\theta\in\bigcup_k C_{n,k}(\hat{\beta}_n,\Sigma_n)\right)\geq 1-\alpha.$$

And for the confidence set $C_{n,k} = \bigcup\limits_{k=1}^{K} C_{n,k}(\hat{\beta}_n, \Sigma_n)$, The above's left side is bounded below by:

$$\underset{\delta\in\triangle,\tau}{inf}\ \underset{k}{inf}\ \underset{\theta\in S(\delta+\tau,\triangle)}{inf}\ P_{\hat{\beta}_n\approx_d\mathcal{N}(\beta,\Sigma_n)}\left(\theta\in C_{n,k}(\hat{\beta}_n,\Sigma_n)\right)$$

Which is at least $1 - \alpha$.

## 3.2 Relaxation of Parallel Trends Assumption

Here we enumerate several practical choices for the parameter $\triangle$ in empirical applications and formalize intuitive arguments frequently presented by applied researchers regarding possible deviations from parallel trends, and define $\delta_{pre} = (\delta_{-\underline{T}}, ...\delta_{-1})'$ and $\delta_{post} = (\delta_1, ...\delta_{\bar{T}})'$, with $\delta_0$ set to $0$.

**Bounding relative magnitudes restriction.** In practical applications, the emergence of confounding factors is somewhat inevitable. Thus in practical research settings, researchers may tend to assume that the influence of confounding factors on non-parallel trends in the post-treatment periods is not substantially greater in magnitude than the non-parallel trends resulting from confounding factors in pre-treatment periods. This situation can be defined as follows: $\delta \in \triangle^{RM}(\bar{M})$, $\bar{M} \geq 0$:

$$\triangle^{RM}(\bar{M}) = \{\delta : \forall t \geq 0, |\delta_{t+1} - \delta_t| \leq \bar{M} \cdot \max_{s<0} |\delta_{s+1} - \delta_s|\} \tag{16}$$

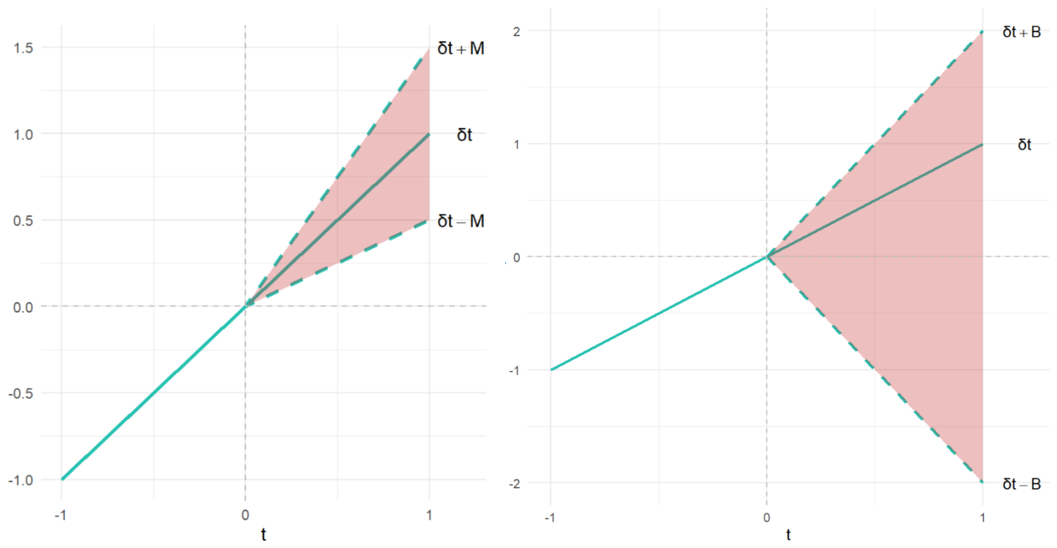
$\triangle^{RM}(\bar{M})$ serves as a restriction on the degree of post-treatment deviation from parallel trends, determined as $\bar{M}$ times the maximum divergence observed before treatment. This choice of restriction might be suitable if researchers suspect that deviations from parallel trends are driven by economic shocks of similar magnitude in both pre and post-treatment periods. For instances where the number of pre-treatment and post-treatment periods align, a typical benchmark could be $M$ = 1, restricting the post-treatment trend deviation to match the maximum observed in the pre-treatment period.

**Smoothness restrictions.** In other empirical situations, secular trends also exist within confounding factors. To address this concern, we can assume it is a linear time trend, just like the approach in Dobkin *et al.* (2018). However, using the concept of slope changes offers a more practical and reliable approach to accommodate these evolving trends. This situation can be formed as follows for $\delta \in \triangle^{SD}(M)$, $M \geq 0$:

$$\triangle^{SD}(M) = \{\delta : |(\delta_{t+1} - \delta_t) - (\delta_t - \delta_{t-1})| \leq M, \forall t\} \tag{17}$$

The parameter $M$ is responsible for regulating how much the slope of a trend can change in consecutive periods, thus it bounds the discrete analog of the second derivative. In the special case where $M = 0$, $SD(0)$ requires that the curve of $\delta$ be exactly linear, which aligns with the assumption commonly used in practical applications, similar to the approach taken in Bhuller *et al.* (2013) and Goodman-Bacon (2018), a linear trend is linear trend is utilized.

**Example of $\triangle^{SD}$ and $\triangle^{RM}$ in three periods:** In this example, three periods are considered, and I assume $\delta$ exhibits linearity concerning $T$. For the restriction on SD in trends, I set $M$ as $0.5$. Additionally, I illustrate $\delta$ under the constraint of relative magnitudes, assuming $\bar{M} = 2$, and determining $B$ as $\bar{M}$ multiplied by $max_{s<0}|\delta_{s+1} - \delta_s|$. In the accompanying graphics, the green line represents the assumed parallel trend violation, while the red area signifies the potential violation region under different restrictions.



(a) SD

(b) RM

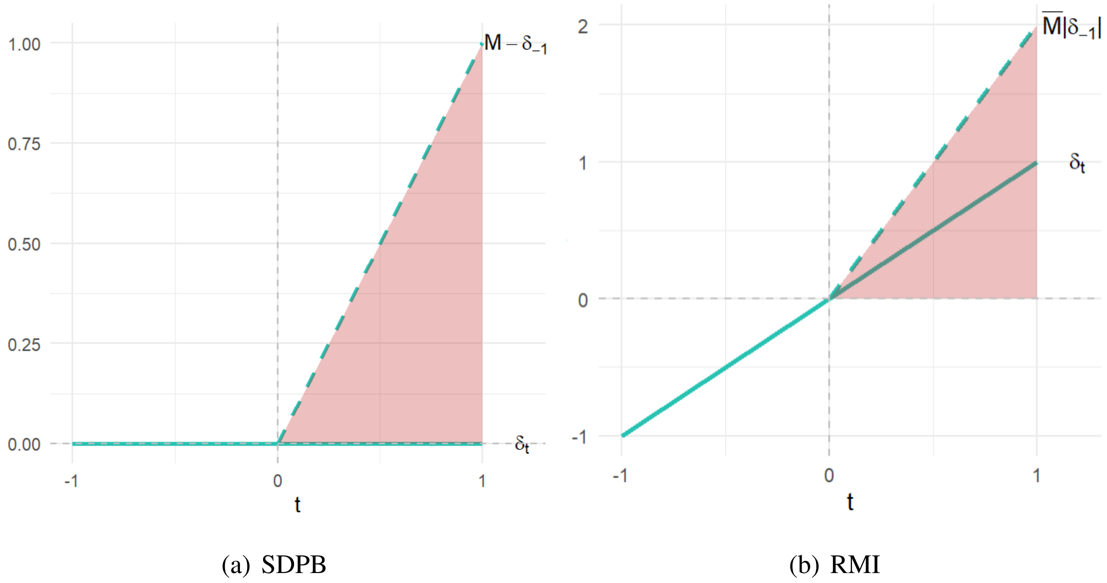
Figure 2: Intuition for $\triangle^{SD}$ and $\triangle^{RM}$

**Combining smoothness and relative magnitudes bounds.** In some contexts, it is necessary to satisfy the aforementioned two restrictions. In such instances, it might be justifiable to infer that the potential deviations from non-linearity in the post-treatment variation could be confined within the bounds of the observed deviations in the pre-treatment variation. This concept can be put into practice by implementing the following constraints:

$$\triangle^{SDRM}(\bar{M}) = \{\delta : \forall t \geq 0, |(\delta_{t+1} - \delta_t) - (\delta_t - \delta_{t-1})| \leq \bar{M} \cdot \max_{s<0} |\delta_{s+1} - \delta_s|\} \tag{18}$$

**Sign and monotonicity restrictions.** $\triangle^{PB}$ means there may be a positive sign of the differential trend in the post-treatment periods. The monotonicity restriction $\triangle^{Mon}$ means that the differential trend may be decreasing or increasing, we may then wish to impose the differential trend be increasing, $\delta \in \triangle^{I} := \{\delta : \delta_t \geq \delta_{t-1}, \forall \delta\}$. The sign and monotonicity restrictions can be combined with $SD$ and $RM$. For example, $\triangle^{SDPM} = \triangle^{SD} \cap \triangle^{PB}$, $\triangle^{RMI}(\bar{M}) := \triangle^{RM}(\bar{M}) \cap \triangle^{I}$.

**Example choices of $\triangle^{SDPB}$ and $\triangle^{RMI}$.** In the three-period difference-in-difference model, assuming the differential trend is exactly linear, this is equivalent to assuming $\triangle = \delta : \delta_1 = -\delta_{-1}$. When $\delta \in \Delta^{SD}$, we can determine the range of $\delta_1$ as follows: $\delta_1 \in [-\delta_{-1} - M, -\delta_{-1} + M]$. Thus, for $\delta \in \Delta^{SDPB}$, meaning $\delta_1$ must be a positive value, so the range of $\delta_1$ now is: $\delta_1 \in [0, -\delta_{-1} + M]$. Assuming $\delta_{-1} = \delta_0 = 0$ and $M = 1$, then we can calculate the exact range of $\delta_1$ when $\delta \in \triangle^{SDPB}$. Similarly, assuming $\triangle \in \triangle^{RM}(\bar{M})$ bounds the magnitude of $\delta_1$ based on the magnitude of $\delta_{-1}$, i.e., $\triangle^{RM}(\bar{M}) = (\delta_1, \delta_{-1})' : |\delta_1| \leq \bar{M}|\delta_{-1}|$. When we assume $\triangle \in \triangle^{RMI}$, the range changes to $\triangle^{RMI}(\bar{M}) = (\delta_1, \delta_{-1})' : \delta_1 \leq \bar{M}|\delta_{-1}|, \delta_{-1} \leq 0$, assuming $\delta_{-1} = -1, \delta_0 = 0$ and $M = 2$. Figure 3 provides a geometric depiction of $\triangle^{SDPB}$ and $\triangle^{RMI}$ in this example, the green line represents the assumed parallel trend violation, while the red area signifies the potential violation region under different restrictions.



(a) SDPB

(b) RMI

Figure 3: Example choices when $\triangle \in \triangle^{SDPB}$ and $\triangle \in \triangle^{RMI}$

Finally, We can summarize the above different kinds of assumptions as written as polyhedral type $\triangle$: $\triangle = \{\delta : A\delta \leq d\}$ for some known matrix $A$ and vector $d$. For example, when $\bar{T} = 1, \underline{T} = 1$, $\triangle^{SD}(M) = \{\delta : A^{SD}\delta \leq d^{SD}\}$, in which $A^{SD} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$, and $d^{SD} = (M, M)'$, this principle extends naturally when considering multiple periods before and after the treatment.

## 3.3 Conditional and Hybrid Inference

In this section, I will present an inference method in Rambachan & Roth (2023))with robust asymptotic properties under diverse restrictions. The inference method can handle different forms of $\triangle$, and it's based on the concept that making inferences about $\theta = l'\tau_{post}$ can be seen as a problem of testing a system of moment inequalities. These moments include a potentially extensive set of nuisance parameters entering linearly. Therefore, we implement this approach based on the conditional and hybrid methods introduced by Andrews *et al.* (2022)(referred to as ARP). By integrating a conditional approach from ARP, confidence sets with robust power properties across various parameter configurations can be generated.

Initiating the analysis, the null hypothesis can be formulated as $H_0$: $\theta = \bar{\theta}, \delta \in \triangle$. Tests can be constructed under this $H_0$ assumption, as long as the normal approximation accurately corresponds to the known variance matrix $\Sigma_n$. Define $L_{post} = [0, I]'$, and $\tau = L_{post}\tau_{post}$, $Y_n = A\hat{\beta}_n - d$, then we can transfer the null hypothesis to as follows, the specific reference process is shown in Appendix B.1:

$$H_0 : \exists \tau_{post} \in R^{\bar{T}} s.t. l'\tau_{post} = \bar{\theta}, \quad and \quad E_{\hat{\beta}_n \sim \mathcal{N}(\beta, \Sigma_n)}[Y_n - AL_{post}\tau_{post}] \leq 0. \tag{19}$$

Then for further testing, we can rewrite the above assumption to moments with unrestricted nuisance parameter $\tilde{\tau}$, $\tilde{\tau} \in R^{\bar{T}-1}$. $\Gamma$ is a square matrix with the vector $l'$ in the first row and remaining rows chosen so that $\Gamma$ has a full rank, we define $\tilde{A} = AL_{post}\Gamma^{-1}$, then $AL_{post}\tau_{post} = \tilde{A}\Gamma\tau_{post} = \tilde{A}\begin{pmatrix} \theta \\ \Gamma_{(-1,)}\tau_{post} \end{pmatrix}$, when $\bar{T} = 1$, then the $\tilde{\tau}$ is the 0-dimensional and should be interpreted as 0.

$$H_0 : \exists \tilde{\tau} \in R^{\bar{T}-1} \quad s.t. E[\tilde{Y}_n(\bar{\theta}) - \tilde{X}\tilde{\tau})] \leq 0 \tag{20}$$

Where $\tilde{Y}(\bar{\theta}) = Y_n - \tilde{A}_{(,1)}\bar{\theta}$, and $\tilde{X} = \tilde{A}_{(,-1)}$, $\tilde{A}_{(,1)}$ is the first column of matrix $\tilde{A}$, $\tilde{A}_{(,-1)}$ are the columns without first column of matrix $\tilde{A}$. So $\tilde{Y}(\bar{\theta})$ 's variance is $\tilde{\Sigma}_n = A\Sigma_n A'$ under the finite sample normal model.

**Example 3.1:** Assume $\bar{T} = 1$, $\tilde{\Sigma}_n = I$ and $\delta \in \triangle^{SD}$, so here the $H_0 : \exists \tau_1 \in R, s.t. \tau_1 = \bar{\theta}, E_{\hat{\beta}_n \sim \mathcal{N}(\beta, I)}[Y_n^{SD} - A^{SD}L_{post}\tau_1] \leq 0$. And $Y_n^{SD} = A^{SD}\hat{\beta}_n - d^{SD}$. There is no nuisance parameter because $\bar{T} = 1$. So here we can get that $\tilde{Y}(\bar{\theta})^{SD} = Y_n^{SD} - A^{SD}L_{post}\tau_1$.

### 3.3.1 Constructing Conditional and Hybrid Confidence Sets

In the above moment inequality test, the challenge arises due to the high dimensionality ($\bar{T} - 1$) of the nuisance parameter $\tau$ in settings with multiple post-treatment periods, thus we consider testing the moment inequalities using ARP's conditional and hybrid methods for computational efficiency. Furthermore, this analysis reveals that ARP's tests demonstrate strong asymptotic power under an LICQ condition.

Firstly, the construction process of the conditional confidence set is as follows. We can transfer the above $H_0$ to as follows, define $\tilde{\sigma}_n = \sqrt{diag(\tilde{\Sigma}_n)}$:

$$\hat{\eta} := \min_{\eta, \tilde{\tau}} \eta \quad s.t. \tilde{Y}_n(\bar{\theta}) - \tilde{X}\tilde{\tau} \leq \tilde{\sigma}_n \cdot \eta \tag{21}$$

By duality programming (e.g.,Schrijver (1998), Section 7.4)) we can transfer it to the following:

$$\hat{\eta} = \max_{\gamma} \gamma' \tilde{Y}_n(\bar{\theta}) \quad s.t. \gamma' \tilde{X} = 0, \gamma' \tilde{\sigma}_n = 1, \gamma \geq 0 \tag{22}$$

$\gamma_*$ is the optimal solution of the dual programming, standard results in linear programming imply that the optimum is always obtained at one of the finite set of vertices, $V(\Sigma_n)$. We denote by $\hat{V}_n \subset V(\Sigma_n)$ the set of optimal vertices of the dual program. The distribution of $\hat{\eta}$ has a truncated normal distribution conditioning on the optimal $\gamma_*$:

$$\hat{\eta} | \{\gamma_* \in \hat{V}_n, S_n = s\} \sim \xi | \xi \in [v^{lo}, v^{up}] \tag{23}$$

$\xi \sim \mathcal{N}(\gamma_*' \tilde{\mu}(\bar{\theta}), \gamma_*' \tilde{\Sigma}_n \gamma_*)$, $\tilde{\mu}(\bar{\theta}) = E[\tilde{Y}_n(\bar{\theta})]$, $S_n = (I - (\tilde{\Sigma}_n \gamma_* / \gamma_*' \tilde{\Sigma}_n \gamma_*) \gamma_*') \tilde{Y}_n(\bar{\theta})$, and $v^{lo}$, $v^{up}$ are functions of $\tilde{\Sigma}_n, s, \gamma_*$(see Lemma 1 of Andrews *et al.* (2022))

The $H_0$ shows that $\gamma_*' \tilde{\mu}(\bar{\theta}) \leq 0$, thus the critical value is $\max\{0, c_{C,\alpha}\}$, $c_{C,\alpha}$ is the $1 - \alpha$ quantile of the truncated normal distribution $\xi | \xi \in [v^{lo}, v^{up}]$ under the worst case assumption that $\gamma_*' \tilde{\mu}(\bar{\theta}) = 0$. We define $\psi_{\alpha}^{C}(\hat{\beta}_n, A, d, \bar{\theta}, \Sigma_n)$ as the indicator for whether the conditional test rejects the $H_0$.

**Example 3.2 following 3.1:** Following the example 3.1, we can get the expression of $\hat{\eta}$: $\hat{\eta} = max\tilde{Y}(\bar{\theta})^{SD})/\tilde{\sigma}_n$, so it is $max\tilde{Y}(\bar{\theta})^{SD}$. When $\delta \in \triangle^{SD}$ in the three periods, we can get the expression of the maximum $\eta \approx max\{\delta_1 + \delta_{-1} - M, -\delta_1 - \delta_{-1} - M\}$, according to the definition of the $\triangle^{SD}$, $\delta_1 + \delta_{-1} - M \leq 0$ and $-\delta_1 - \delta_{-1} - M \leq 0$, the inference process is in Appendix B.2. Here we can set the $v^{lo}$ as second biggest value of $\tilde{Y}(\bar{\theta})^{SD}$, and the $v^{up} = \infty$. Then we can get the truncated $\hat{\eta}$ distribution. Based on the critical value, we can check if we can reject the $H_0$ assumption based on $1 - \alpha$ significance level. So if the second biggest value is far smaller than the first biggest value, the conditional test tends to reject $H_0$. But we also need to consider that if the value of the maximum and second-largest sample moment is close, in this situation the condition test will have poor power. Thus in order to fix this problem, the hybrid test of ”ARP” is introduced.

We can construct a size $k$ hybrid test based on the least favorable (LF) assumption: $\gamma_*' \tilde{\mu}(\bar{\theta}) = 0$. APR shows that under the null hypothesis, $\hat{\eta}$'s distribution is bounded above in the worst assumption. Thus we define $0 < k < \alpha$, $c_{LF,k}$ is the critical value. If the first test is not rejected, then in the second stage, we construct a new size: $-((\alpha - k)/(1 - l))$. In particular, by similar logic as for the conditional confidence sets, we have that:

$$\hat{\eta} | \{\gamma_* \in \hat{V}_n, S_n = s, \hat{\eta} \leq c_{LF,k}\} \sim \xi | \xi \in [v^{lo}, v_H^{up}] \tag{24}$$

$v_H^{up} = \min\{v^{up}, c_{LF,k}\}$. We reject the hybrid test if the $\hat{\eta}$ is bigger than the critical value of the size $-((\alpha - k)/(1 - l))$. $\psi_{\alpha}^{C-LF}(\hat{\beta}_n, A, d, \bar{\theta}, \Sigma_n)$ is the indicator shows that if the test is rejected based on a fix $\bar{\theta}$. The confidence set for the hybrid test is: $C_{k,\alpha,n}^{C-LF}(\hat{\beta}_n, \Sigma_n)$.

### 3.3.2 Uniform Asymptotic Size Control and Consistency

Uniform asymptotic size control refers to the property where the size of a statistical test remains under control or within a specified limit across various scenarios or data-generating processes as sample sizes grow indefinitely large for all parameters we considered. Essentially, it ensures that the probability of incorrectly rejecting a null hypothesis (Type I error) stays bounded even when dealing with different or more complex datasets. In statistical inference, particularly when dealing with hypothesis testing or constructing confidence intervals, maintaining size control is crucial. When a test is said to have uniform asymptotic size control, it means that as the sample size increases or as the conditions change within a certain class of data-generating processes, the test's Type I error rate remains controlled, not escalating unreasonably. This property is significant because it ensures the reliability of the statistical test or procedure across diverse scenarios, making it more robust and dependable for concluding data, especially in complex or varying situations. We establish conditions ensuring size control within the finite sample normal model that extends uniformly across diverse data-generating processes $\mathcal{P}$ under the condition: $\hat{\beta}_n \sim \mathcal{N}$ and the $\Sigma_n$ is replaced by $\hat{\Sigma}_n$.

Here, we fix $\triangle = \{A\delta \leq d\}$ for $A$ with non-zero rows, $\triangle$ is non-empty. The data generation process is $P \in \mathcal{P}$, $\beta_P$ is the true $\beta$ of the data process $P$, and $\sqrt{n}(\hat{\beta}_n - \beta_P)$ is asymptotically normal, $\beta_P = \delta_P + L_{post}\tau_{P,post}$ for $\delta_P \in \triangle$ and $\tau_{P,post} \in R^{\underline{T}}$, $\theta_P := l'\tau_{P,post}$, for some fixed $l \neq 0$.

**Assumption 3.1** *Convergence in distribution is equivalent to convergence in bounded Lipschitz metric (see Theorem 1.12.4 in van der Vaart & Wellner (1996)). $f$ is the Lipschitz function bounded by 1 in absolute value and the constant bounded by* $1$. *So for this function, for* $\forall a, b, |f(a) - f(b)| \leq |a - b|$, *and* $|f(a)| \leq 1$. $BL_1$ *is the set of the function.*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{f\in BL_1} |E_P[f(\sqrt{n}(\hat{\beta}_n - \beta_P))] - E[f(\xi_P)]| = 0 \tag{25}$$

Where $\xi_P \sim \mathcal{N}(0, \Sigma_P)$, and $\beta_P = \delta_P + L_{post}\tau_{P,post}$ for $\delta_P \in \triangle$ and $\tau_{P,post} \in R^{\bar{T}}$.

**Assumption 3.2** *Let $S$ denote the set of matrices with eigenvalues bounded below by $\underline{\lambda} > 0$ and above by $\bar{\lambda} \geq \underline{\lambda}$, for all $P \in \mathcal{P}, \Sigma_P \in S$.*

**Assumption 3.3** *$\hat{\Sigma}_n$ is uniformly consistent for $\Sigma_P$: $\lim_{n\to\infty} \sup_{P\in\mathcal{P}} P_P(\|\hat{\Sigma}_n - \Sigma_P\| > \epsilon) = 0$, for all $\epsilon > 0$.*

**Assumption 3.4** *at least one of the following holds:*

*a. For $k_1 + k_2 = dim(\delta)$, the matrix $A$ can be written as $TQ$, where $Q$ has full rank and $T = \begin{pmatrix} I_{k1} & 0 \\ -I_{k1} & 0 \\ 0 & I_{k2} \end{pmatrix}$*

*b. let $\bar{\gamma}_1, ..., \bar{\gamma}_K$ be the elements of $V(I)$. Then for all k, either $\bar{\gamma}_k' A = 0$, or $\inf_{a\geq 0} \inf_{j\neq k} \|(\bar{\gamma}_k - a\bar{\gamma}_j)' A\| > 0$.*

The first assumption means that the only source of degeneracy for $A$ arises from matching inequalities of opposite signs. Degeneracy in the asymptotic distribution occurs when a statistical model or estimator fails to converge to a specific distribution, the estimator might converge to a degenerate or uninformative distribution, such as a point mass at a particular value or a distribution with zero variance. The degeneracy of matrix $A$ here means that the vectors in $A$ might be correlated. The second assumption means that $\forall k, \bar{\gamma}_k$ and $\bar{\gamma}_{j\neq k}$ are not linearly correlated. From which we can infer that $\bar{\gamma}_k \tilde{Y}_n$ and $\bar{\gamma}_j \tilde{Y}_n$ are not linearly correlated. So for the dual problem, the only source of the degeneracy for $\gamma' \tilde{Y}_n(\bar{\theta})$ is from matrix $A$.

**Proposition 3.1** *When the assumptions 3.1 - 3.4 hold, then we can get the conclusion:*

$$\lim_{n\to\infty} \sup_{n\to\infty} \sup_{P\in\mathcal{P}} E_P[\psi_a^C(\hat{\beta}_n, A, d, \theta_P, (1/n)\hat{\Sigma}_n)] \leq \alpha$$

$$\lim_{n\to\infty} \sup_{n\to\infty} \sup_{P\in\mathcal{P}} E_P[\psi_{k,a}^{C-LF}(\hat{\beta}_n, A, d, \theta_P, (1/n)\hat{\Sigma}_n)] \leq \alpha$$

Uniform asymptotic consistency means that as the sample size increases, a statistical test becomes more reliable in detecting true effects, ensuring that its power approaches 1 uniformly across all possible scenarios outside the null hypothesis. Next, to establish the uniform consistency of the conditional and hybrid tests, two additional assumptions are provided:

**Assumption 3.5** *Let $W_n = ((\hat{\beta}_n - \beta_P)', (vec(\hat{\Sigma}_n) - vec(\Sigma_P))')'$, the $vec(\sigma)$ is the vector of the elements of matrix $\Sigma$.*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{f\in BL_1} \|E_P[f(\sqrt{n}W_n)] - E[f(\xi_P^+)]\| = 0$$

$$\xi_P^+ \sim \mathcal{N}(0, V_P), V_P = \begin{pmatrix} \Sigma_P & V_{P,\beta\Sigma} \\ V_{P,\Sigma\beta} & V_{P,\Sigma} \end{pmatrix}$$

The $\Sigma_P$ is the true variance of the $\beta$, the $V_{P,\beta\Sigma}$ and $V_{P,\Sigma\beta}$ are the covariance of between $\beta$'s bias and vectorized $\Sigma_P$'s bias. The assumption means that the $\hat{\beta}_n$ and the $\hat{\Sigma}$ have a joint normal asymptotic distribution.

**Assumption 3.6** *For all $P \in \mathcal{P}, \Sigma_P \in S$ and the matrix $V_P$ defined in assumption 3.5 lies in a compact set $V$. Additionally, $(\Sigma_P - V_{P,\beta\Sigma} V_{P,\Sigma}^{\dagger} V_{P,\Sigma\beta})$ has eigenvalues bounded below by $\tilde{\lambda} > 0$, where $\dagger$ denotes the Moore-Penrose inverse.*

This assumption means that even though $\hat{\beta}_n$ and the $\hat{\Sigma}$ have a joint normal asymptotic distribution, $\hat{\beta}_n$ and $\hat{\Sigma}_n$ is not totally colinear, the proving process is in Appendix B.3. Under these assumptions, we obtain uniform consistency of the conditional and hybrid tests.

**Proposition 3.2** *Suppose the assumption $3.3 - 3.6$ hold, then we can get that for any $x > 0$ and $\alpha < 0.5$,*

$$\lim_{n\to\infty} \inf_{P\in\mathcal{P}} E_P[\psi_{\alpha}^{C}(\hat{\beta}_n, A, d, \theta_P^{ub} + x, (1/n)\hat{\Sigma}_n)] = 1$$

$$\lim_{n\to\infty} \inf_{P\in\mathcal{P}} E_P[\psi_{K,\alpha}^{C-LF}(\hat{\beta}_n, A, d, \theta_P^{ub} + x, (1/n)\hat{\Sigma}_n)] = 1$$

Where the $\theta_P^{ub} = supS(\beta_P, \triangle)$ is the upper bound of the identified set. The analogous result holds also for $\theta_P^{lb} - x$, $\theta_P^{lb} = infS(\beta_P, \triangle)$.

### 3.3.3 Optimal Local Asymptotic Power

The upper bound of the identified set is given by:

$$\theta^{ub}(\beta, \triangle) = l^{'}\beta_{post} - \Big((\min_{\delta} l^{'}\delta_{post}), s.t. A\delta \leq d, \delta_{pre} = \beta_{pre}\Big)$$

$\delta_{post} = \beta_{post} - \tau_{post}$, we can then rewrite the upper bound as the following, $A_{(,post)}$ contains the columns of $A$ corresponding to $\delta_{post}$:

$$\theta^{ub}(\beta, \triangle) = \max_{\tau_{post}} l^{'}\tau_{post}, s.t. - A_{(,post)}\tau_{post} \leq d - A\beta \tag{26}$$

The optimal $\tau_{post}$ is $\tau_{post}^{*}$, and $B^{*}$ is the indices of the binding constraints: $-A_{(B^{*},post)}\tau_{post}^{*} = d_{B^{*}} - A_{(B^{*},)}\beta$.

We define the LICQ holds in direction $l$ when there exists a solution $\tau^{*}$ fulfill the above binding restriction and $-A_{(B^{*},post)}$ has a full rank.

**Example 3.3:** In the three periods, assuming $\delta \in \triangle^{SD}$ and $M > 0$, so we can get the upper-bound as follows:

$$\theta^{ub}(\beta, \triangle) = \max_{\tau_1} \tau_1, s.t. - A^{SD}_{(,post)}\tau_1 \leq d^{SD} - A^{SD}\beta$$

We have identified $A^{SD}$ across three time periods, and we can organize the constraint, then we get the following results: $\delta_1 + \delta_{-1} \leq M$, and $\delta_1 + \delta_{-1} \geq -M$. When $M > 0$, LICQ can hold in this situation. But when $M = 0$, both the upper and lower bounds are binding, so LICQ is not satisfied.

For $\epsilon > 0$, we define $\mathcal{P}_\epsilon$ to be the set of the distribution $P \in \mathcal{P}$ with the above binding holds and the non-binding slack at least $\epsilon$: $-A_{(-B^*,post)}\tau^* < d_{B^*} - A_{(B_P)} - \epsilon$

Then we define $I(\triangle, \Sigma_n)$ as the collection of the confidence sets that control size in the finite sample model. The next proposition shows that as the sample size increases, the ability of the conditional test to detect deviations from the null hypothesis approaches the best possible performance while controlling for control size in a finite sample normal model.

**Proposition 3.3** *Suppose Assumption* 3.1 − 3.3 *hold,* $\theta^{ub}_P = supS(\beta_P, \triangle)$*, then for any* $\epsilon > 0, x > 0$ *and* $\alpha < 0.5$*,*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_\epsilon} \left| E_P \left[ \Psi^C_\alpha \left( \hat{\beta}_n, A, d, \theta^{ub}_P + 1/\sqrt{n}x, (1/n)\hat{\Sigma}_n \right) \right] - \rho^*_\alpha(P, x) \right| = 0$$

$$\rho^*_\alpha(P, x) = \lim_{n\to\infty} \sup_{C_{\alpha,n}\in\mathcal{I}_\alpha(\triangle,1/n\Sigma_P)} P_{\hat{\beta}_n\sim\mathcal{N}(\beta_P,(1/n)\Sigma_P)} \left( (\theta^{ub}_P + (1/\sqrt{n})x) \notin C_{\alpha,n} \right)$$

$\rho^*_\alpha(P, x)$ is the optimal local asymptotic power of a size $\alpha$ test in the finite sample normal model. And from the correlation of the LF-hybrid test and $(\alpha - k)/(1 - k)$, the local asymptotic power of the LF-hybrid is smaller than the value of $(\alpha - k)/(1 - k)$ test.

## 3.4 Inference using Fixed Length Confidence Internals (FLCIs)

A fixed-length confidence interval (FLCIs) is an interval estimate for a parameter in statistics where the width of the interval is predetermined or fixed. This type of confidence interval is particularly useful when we want to control the precision of the estimation. In a regular confidence interval, a confidence level (e.g., 95%) is specified and then the interval is calculated based on the data. However, in a fixed-length confidence interval, the desired width of the interval is settled beforehand, and the interval calculated will have a fixed width. In this chapter, We consider FLCIs based on the affine estimator, FLCIs can offer finite-sample power guarantees (in the normal model) for certain $\triangle$. We also show that the FLCIs perform well when $\triangle \in \triangle^{SD}$, but may perform poorly under other types of restrictions.

### 3.4.1 Constructing FLCIs

We denote the $C_{\alpha,n}(a, v, \gamma) := (a + v^{'}\hat{\beta}_n) \pm \chi$ as the affine estimator, in which $\chi$ is the half length confidence interval, and the $C_{\alpha,n}$ satisfied the convergence requirement (equation 15).

According $\hat{\beta} \sim \mathcal{N}(\beta, \Sigma_n)$, then $a + v^{'}\hat{\beta}_n \sim \mathcal{N}(a + v^{'}\beta, v^{'}\Sigma_n v)$, then when the value is absolute: $|a + v^{'}\beta - \theta| \sim |\mathcal{N}(a + v^{'}\beta - \theta, v^{'}\Sigma_n v)|$, define $b = a + v^{'}\beta - \theta$ as the bias between affine estimator and $\theta$, so when $|a + v^{'}\beta - \theta| \leq \gamma$, $\theta \in C_{\alpha,n}$. Define $\bar{b}$ is the affine estimator's worst-case bias:

$$\bar{b}(a, v) := \sup_{\delta \in \triangle, \tau_{post} \in R^{\bar{T}}} |a + v^{'}(\delta + L_{post}\tau_{post}) - l^{'}\tau_{post}| \tag{27}$$

We define $cv_\alpha(t)$ as the critical value of $1 - \alpha$ quantile of the folded normal distribution $|\mathcal{N}(t, 1)|$. The smallest $\chi$ satisfying the coverage requirement is :

$$\chi_n(a, v; \alpha) = \sigma_{v,n} \cdot cv_\alpha(\bar{b}(a, v)/\sigma_{v,n}) \tag{28}$$

$\sigma_{v,n} := \sqrt{v^{'}\Sigma_n v}$, $\chi$ is a function of $a, v$. So we can choose the $a$ and $v$ to minimize the $\chi_n(a, v; \alpha)$ to construct minimum-length FLCI, this minimization optimally trades off bias and variance, since the half-length $\chi_n(a, v; \alpha)$ is increasing in both the worse-case bias $\bar{b}$ and the variance $\sigma^2_{v,n}$(assuming $\alpha \in (0, 0.5]$). When $\triangle$ is convex, the minimization can be solved as a nested optimization problem, where both the inner and outer minimization are convex( see Armstrong & Kolesár (2018)). We denote the $C^{FLCI}_{\alpha,n}$ as the $1 - \alpha$ level, FLCI with the shortest length: $C^{FLCI}_{\alpha,n}(\hat{\beta}_n, \Sigma_n) := (a_n + v^{'}_n\hat{\beta}_n) \pm \chi_n$, $\chi_n := inf_{a,v}\chi_n(a, v; \alpha)$ and $a_n, v_n$ are the optimal values in the minimization.

### 3.4.2 Finite-Sample Near Optimality

The following result, a direct outcome of Armstrong & Kolesár (2018), bounds the ratio of the expected length of the shortest possible confidence interval relative to the length of the optimal FLCIs.

**Assumption 3.7** *Assume i) $\triangle$ is convex and centro-symmetric, and ii) $\delta_A \in \triangle$ is such that $(\delta - \delta_A) \in \triangle$ for all $\delta \in \triangle$.*

Not all $\triangle$ satisfy this assumption: for i) $\triangle^{SD}$ is satisfied, but not for $\triangle^{SDPM}$, $\triangle^{RM}$ and $\triangle^{RMI}$. For ii) it just holds when the $\delta_t$ is a linear trend for $\triangle^{SD}$, or $\delta_t = 0, \forall t$.

**Proposition 3.4** *Suppose $\delta$ and $\triangle$ satisfy Assumption 3.7. Let $I_a(\triangle, \Sigma_n)$ denote the class of the confidence sets that satisfy the coverage criterion at the $1-\alpha$ level, then for any $\tau$ with $\tau_{pre} = 0$ and $\Sigma_n$ positive definite, $\tilde{z}_\alpha = z_{1-\alpha} - z_{1-\alpha/2}$.*

$$\frac{\inf_{C_{\alpha,n} \in I_a(\triangle, \Sigma_n)} E_{\hat{\beta}_n \sim N(\delta+\tau, \Sigma_n)}[\lambda(C_{\alpha,n})]}{2\chi_n} \geq \frac{z_{1-\alpha}(1-\alpha) - \tilde{z}_\alpha \Phi(\tilde{z}_\alpha) + \phi(z_{1-\alpha}) - \phi(\tilde{z})}{z_{1-\alpha/2}}$$

When $\triangle = \triangle^{SD}$, $\alpha = 0.5$, the lower bound in the Proposition is 0.72, which means the length of the shortest CI that satisfies the coverage requirement is at most 28% less than the optimal FLCI.

### 3.4.3 Inconsistency of FLCIs

The discussed finite-sample guarantees do not hold for various restrictions, such as those involving sign and shape constraints. We demonstrate that the FLCIs may exhibit suboptimal performance in such scenarios, consider the following example:

**Example of $\triangle^{RMI}(\bar{M})$:** Suppose $\theta = \tau_1$. When $\triangle = \triangle^{RMI}(\bar{M})$ and $\bar{M} > 0$, then all affine estimators for $\tau_1$ have infinite worst-case bias, so the FLCI is the entire real line. To see this, following the lemma B.19 of Rambachan & Roth (2023), since $\tau_{post}$ is unrestricted in equation (27), the worst-case bias will be infinite if $v_{post} = e_1$. For $v_{post} = e_1$, the bias of the affine estimator $a + v'\hat{\beta}_n$ equals $|a + v'_{pre}\delta_{pre} + \delta_1|$ regardless of the value of $\tau$. By the triangle inequality, $max\{|a + v'_{pre}\delta_{pre} + \tilde{\delta}_1|, |a + v'_{pre}\delta_{pre} + \bar{\delta}_1|\} \geq \frac{1}{2}|\tilde{\delta}_1 - \bar{\delta}_1|$ for any two values $\tilde{\delta}_1$ and $\bar{\delta}_1$. The range of values of $\delta_1$ are feasible given a pre-treatment periods trend $\delta_{pre}$ can be made arbitrarily large by setting $\delta_{pre}$ such that $|\delta_{-1}|$ is large. Hence $\bar{b}(a, v) = \infty$ for all choices of $(a, v)$.

Next, a formal result on the (in)consistency of the FLCIs is presented. When $\triangle$ is convex, the identified set $S(\beta, \triangle)$ is an interval with a length defined as $\theta^{ub}(\beta, \triangle) - \theta^{lb}(\beta, \triangle) = b^{max}(\beta_{pre}, \triangle) - b^{min}(\beta_{pre}, \triangle)$. This length is determined solely by $\triangle$ and $\beta_{pre}$, so here we can denote the length as $LID(\beta_{pre}, \triangle)$. Our subsequent finding reveals that $C^{FLCI}_{\alpha,n}$ is reliable only when $LID(\beta_{pre}, \triangle)$ achieves its maximum value, provided that the identified set doesn't span the entire real line (in which case any procedure is trivially consistent).

**Assumption 3.8 (Identified Set Maximal Length and Finite)** *Suppose $\delta \in \triangle$ is such that: $LID(\delta_{pre}, \triangle) = \sup_{\tilde{\delta}_{pre} \in \triangle_{pre}} LID(\tilde{\delta}_{pre}, \triangle) < \infty$, where $\triangle_{pre} = \{\delta_{pre} \in R^{\underline{T}} : \exists \delta_{post} s.t.$*

$(\delta'_{pre}, \delta'_{post})' \in \triangle\}$ *is the set of positive values for* $\delta_{pre}$.

**Proposition 3.5** *Suppose* $\triangle$ *is convex and* $\alpha \in (0, 0.5]$. *Fix* $\delta \in \triangle$ *and* $\tau_A \in R^{\bar{T}}$, *and suppose* $S(\delta + \tau, \triangle) \neq R$. *Then* $(\delta, \triangle)$ *satisfies Assumption 3.8 if and only of* $C^{FLCI}_{\alpha,n}$ *is consistent, meaning that:*

$$\lim_{n\to\infty} P_{\hat{\beta}_n \sim \mathcal{N}(\delta+\tau,\Sigma_n)}(\theta_{out} \in C^{FLCI}_{\alpha,n}(\hat{\beta}_n, \Sigma_n)) = 0, \forall \theta_{out} \notin S(\delta + \tau, \triangle).$$

So if assumption 3.8 is not satisfied, then $C^{FLCI}_{\alpha,n}$ is not consistent, and it may include the fixed points outside the identified set. The lemma B.26 of Rambachan & Roth (2023) proves that the conditions of Proposition 3.4 imply that assumption 3.8 holds. Thus FLCIs achieve near-optimal results with a limited sample size, but this occurs only in some situations where they are consistent.

**Example of SDPB, RMI, and SD:** In the context of three-period difference-in-difference examples, the length of the identified set corresponds to the height of $\triangle$ in Figure 4. Consequently, Assumption 3.8 holds if and only if $\triangle$ is the largest height at $\delta_{-1}$. This assumption is satisfied for $\triangle^{SD}$, while for $\triangle^{SDPB}$, the sign restriction is not binding. However, for $\triangle^{RMI}$, the assumption does not hold. The graphic below illustrates these results: $\delta$ values are color-coded as light green (neither assumption holds), red (Assumption 3.8 holds), and Lime green (both Assumption 3.7 and 3.8 hold).

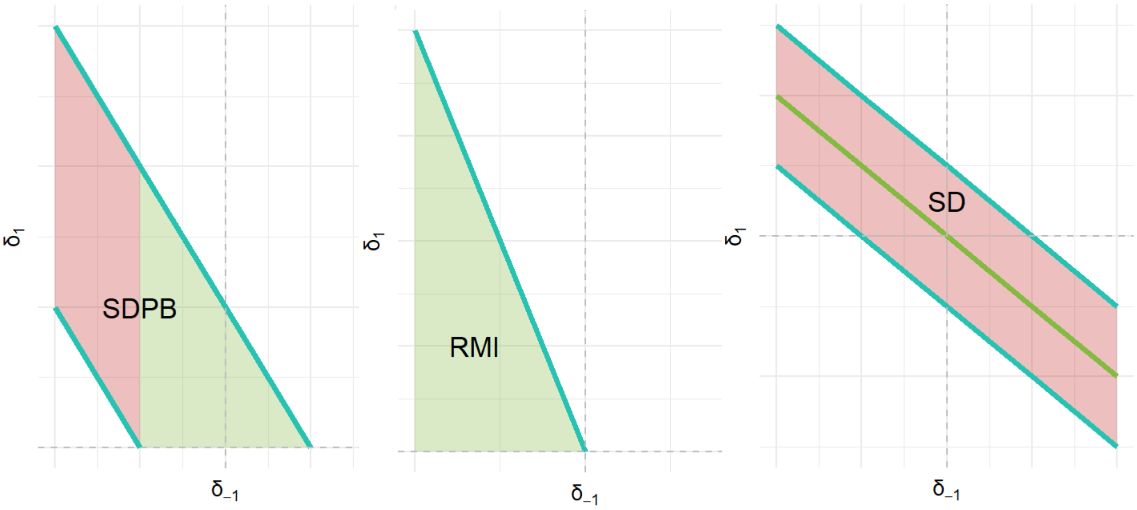


Figure 4: Diagram of Assumption 3.7 and 3.8 hold

# 4 Monte Carlo Simulation

In this section, I'll start by introducing the data-generating process, and provide a detailed exploration of the variables, parameters, and relationships that form the basis of my simulated dataset. Following this, I will illustrate the simulation process. Finally, I will present the simulation results, showing the key findings that align with the theory discussed in the previous sections.

## 4.1 Simulation Design

### 4.1.1 Data Generation Process

In the simulation, the non-staggered DiD model is used, assuming the number of the pre-treatment periods is $\underline{T}$, and the number of post-treatment periods is $\bar{T}$. The simulation DiD model is as follows:

$$Y_{it} = \lambda_i + \phi_t + \sum_{s \neq (\underline{T}+1)} \beta_s \times 1[t = s] \times D_i + \epsilon_{st}$$

$\epsilon_{st}$ is the error term, in the data generation process, I create the $\epsilon_{st}$ by generating random values from a normal distribution with a mean of 0 and a standard deviation of 1. About the choice of the standard deviation, in this context, using a very small standard deviation is inappropriate due to the methodology used in the HonestDiD package, specifically in relation to confidence interval calculation via test inversion over a grid. The default grid bounds are set as [lower bound of identified set under parallel trends - 20 times the standard deviation of the point estimate, the upper bound of identified set under parallel trends + 20 times the standard deviation of point estimate]. When the standard deviation is very small, the default grid bounds become very small. However, the parameter $\bar{M}$ typically takes values like 1, 2, or 3, which are considerably larger than the standard deviation. Consequently, the credible confidence interval should be relatively large. This creates a problem: calculating a relatively large confidence interval using a grid that is relatively small due to the small deviation value.

$\lambda_i$ is the individual-fixed effect, $\phi_t$ is the time-fixed effect. $\lambda_i$ is generated by drawing random values from a normal distribution with a mean of 0 and a standard deviation of 0.5. $\phi_t$ is created based on the given grid points and a mean of 0 using a mean-shift function. In assigning binary treatments, denoted as $D_i$, I randomly sample from the set $\{0, 1\}$ with replacement, ensuring each individual receives either 0 or 1 as treatment designation. True $\beta$ coefficients are defined based on different assumptions: under the parallel trends assumption, $\beta$ is defined as $\beta = c(0, 0, \ldots, 0)$. Under the pulse pre-trend assumption, $\beta$ is defined as $\beta = c(\text{rep}(0, \underline{T} - 1), 0.5, \text{rep}(0, \bar{T} + 1))$. Finally, the outcome variable $Y$ is generated using the above formula, capturing the intricate interplay of individual-fixed effects, time-fixed effects, treatment effects, and random errors.

The dataset is organized into a data frame, incorporating individual ID, years, treatment assignments, and the outcome variable. Subsequently, using TWFE to estimate the $\hat{\beta}$ and $\hat{\Sigma}$, the variance and covariance matrix of the estimated coefficient. In the Rambachan & Roth (2023), the $\hat{\beta}$ and $\hat{\sigma}$ are collected from the 12 recently published papers surveyed in Roth (2022). The most common criterion for assessing pre-trends typically involves ensuring that none of the coefficients from the pre-period are individually statistically significant. However, this criterion is not met among the 12 papers examined. There is at least one statistically significant pre-period coefficient in 3 of the 12 papers, and in 2 papers the pre-period coefficients are also jointly significant (see Table 2 in Roth (2022)).

In my data generation process, the true $\beta$ can be adjusted according to two different assumptions, leading to the production of different estimations of $\hat{\sigma}$ based on these assumptions. Consequently, my data generation process can yield more reliable estimates of $\hat{\beta}$ and $\hat{\Sigma}$ that align with the design assumptions.

The goal parameter $\theta$ is designed in two situations: first, as the causal effect in the first post-treatment period ($\theta = \tau_1$), and second, as the average post-treatment periods ($\theta = \bar{\tau}$).

### 4.1.2 Monte Carlo Simulation Process

In the simulation design, I consider four types of restrictions: $\triangle^{SD}(M)$,$\triangle^{SDPB}(M)$,$\triangle^{SDRM}(\bar{M})$ and $\triangle^{RM}(\bar{M})$. $\tau$ is set as 0 in the simulation, so $\beta = \delta$. Two simulation scenarios are considered: one where $\delta$ is set to 0, ensuring parallel trends, and another where $\delta_{-1}$ takes a non-zero value while other $\delta$ parameters are set to 0 (pulse pre-trend).

For the given $\triangle$, in the simulation phase, the identified set $LID(\beta^*, \triangle)$ is computed for each restriction type. Subsequently, the expected excess length is determined, and the optimal bound on excess length is calculated. The excess length ratio (efficiency ratio) is then derived by the optimal bound on the excess length divided by the expected excess length.

The expected excess length is calculated by the difference between the length of the identified set and the estimated confidence length by the HonestDiD package. Here is the link to the introduction of the HonestDiD: (https://github.com/asheshrambachan/HonestDiD). The computation process of the optimal bound on the excess length of confidence intervals that satisfy the uniform coverage requirement is in the Appendix B.4:

Three methods in my simulation are used for constructing confidence intervals: FLCIs, conditional confidence sets, and conditional-least favorable hybrid confidence sets. Specifically, for the conditional-least favorable hybrid confidence sets, I utilize a first-stage least-favorable test with a size of $k = \alpha/10$. The units of $M$ and $\delta_{-1}$ are normalized by the standard deviation of $\hat{\beta}_1$. When considering $\triangle^{RM}$ and $\triangle^{SDRM}$, the results of FLCIs are excluded due to the infinite length of FLCIs under these two restrictions. Table 1 presents the anticipated properties based on different $\triangle$ and assumptions when both $M$ and $\bar{M}$ are non-zero.

Table 1: *Summary of expected properties for each simulation design*

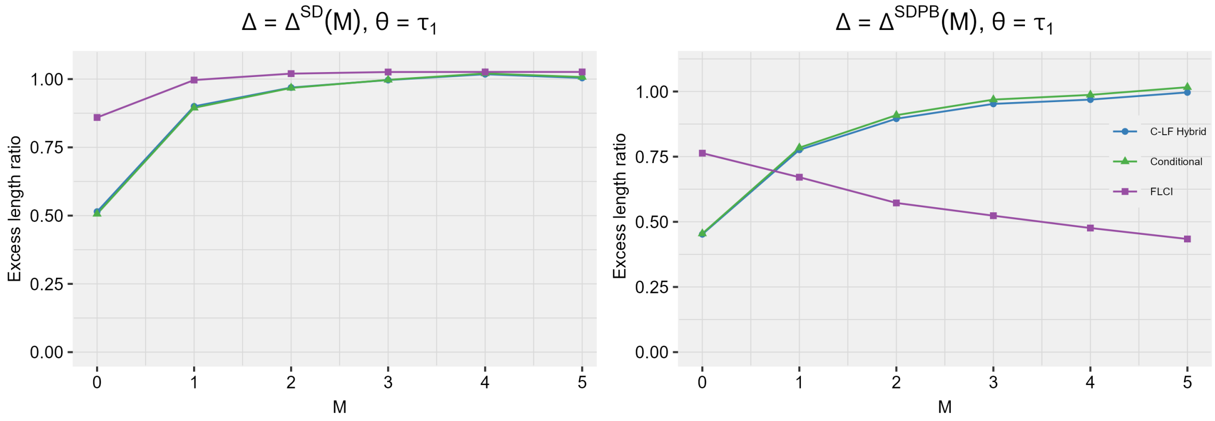
| | Parallel Trends | | Pulse Pre-Trend | |
|---|---|---|---|---|
| | $\triangle^{SD}(M)$ | $\triangle^{SDPB}(M)$ | $\triangle^{SDRM}(\bar{M})$ | $\triangle^{RM}(\bar{M})$ |
| **Conditional and Hybrid** | | | | |
| Consistent | ✓ | ✓ | ✓ | ✓ |
| Asymptotically (near-) optimal | ✓ | ✓ | ✓ | ✗ |
| **FLCI** | | | | |
| Consistent | ✓ | ✗ | ✗ | ✗ |
| Finite-sample near-optimal | ✓ | ✗ | ✗ | ✗ |

## 4.2 Simulation under Parallel Trends

In this section, I will show the simulation results under parallel trends assumption and the $\triangle$ restriction including $\triangle^{SD}$ and $\triangle^{SDPB}$. First, the simulation results of the two restrictions when $\theta = \tau_1$ are shown.

Figure 5 illustrates the simulation results for both $\triangle^{SD}$ and $\triangle^{SDPB}$ under the assumption of a parallel trend and zero treatment effects. The results of FLCIs are represented in purple, the results of conditional-least favorable hybrid confidence sets are in blue, and the conditional confidence sets results are in green.

The left panel provides an overview of the excess length ratio when $\triangle = \triangle^{SD}$ across various values of $M$. A noticeable trend emerges, with all three curves exhibiting an increased trend. The performance of conditional and conditional-least favorable hybrids confidence sets are similar. Notably, when $M \geq 4$, the ratios for all three methods converge to 1, indicating the asymptotic (near-)optimality. To delve into specifics, let's check the behavior for different



$\Delta = \Delta^{SD}(M), \theta = \tau_1$

$\Delta = \Delta^{SDPB}(M), \theta = \tau_1$

Figure 5: Simulation results for $\triangle^{SD}$ and $\triangle^{SDPB}$ when $\theta = \tau_1$: excess length ratio

ranges of $M$: the ratio of FLCI surpasses that of both conditional and conditional-least favorable hybrid methods for all $M$, implying finite-sample near-optimality under Assumptions 3.7 in Section 3. Despite the ratios of conditional and conditional-least favorable hybrid methods approaching 1 when $M \geq 4$, the ratio for these methods is around $50\%$ when $M = 0$. This peculiarity stems from the fact that, at $M = 0$, both upper and lower bounds for $\triangle^{SD}$ are zero, leading to the non-satisfaction of LICQ.

The right panel of Figure 5 provides insights into the excess length ratio when $\triangle = \triangle^{SDPB}$ across a range of $M$ values. Notably, the trends for the conditional and conditional-least favorable hybrid methods showcase an increase that eventually converges to 1 with the growing values of $M$, indicating asymptotic (near-)optimality in the simulation results. In contrast, the FLCI curve demonstrates a decreasing pattern as $M$ increases, revealing that FLCIs lack consistency in this simulation design for $M > 0$. Additionally, it is observed that the excess length ratios of the conditional method are slightly larger than those of the conditional-least favorable hybrid method as $M$ increases.

Then I present the additional results using the average of post-period treatment effects, $\theta = \bar{\tau}$ as the target parameter when $\triangle = \triangle^{SD}$ and $\triangle^{SDPB}$.

Figure 6 shows the results of the excess length ratio for $\theta = \bar{\tau}$ concerning both $\triangle^{SD}$ and $\triangle^{SDPB}$. Similar to the simulations conducted for $\theta = \tau_1$, these analyses are carried out under the assumption of parallel trends and zero treatment effects ($\tau = 0$) based on various $M$ values. The figure shows that the conclusions align with those observed for $\theta = \tau_1$. But there are still some differences:

In the left panel of Figure 6, the ratios are presented when $\triangle = \triangle^{SD}$. The trends of the three curves remain consistent with the observed trends in the previous analysis. When $M \geq 3$,

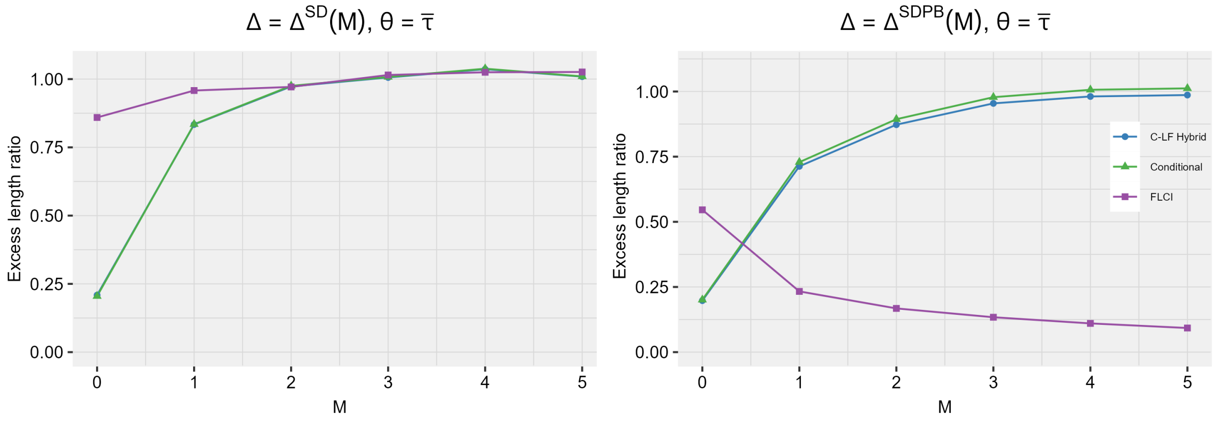


Figure 6: Simulation results for $\triangle^{SD}$ and $\triangle^{SDPB}$ when $\theta = \bar{\tau}$: excess length ratio

the ratios for all three methods converge to 1, indicating the asymptotic (near-)optimality of this design. When $M \leq 2$, the performance of FLCIs is best compared to conditional confidence sets and conditional-least favorable hybrid methods. However, a notable distinction from Figure 5 is observed when $M = 0$: in this scenario, the ratios of conditional and conditional-least favorable hybrid methods are smaller compared to the case where $\theta = \tau_1$.

On the right panel of Figure 6, the ratios are displayed when $\triangle = \triangle^{SDPB}$. The trends of the three curves mirror those observed trends in the analysis where $\theta = \tau_1$. Similarly, the only notable difference occurs when $M = 0$, with the ratios of conditional and conditional-least favorable hybrid methods being smaller than the ratios in $\theta = \tau_1$.

## 4.3 Simulation under Pulse Pre-Trend

In this simulation, I will show the simulation results under pulse pre-trend when $\triangle = \triangle^{RM}(\bar{M})$ and $\triangle = \triangle^{SDRM}(\bar{M})$. Firstly, $\bar{M}$ is set as 1 for both $\triangle^{RM}$ and $\triangle^{SDRM}$, and I will show the simulation results when $\theta = \tau_1$:

In Figure 7, the simulation results for $\triangle^{SDRM}$ and $\triangle^{RM}$ are depicted, assuming a pulse pre-trend and zero treatment effects. The FLCIs are excluded from the presentation due to their infinite length. The outcomes of the conditional-least favorable hybrid confidence sets approach are denoted in blue, while the conditional confidence sets results are presented in green.

In the left panel, an analysis of the excess length ratio is presented for $\triangle = \triangle^{SDRM}$ across different values of $\delta_{-1}$. A trend is evident, with both curves demonstrating an upward trend. The performance of the conditional and conditional-least favorable hybrid methods appears similar. Notably, as $\delta_{-1} \geq 2$, the ratios for both methods converge to 1, signaling the asymptotic (near-)optimality of this design.

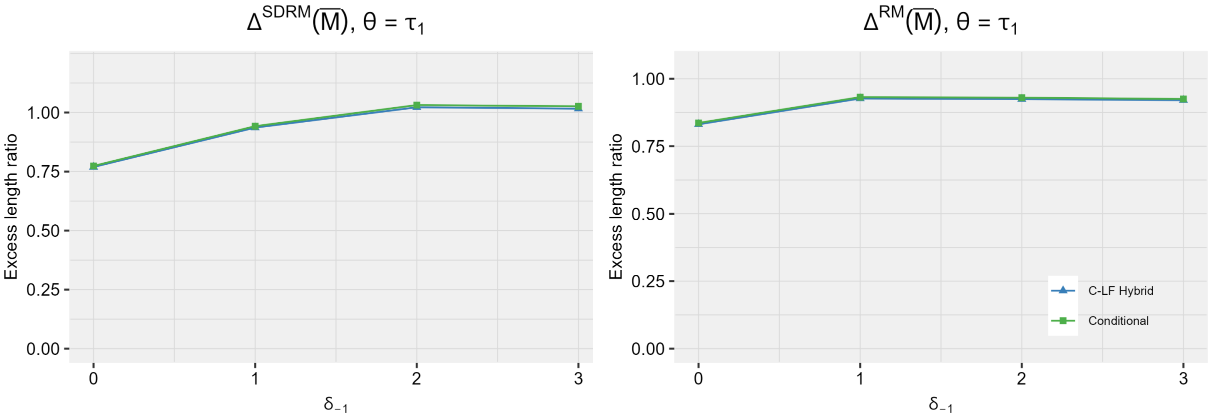


Figure 7: Simulation results for $\triangle^{SDRM}$ and $\triangle^{RM}$ when $\theta = \tau_1$: excess length ratio

In the right panel of Figure 7, the excess length ratio is shown for $\triangle = \triangle^{RM}$ over varying $\delta_{-1}$ values. It is observed that both curves for the two methods exhibit an upward trend as $\delta_{-1}$ increases. However, the ratios do not reach 1, indicating a lack of asymptotic (near-)optimality results for this design. The excess length ratio of the two methods stays around at $92\%$ with the increasing of the $\delta_{-1}$, which means that these procedures can have access length within $8\%$ of the optimal bound in cases when LICQ does not hold.

Then I report additional results using the average of post-period treatment effects, $\theta = \bar{\tau}$ as the target parameter:

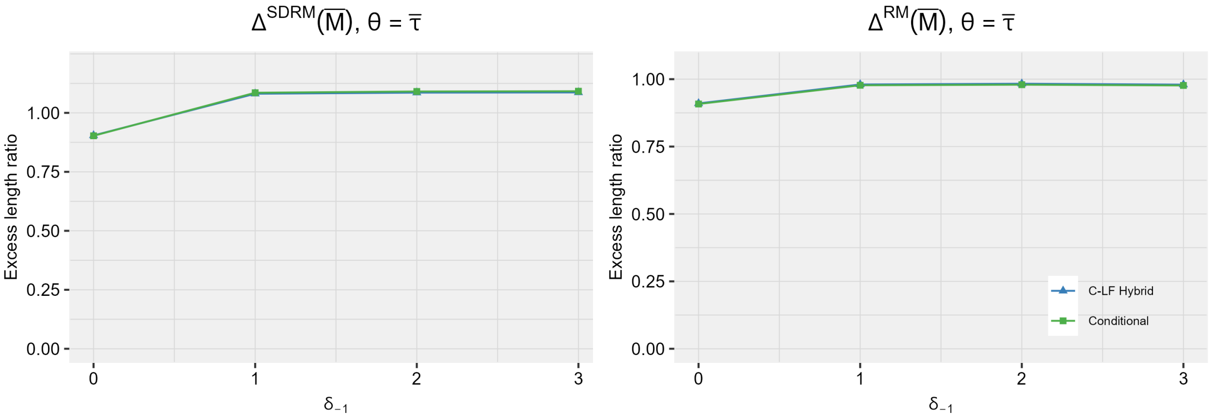


Figure 8: Simulation results for $\triangle^{SDRM}$ and $\triangle^{RM}$ when $\theta = \bar{\tau}$: excess length ratio

In the left panel of Figure 8, the excess length ratio is graphed for both the conditional and conditional-least favorable hybrid confidence sets, with $\delta_{-1}$ being the variable of interest and $\triangle$ set as $\triangle^{SDRM}(\bar{M})$. It's noteworthy that when $\delta_{-1} \geq 2$, the excess length ratios are slightly bigger than 1, signifying an asymptotic (near-)optimality result for this specific design.

In the right panel of Figure 8, the excess length ratio is presented for both the conditional and conditional-least favorable hybrid confidence sets, considering $\delta_{-1}$ while having $\triangle$ set as $\triangle^{RM}(\bar{M})$. Both curves display an upward trend as $\delta_{-1}$ increases. However, it's crucial to note that these ratios do not reach 1, indicating an absence of asymptotic (near-)optimality results for this design. The excess length ratios for both methods consistently hover around $98\%$ with increasing $\delta_{-1}$. This implies that these procedures maintain an excess length within $2\%$ of the optimal bound in cases where LICQ does not hold.

The aforementioned results were obtained with $\bar{M} = 1$. Now, I will modify $\bar{M}$ to 3 and observe the variations in the results. Figure 9 illustrates the simulation results of $\triangle^{RM}$ and $\triangle^{SDRM}$, where $\theta = \tau_1$ serves as the target parameter.

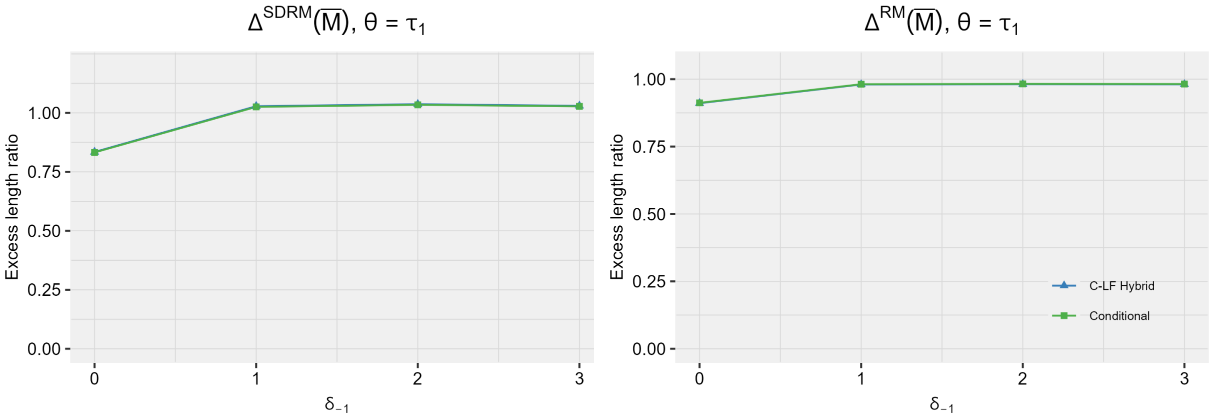


Figure 9: Simulation results for $\triangle^{SDRM}$ and $\triangle^{RM}$ when $\bar{M} = 3$: excess length ratio

The outcomes exhibit similarity with $\bar{M} = 1$, indicating that the selection of $\bar{M}$ does not seem to exert a significant influence on the effectiveness of our proposed procedures.

## 4.4 Key Points from Simulations

The simulation results above demonstrate that under the parallel trend assumption, both the conditional confidence sets and the conditional-least favorable hybrid confidence sets perform well. FLCIs also exhibit strong performance when $\triangle = \triangle^{SD}$, but FLCIs perform poorly for $\triangle = \triangle^{SDPB}$. Under the pulse-parallel trend assumption, the performance of conditional confidence sets and conditional-least favorable hybrid are well when $\triangle = \triangle^{SDRM}$. when $\triangle = \triangle^{RM}$, for both methods, this procedure can have an excess length within a certain degree of the optimum even in cases where LICQ fails. Moreover, the increase of $\bar{M}$ does not significantly influence effectiveness.

Therefore, it can be concluded that conditional confidence sets and conditional-least favorable hybrid confidence sets perform well, and FLCIs exhibit optimal performance when

$\triangle = \triangle^{SD}$. So in practice, when $\triangle = \triangle^{SD}$, utilizing FLCIs will yield robust confidence sets. Under other restrictions, conditional-least favorable hybrid confidence sets are a preferable approach, as discussed in Section 3.

The conclusions align with those of the original paper despite differences in the data generation process. In the original paper, the efficiency ratio is determined by taking the median of the 12 excess length ratios. Taking the median helps to mitigate the influence of outliers. In my simulation, the data generation process adheres to the (non)parallel assumptions made in the study, resulting in more stable results, and reinforcing the validity of the findings.

During the simulation process, to replicate the simulation of the original paper, some points need to be considered: i)The number of post-treatment periods should be large. Theoretically, the conditional-least favorable hybrid confidence sets perform better than conditional confidence sets in constructing robust confidence intervals when $\bar{T}$ is larger. Therefore, setting a larger $\bar{T}$ allows for thorough testing of this aspect. ii)The choice of the error term's standard deviation should consider the value of $\bar{M}$. If the standard deviation is fixed and small, it's important to adjust the grid bounds in the replication package accordingly to ensure accurate results iii) The values of $M$ and $\delta_{-1}$ are normalized in the replicate package of the original paper. If there's a preference for a non-normalized version, alterations can be made accordingly.

# 5 Application

In this chapter, I will apply the aforementioned method to estimate the treatment effect while integrating diverse pre-trend restrictions into the model. Subsequently, a sensitivity analysis will be conducted to monitor the length of the confidence interval and assess the breakdown point.

## 5.1 Estimating the Effect of Increasing Minimum Wage

Callaway & Sant'Anna (2021) investigate the identification, estimation, and inference procedures for treatment effect parameters using DiD with multiple time periods in the context of the minimum wage policy's impact on teen employment from 2001 to 2007. In this study, I utilize a subset of the data employed by Callaway and Sant'Anna to estimate confidence intervals based on different values $\bar{M}$. The formula used for this estimation is:

$$Y_{it} = \sum_{s \neq 2004} \beta_s \times 1[t = s] \times D_i + \phi_t + \lambda_i + \epsilon_{it}$$

$Y_{it}$ is the log of county-level teen employment in state $i$ in period $t$. It is the outcome variable. $D_i$ is an indicator of whether the state $i$ get the treatment. $\phi_t$ is the fixed effect for the year, $\lambda_i$ is the fix effect for state. The dataset comprises 500 observations of county-level teen employment rates spanning the years 2003 to 2007. We specifically selected counties that experienced a policy treatment in 2004. We proceed to estimate the effect on the county-level teen employment using TWFE:

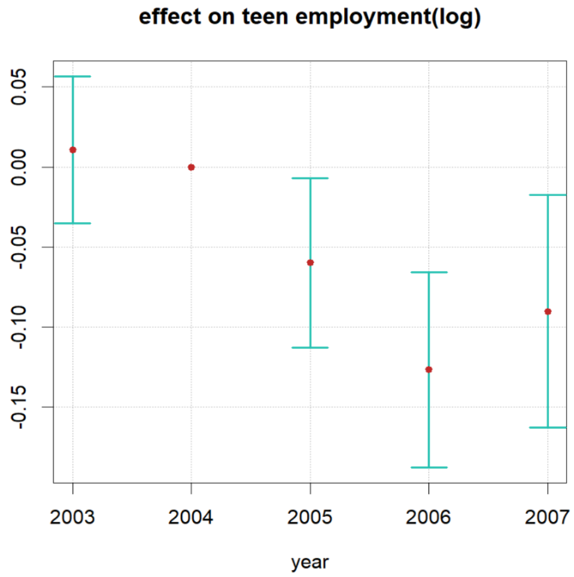
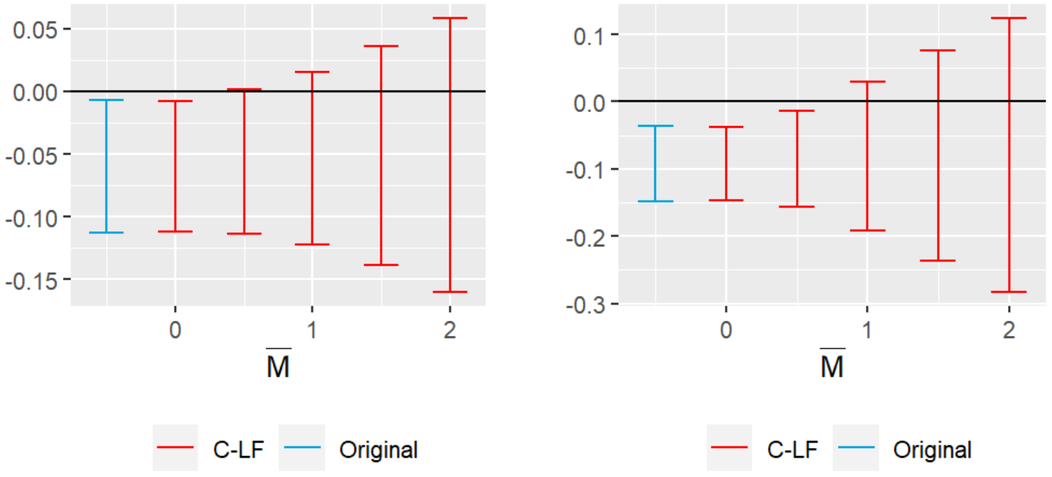


Figure 10: Event study coefficient by TWFE of Callaway and Sant'Anna

The figure above displays the event coefficient estimations using TWFE. From the figure, it is evident that the pre-trend is violated as indicated by the non-zero value of $\beta_{2003}$. Additionally, the magnitude change in the post-period $\hat{\beta}$ is larger than that observed in the pre-treatment periods.

In addition to the minimum wage policy, several other factors also play a role in influencing the log of county-level teen employment. These encompass economic conditions, education policies, and the impact of technology and automation, among others. Considering the potential influence of other macroeconomic events on the log of county-level teen employment, I incorporate the $\triangle^{RM}$ restriction for $\delta$ to account for these dynamics. The parameter $\bar{M}$ varies from 0 to 2, where $\bar{M} = 1$ indicates that post-period violations of parallel trends do not exceed the maximal pre-treatment period violation. The sensitivity analysis is conducted for different $\theta$ values: $\theta = \tau_{2005}$ and $\theta = \bar{\tau}$. Appendix C.1 provides a detailed overview of robust conditional confidence set results for $\delta = \triangle^{RM}(\bar{M})$. The information includes the lower and upper bounds, the values of $\bar{M}$, and the methodology employed: conditional-least favorable hybrid (C-LF) method.



(a) $\theta = \tau_{2005}$

(b) $\theta = \bar{\tau}$

Figure 11: Senstivity Analysis of RM for Callaway & Sant'Anna (2021)

The figure above illustrates robust confidence intervals for various values of $\bar{M}$. In the left panel, the confidence sets for the first year after treatment are estimated with different $\bar{M}$ values. The original interval, calculated through OLS estimation which is only valid under parallel trends is given as $[-0.113, -0.007]$. Notably, the 'breakdown value' for a significant effect is observed at $\bar{M} = 0.5$. This suggests that the significant result remains robust when allowing for violations of parallel trends up to half as large as the maximum violation in the pre-treatment period. Therefore, to maintain a significant confidence interval result, post-period violations should not exceed half of the maximal pre-treatment periods violation.

The right panel illustrates another scenario with $\theta = \bar{\tau}$. In this case, the 'breakdown value' is observed at $\bar{M} = 1$, which is larger than the corresponding value for $\theta = \tau_{2005}$. The CI of the original (OLS) now stands at $[-0.148, -0.037]$, which is bigger than the $\tau_{2005}$. Besides, the CI at different $M$ is also bigger when looking at $\bar{\tau}$ than $\tau_{2005}$, this difference is attributed to the RM restriction($\delta : \forall t \geq 0, |\delta_{t+1} - \delta_t| \leq \bar{M} \times \max_{s<0} |\delta_{s+1} - \delta_s|$), it should bind the violation of parallel trends across consecutive periods. As a consequence, the values of $\delta$ in later periods tend to be larger, leading to a broader identified set. Therefore, opting to constrain the magnitude of economic shocks based on the maximum value in the pre-treatment period generally results in broader confidence sets for parameters associated with subsequent periods.

## 5.2 Estimating the Fiscal Cost of Hurricanes

Deryugina (2017) investigates US hurricanes' influence on substantial increases in non-disaster government transfers, such as unemployment insurance and public medical payments, in affected counties in the decade after a hurricane. The model is as follows:

$$O_{ct} = \sum_{\tau=-10,\tau\neq-1}^{10} \beta_\tau H_{c\tau} + \alpha_t + \alpha_c + X_{c,1969}^{'}\alpha_t + \beta_{-11}H_{c,-11} + \beta_{-11}H_{c,-11} + \epsilon_{ct} \tag{29}$$

$O_{ct}$ denotes the per capita government transfers (log) outcome in the country $c$ at time $t$. The variable $H_{c\tau}$ serves as the hurricane indicator, taking the value of 1 if country $c$ experienced a hurricane $\tau$ years ago. Time fixed effect is represented by $\alpha_t$, while country fixed effect is denoted as $\alpha_c$. Additionally, the term $X'_{c,1969}\alpha_t$ allows for variation in the year-fixed effect based on linear 1969 characteristics. Finally, $H_{c,-11}$ and $H_{c,11}$ are indicators equal to 1 if the country experienced a hurricane before or after the time periods of interest.

Another point needs to be considered: results using year-by-year hurricane indicators are very similar, but noisier. Hence, to enhance statistical power, hurricane indicators are aggregated into two-year bins. The combined lags considered are $\tau = 1$ and 2, 3 and 4, 5 and 6, 7 and 8, as well as 9 and 10. In the context of this analysis, the reference year, denoted as year 0, corresponds to the year when a hurricane makes landfall in a country. For the pre-treatment periods, it is assumed that the effects of lead 1 and lead 2 are both zero. Therefore, the estimated coefficients should be interpreted as changes relative to the two years preceding the hurricane's occurrence. In the final data, the number of pre-treatment periods is 4, and the number of post-periods is 6.

The event coefficient estimation ($\hat{\beta}_\tau$) derived from the model formula are presented in Figure 12. Upon examination of the figure, it is shown that the values of $\hat{\beta}_{-5}, \ldots, \hat{\beta}_{-2}$ are not zero, thereby indicating a deviation from the parallel trend assumption ($\beta_{pre} = 0$). Additionally, $\hat{\beta}_\tau$ exhibits a decreasing trend during the pre-treatment periods and an increasing trend in the post-treatment periods. Besides, it's noteworthy that the magnitude of the estimated coefficient surpasses that of any of the pre-trend coefficients.

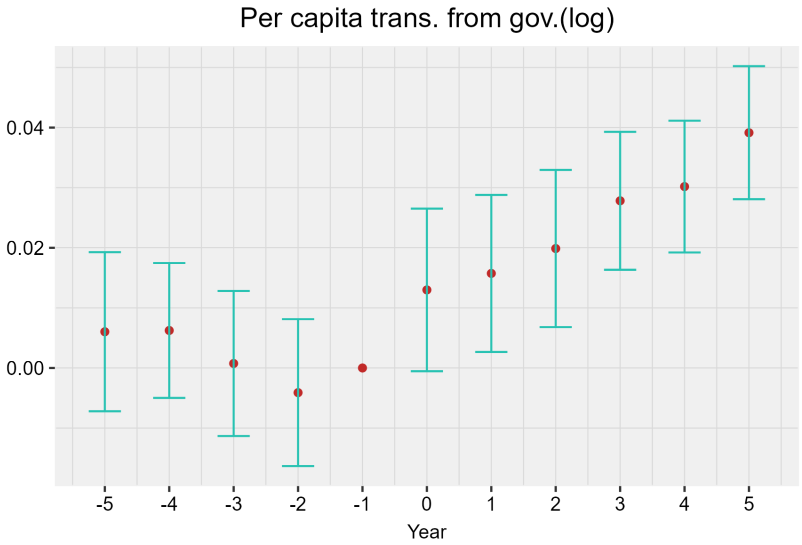
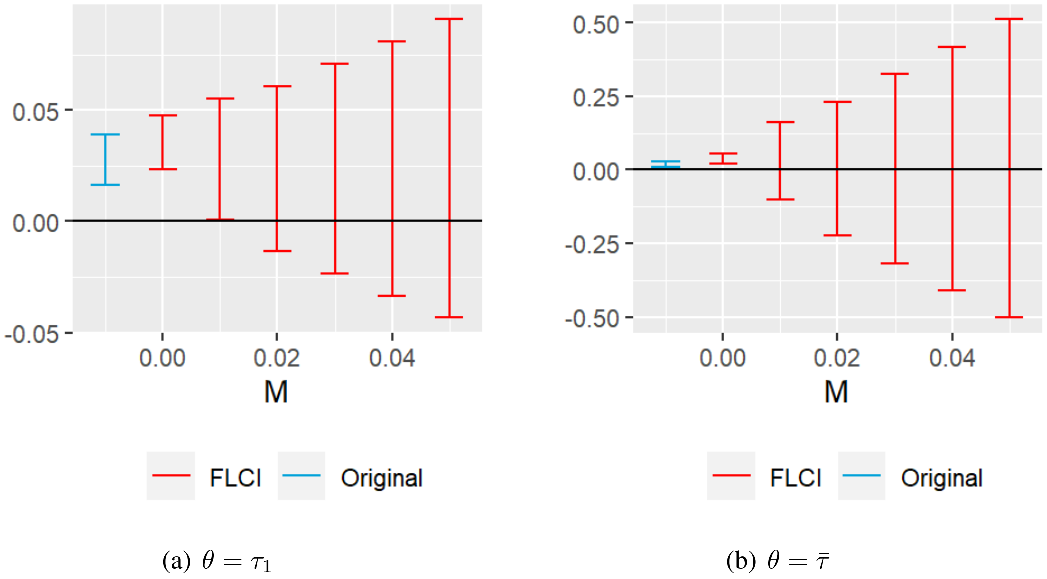


Figure 12: Event study coefficient of Deryugina (2017)

I did the sensitivity analysis when $\triangle = \triangle^{SD}$: $M = [0, 0.01, 0.02, ...0.05]$, and two scenarios were considered for $\theta$: $\theta = \tau_1$ and $\theta = \bar{\tau}$. Within the SD constraint, a setting of $M = 0$ implies a linear trend for $\delta$, while an increase in $M$ corresponds to a bigger deviation from linearity. The results are visualized in sensitivity analysis in Figure 13. Appendix C.2 presents comprehensive conditional confidence set information, including the lower bound, upper bound, the value of $M$, and the employed method.



(a) $\theta = \tau_1$

(b) $\theta = \bar{\tau}$

Figure 13: Senstivity Analysis of SD

Figure 13 illustrates the robust confidence sets for the treatment effect corresponding to different values of $M$ when $\theta = \tau_1$ and $\theta = \bar{\tau}$. The blue confidence interval is estimated using OLS, while the red confidence intervals are estimated using FLCIs. The length of the confidence interval for OLS is a constant: [0.0163, 0.0393], whereas the length of the FLCIs grows progressively larger with increasing values of $M$.

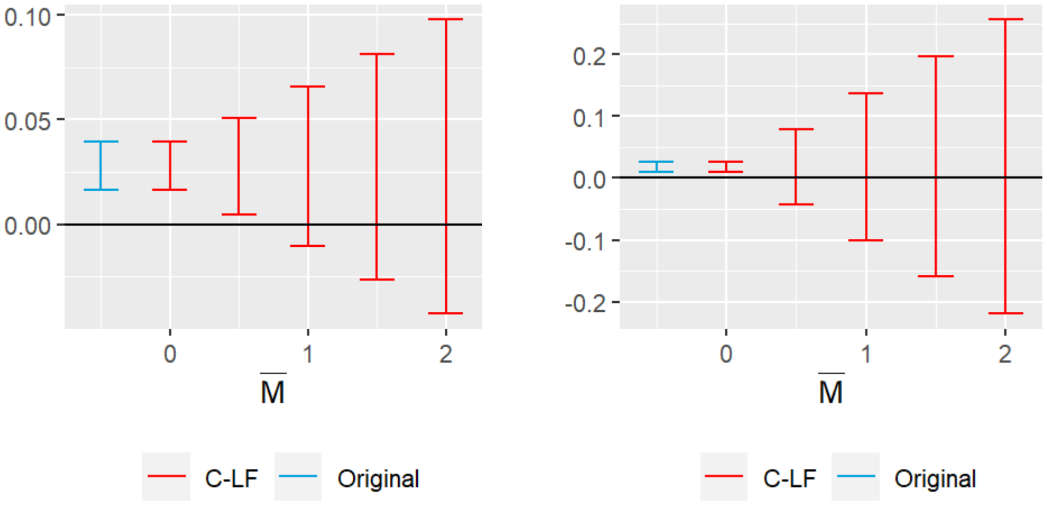
The left panel of Figure 13 describes the confidence sets for treatment effect when $\theta = \tau_1$. The confidence interval under $M = 0$ is slightly bigger than the confidence interval length of OLS. With the growth of the $M$, the difference in the confidence intervals between FLCIs and OLS becomes more pronounced. The breakdown point is observed for the significant effect, occurring at $M = 0.02$, this implies that when $M \geq 0.02$, the estimated confidence interval is no longer statistically significant.

The right panel of Figure 13 shows the confidence sets for the treatment effect when $\theta = \bar{\tau}$. When $M = 0$, the confidence intervals for OLS and FLCIs appear similar, and the confidence interval when $M = 0$ closely resembles that when $\theta = \tau_1$. However, as $M$ increases, the difference between OLS and FLCIs becomes bigger compared to the scenario where $\theta = \tau_1$. The breakdown point is observed to occur earlier, specifically at $M = 0.01$, which implies that when $M \geq 0.01$, the estimated confidence interval is no longer statistically significant. The reasoning behind this lies in the fact that $\triangle^{SD}$ imposes constraints on the slope of the violation of parallel trends over time. As $M$ increases, the slope grows at a faster rate, leading to a larger identified set.

I also did the sensitivity analysis when $\delta = \delta^{RM}$. The parameter $\bar{M}$ is varied from 0 to 2, where $\bar{M} = 1$ indicates that post-period violations of parallel trends do not exceed the maximal pre-treatment periods violation. The sensitivity analysis is conducted for different $\theta$ values: $\theta = \tau_1$ and $\theta = \bar{\tau}$. Appendix C.2 presents comprehensive results, including the lower bound, upper bound, the value of $\bar{M}$, and the method for RM estimation.

Figure 14 displays robust confidence sets for the treatment effect for $\triangle = \triangle^{RM}$ using various values of $\bar{M}$ and different choices of $\tau$. The blue confidence interval is estimated using OLS, while the red intervals are determined using the conditional-least favorable hybrid (C-LF hybrid). The length of the confidence interval for OLS remains consistent: [0.0096, 0.0271], while the length of the conditional-least favorable hybrid (C-LF hybrid) interval increases systematically with higher values of $M$.

The left panel of Figure 14 depicts confidence sets for the treatment effect when $\theta = \tau_1$. The confidence intervals for OLS and C-LF hybrid are initially similar, and the discrepancy widens with an increase in $\bar{M}$. The breakdown point for a significant effect is observed at $\bar{M} = 1$. This suggests that the significant result remains robust when allowing for violations of parallel trends up to as large as the maximum violation in the pre-treatment period. Hence, in order to maintain a significant confidence interval result, post-period violations should not exceed the



(a) $\theta = \tau_1$

(b) $\theta = \bar{\tau}$

Figure 14: Senstivity Analysis of RM for Deryugina (2017)

maximum violation observed in the pre-treatment period.

In the right panel of Figure 14, an alternative scenario is presented with $\theta = \bar{\tau}$. When $M = 0$, the difference in confidence intervals between C-LF and OLS is small. Compared to the situation when $\theta = \tau_1$, the difference in confidence intervals is also slight. However, as $M$ increases, the distinction between OLS and C-LF hybrid grows, contrasted with the scenario where $\theta = \tau_1$. In this case, the breakdown point is noted at $\bar{M} = 0.5$, a value smaller than the corresponding breakdown point observed for $\theta = \tau_1$. The intuition for the larger confidence sets when examining $\bar{\tau}$ compared to $\tau_1$ lies in the fact that $\triangle^{RM}$ constrains the violation of parallel trends across consecutive periods by $\bar{M}$ times the maximum observed in the pre-treatment period. Consequently, the identified set is larger for later periods, since the treatment and control groups have more time to diverge.

# 6 Conclusion

In this dissertation, I elucidate the methodology for constructing robust confidence sets in DiD and event estimation scenarios, particularly when the parallel trends assumption may face violations. Various restrictions are introduced to capture potential deviations in empirical studies. The utilization of moment inequalities for inference is presented, showcasing their robust asymptotic properties across a diverse range of restrictions. Furthermore, FLCIs are introduced, while the prior approach demonstrates that conditional and conditional hybrid confidence sets offer compelling asymptotic power guarantees, FLCIs provide finite-sample power guarantees specifically when $\triangle = \triangle^{SD}$.

In the extensive simulation study encompassing four types of restrictions in varied scenarios, unlike the original paper, the data generation process corresponds to the (non)parallel trends assumption, and the simulation conclusions are similar: conditional and conditional-least favorable hybrid confidence sets exhibit superior performance for generic forms of $\triangle$, whereas FLCIs excel when $\triangle^{SD}$ restriction is considered. Thus, the simulation's conclusion aligns with the theoretical assertions made earlier. To replicate the simulation process, please pay attention to the three points I mentioned in section 4.

In practice, robust confidence sets can be constructed through sensitivity analyses considering economic-driven restrictions. These sensitivity analyses provide transparency regarding the assumptions required to draw specific conclusions.

# References

Abadie, Alberto. 2005. Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, **72**(1), 1–19.

Andrews, Isaiah, Roth, Jonathan, & Pakes, Ariel. 2022. *Inference for Linear Conditional Moment Inequalities*.

Armstrong, Timothy B., & Kolesár, Michal. 2018. Optimal Inference in a Class of Regression Models. *Econometrica*, **86**(2), 655–683.

Bhuller, Manudeep, Havnes, Tarjei, Leuven, Edwin, & Mogstad, Magne. 2013. Broadband internet: An information superhighway to sex crime? *Review of Economic studies*, **80**(4), 1237–1266.

Bilinski, Alyssa, & Hatfield, Laura A. 2018. Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions. *arXiv preprint arXiv:1805.03273*.

Burde, Dana, & Linden, Leigh L. 2013. Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools. *American Economic Journal: Applied Economics*, **5**(3), 27–40.

Callaway, Brantly, & Sant'Anna, Pedro H.C. 2021. Difference-in-Differences with multiple time periods. *Journal of Econometrics*, **225**(2), 200–230. Themed Issue: Treatment Effect 1.

Callaway, Brantly, Li, Tong, & Oka, Tatsushi. 2018. Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, **206**(2), 395–413. Special issue on Advances in Econometric Theory: Essays in honor of Takeshi Amemiya.

Deryugina, Tatyana. 2017. The Fiscal Cost of Hurricanes: Disaster Aid versus Social Insurance. *American Economic Journal: Economic Policy*, **9**(3), 168–98.

Dobkin, Carlos, Finkelstein, Amy, Kluender, Raymond, & Notowidigdo, Matthew J. 2018. The economic consequences of hospital admissions. *American Economic Review*, **108**(2), 308–352.

Goodman-Bacon, Andrew. 2018. Public insurance and mortality: evidence from Medicaid implementation. *Journal of Political Economy*, **126**(1), 216–262.

Kahn-Lang, Ariella, & Lang, Kevin. 2020. The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics*, **38**(3), 613–620.

Lee, Jin Young, & Solon, Gary. 2011. The Fragility of Estimated Effects of Unilateral Divorce Laws on Divorce Rates. *The B.E. Journal of Economic Analysis and Policy*, **11**(1).

Malani, Anup, & Reif, Julian. 2015. Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform. *Journal of Public Economics*, **124**, 1–17.

Qin, Jing, & Zhang, Biao. 2008. Empirical-Likelihood-Based Difference-in-Differences Estimators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **70**(2), 329–349.

Rambachan, Ashesh, & Roth, Jonathan. 2023. A More Credible Approach to Parallel Trends. *The Review of Economic Studies*, **90**(5), 2555–2591.

Roth, Jonathan. 2022. Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights*, **4**(3), 305–22.

Sant'Anna, Pedro H.C., & Zhao, Jun. 2020. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, **219**(1), 101–122.

Schrijver, Alexander. 1998. *Theory of linear and integer programming*. John Wiley & Sons.

van der Vaart, Aad W., & Wellner, Jon A. 1996. *Weak Convergence*. New York, NY: Springer New York. Pages 16–28.

# A Canonical DiD ATT identification

$$y_{it} = d_i y_{it}(1) - (1 - d_i) y_{it}(0)$$

$$y_{it} = y_{it}(0) + d_i(y_{it}(1) - y_{it}(0))$$

$$\triangle_i = y_{it}(1) - y_{it}(0)$$

$$y_{it}(0) = v_t + u_i + \epsilon_{it}$$

$$y_{it} = d_i \triangle_i + v_t + u_i + \epsilon_{it}$$

$$y_{i,2} - y_{i,1} = (v_2 - v_1) + \triangle_i(d_{i,2} - d_{i,1}) + \epsilon_{i,2} - \epsilon_{i,1}$$

Here is the proving process, assume $d_{i,1} = 0$ and $d_{i,2} = d_i$:

$$E[y_{i,2}(0)|d_i = 1] = E[y_{i,1}(0)|d_i = 1] + E[y_{i,2}(0) - y_{i,1}(0)|d_i = 1]$$

According to the parallel trends

$$E[y_{i,2}(0)|d_i = 1] = E[y_{i,1}(0)|d_i = 1] + E[y_{i,2}(0) - y_{i,1}(0)|d_i = 0]$$

According to the no-anticipation, we can get:

$$ATT = E[y_{i,2}(1) - y_{i,2}(0)|d_i = 1] = E[y_{i,2} - y_{i.1}|d_i = 1] - E[y_{i,2} - y_{i,1}|d_i = 0]$$

# B Supplementary Material: Theoretical Derivations of Section 3

## B.1 $H_0$ transformation

$H_0 : \theta = \hat{\theta}, \delta \in \triangle$, and $\hat{\beta}_n \sim \mathcal{N}(\beta, \Sigma_n)$. $\beta = (\tau + \delta)$, so we can get:

$$E_{\hat{\beta}_n \sim \mathcal{N}(\beta, \Sigma_n)}[\hat{\beta}_n - \tau] = \delta$$

And $\delta \in \triangle = \{\delta : A\delta \leq d\}$, so we get:

$$E_{\hat{\beta}_n \sim \mathcal{N}(\beta, \Sigma_n)}[A\hat{\beta}_n - A\tau] \leq d$$

Define $Y_n = A\hat{\beta}_n - d$, and $\tau = L_{post}\tau_{post}$, finally the $H_0$ is:

$$H_0 : \exists \tau_{post} \in R^{\bar{T}} s.t. l^{'}\tau_{post} = \bar{\theta}, and \quad E_{\hat{\beta}_n \sim \mathcal{N}(\beta,\Sigma_n)}[Y_n - AL_{post}\tau_{post}] \leq 0$$

## B.2 Example 3.2 inference process

In three periods DiD and the bounding type of $\delta$ is $\triangle^{SD}$, we can get that the $A^{SD} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$, and $d^{SD} = \begin{pmatrix} M \\ M \end{pmatrix}$.And the expression of the $\hat{\eta}$ is:$\hat{\eta} = \max \tilde{Y}(\bar{\theta})^{SD}$. Then the following are the steps to get the exact expression:

$$\begin{aligned}
\tilde{Y}_n(\bar{\theta})^{SD} &= Y_n^{SD} - A^{SD}L_{post}\tau_1 \\
&= A^{SD}\hat{\beta}_n - d^{SD} - A^{SD}L_{post}\tau_1 \\
&= \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_{-1} \end{pmatrix} - \begin{pmatrix} M \\ M \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} 0 \\ \tau_1 \end{pmatrix} \\
&= \begin{pmatrix} \hat{\beta}_1 + \hat{\beta}_{-1} - M - \tau_1 \\ -\hat{\beta}_1 - \hat{\beta}_{-1} - M + \tau_1 \end{pmatrix} \\
&= \begin{pmatrix} \hat{\delta}_1 + \hat{\delta}_{-1} + \hat{\tau}_1 - M - \tau_1 \\ -\hat{\delta}_1 - \hat{\tau}_1 - \hat{\tau}_1 - M + \tau_1 \end{pmatrix} \\
&\overset{d}{\sim} \begin{pmatrix} \hat{\delta}_1 + \hat{\delta}_{-1} - M \\ -\hat{\delta}_1 - \hat{\tau}_1 - M \end{pmatrix}
\end{aligned}$$

## B.3 Assumption 3.6 proving

In assumption 3.5, we can get(We cannot directly infer the following formula from Assumption 3.5, but we can understand it this way):

$$\sqrt{n}\left((\hat{\beta} - \beta_P)^{'}, (vec(\hat{\Sigma}_n) - vec(\Sigma_P))^{'}\right)^{'} \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

$$\sigma = \begin{pmatrix} \Sigma_P & V_{P,\beta\Sigma} \\ V_{P,\Sigma\beta} & V_{P,\Sigma} \end{pmatrix}$$

By standardization, we then have:

$$\sqrt{n}\left(\sqrt{\Sigma_P^{\dagger}}(\hat{\beta} - \beta_P)^{'}, \sqrt{\Sigma_{P,\Sigma}^{\dagger}}(vec(\hat{\Sigma}_n) - vec(\Sigma_P))^{'}\right)^{'} \xrightarrow{d} \mathcal{N}(0, \Sigma^{*})$$

$$\Sigma^* = \begin{pmatrix} I & \sqrt{\Sigma_P^\dagger} V_{P,\beta\Sigma} \sqrt{V_{P,\Sigma}^\dagger} \\ \sqrt{\Sigma_P^\dagger} V_{P,\Sigma\beta} \sqrt{V_{P,\Sigma}^\dagger} & I \end{pmatrix}$$

Where $I$ is identity matrix. $\Sigma_P^\dagger \Sigma_P = I$ because $\Sigma_P$ has linearly independent columns. Analogously,$V_{P,\Sigma}^\dagger V_{P,\Sigma} = I$.

$\sqrt{\Sigma_P^\dagger} V_{P,\beta\Sigma} \sqrt{V_{P,\Sigma}^\dagger}$ should be a diagonal matrix with all elements equal to $a \in [-1, 1]$. Then we have:

$$\begin{aligned}
\sqrt{\Sigma_P^\dagger} V_{P,\beta\Sigma} \sqrt{V_{P,\Sigma}^\dagger} &= aI \\
V_{P,\beta\Sigma} \sqrt{V_{P,\Sigma}^\dagger} &= a\sqrt{\Sigma_P} \\
V_{P,\beta\Sigma} V_{P,\Sigma}^\dagger V_{P,\Sigma\beta} &= a^2 \Sigma_P \\
a^2 \Sigma_P - V_{P,\beta\Sigma} V_{P,\Sigma}^\dagger V_{P,\Sigma\beta} &= 0 \\
\Sigma_P - V_{P,\beta\Sigma} V_{P,\Sigma}^\dagger V_{P,\Sigma\beta} &= (1 - a^2)(\Sigma_P)
\end{aligned}$$

Assumption 3.5 requires that $V_{P,\beta\Sigma} V_{P,\Sigma}^\dagger V_{P,\Sigma\beta}$ has eigenvalues bounded below by $\tilde{\lambda} \geq 0$. That means $a \neq 1$ and $-1$. So this condition shows that the asymptotical distribution of $\hat{\beta}_n$ is not perfectly asymptotically colinear with $\hat{\Sigma}_n$, which is illustrated in the paper.

## B.4 Optimal bounds on excess length

According to the Theorem 3.2 of Armstrong & Kolesár (2018), we can get the following conclusions:

**Lemma B.1** *Suppose that $\triangle$ is convex. Let I denote the set of confidence sets that satisfy the coverage requirement. Then for any $\delta^* \in \triangle$, $\tau_{post}^* \in R^{\bar{T}}$, and $\Sigma_n$ positive definite,*

$$\inf_{C \in I_\alpha} E_{\hat{\beta}_n \sim \mathcal{N}(\delta^* + L_{post}\tau^*, \Sigma_n)}[\lambda(C)] = (1 - \alpha) E[\bar{w}(z_{1-\alpha} - Z) - \underline{w}(z_{1-\alpha} - Z) | Z < z_{1-\alpha}]$$

*where $Z \sim \mathcal{N}(0, 1)$, $z_{1-\alpha}$ is the $1 - \alpha$ quantile of Z, and*

$$\bar{w} := \sup\{l'\tau | \tau \in R^{\bar{T}}, \exists \delta \in \triangle, s.t. \|\delta + L_{post}\tau - \beta^*\|_{\Sigma_n}^2 \leq b^2$$

$$\underline{w} := \inf\{l'\tau | \tau \in R^{\bar{T}}, \exists \delta \in \triangle, s.t. \|\delta + L_{post}\tau - \beta^*\|_{\Sigma_n}^2 \leq b^2$$

*for $\beta^* := \delta^* + L_{post}\tau_{post}^*$, and $\|x\|_\Sigma = x'\Sigma^{-1}x$*

**Corollary B.1** *We can get the following Corollary under the Lemma 6.1:*

$$\inf_{C \in I_\alpha} E_{\hat{\beta}_n \sim \mathcal{N}(\beta^*, \Sigma_n)}[EL_{opt}(C;\beta^*)] = \inf_{C \in I_\alpha} E_{\hat{\beta}_n \sim \mathcal{N}(\beta^*, \sigma_n)}[\lambda(C)] - (1-\alpha) LID(\beta^*, \triangle) \tag{30}$$

*where $EL_{opt}(C;\beta^*) = \lambda(C \backslash S(\beta, \triangle))$ is the excess length of the confidence set C, i.e. the length of the part of the confidence set that falls outside of the identified set.*

When $\triangle = \bigcup_{k=1}^{K} \triangle_k$, the identified set is also the union of the identified sets for each of the $\triangle_k$. Thus, if the $C_\alpha$ satisfies the uniform coverage requirement for $\triangle$, then for each $\triangle_k$, the $C_\alpha$ also satisfies the uniform coverage requirement. Consequently, the expected excess length for $C$ is bounded below by the optimal excess length for confidence sets that satisfy the uniform coverage requirement for each $\triangle_k$.

# C Empirical results

## C.1 Empirical 1

Table 2: Result when $\triangle = \triangle^{RM}$ and $\theta = \tau_{2005}$

| lb | ub | method | delta | $\bar{M}$ |
|---|---|---|---|---|
| -0.1116 | -0.0081 | C-LF | RM | 0.00 |
| -0.1137 | 0.0016 | C-LF | SD | 0.50 |
| -0.1224 | 0.0156 | C-LF | SD | 1.00 |
| -0.1385 | 0.0361 | C-LF | SD | 1.50 |
| -0.1601 | 0.0588 | C-LF | SD | 2.00 |

Table 3: Result when $\triangle = \triangle^{RM}$ and $\theta = \bar{\tau}$

| lb | ub | method | delta | $\bar{M}$ |
|---|---|---|---|---|
| -0.1473 | -0.0370 | C-LF | SD | 0.00 |
| -0.1564 | -0.0131 | C-LF | SD | 0.50 |
| -0.1916 | 0.0301 | C-LF | SD | 1.00 |
| -0.2361 | 0.0768 | C-LF | SD | 1.50 |
| -0.2826 | 0.1246 | C-LF | SD | 2.00 |

## C.2 Empirical 2

Table 4: Result when $\triangle = \triangle^{SD}$ and $\theta = \tau_1$

| lb | ub | method | delta | $M$ |
|---|---|---|---|---|
| 0.0233 | 0.0475 | FLCI | SD | 0.00 |
| 0.0009 | 0.0550 | FLCI | SD | 0.01 |
| -0.0134 | 0.0608 | FLCI | SD | 0.02 |
| -0.0234 | 0.0708 | FLCI | SD | 0.03 |
| -0.0334 | 0.0809 | FLCI | SD | 0.04 |
| -0.0434 | 0.0909 | FLCI | SD | 0.05 |

Table 5: Result when $\triangle = \triangle^{SD}$ and $\theta = \bar{\tau}$

| lb | ub | method | delta | $M$ |
|---|---|---|---|---|
| 0.0207 | 0.0546 | FLCI | SD | 0.00 |
| -0.1030 | 0.1619 | FLCI | SD | 0.01 |
| -0.2237 | 0.2317 | FLCI | SD | 0.02 |
| -0.3171 | 0.3251 | FLCI | SD | 0.03 |
| -0.4104 | 0.4184 | FLCI | SD | 0.04 |
| -0.5037 | 0.5117 | FLCI | SD | 0.05 |

Table 6: Result when $\triangle = \triangle^{RM}$ and $\theta = \tau_1$

| lb | ub | method | delta | $\bar{M}$ |
|---|---|---|---|---|
| 0.0162 | 0.0393 | C-LF | RM | 0.00 |
| 0.0048 | 0.0505 | C-LF | RM | 0.50 |
| -0.0104 | 0.0655 | C-LF | RM | 1.00 |
| -0.0263 | 0.0814 | C-LF | RM | 1.50 |
| -0.0428 | 0.0978 | C-LF | RM | 2.00 |

Table 7: Result when $\triangle = \triangle^{RM}$ and $\theta = \bar{\tau}$

| lb | ub | method | delta | $\bar{M}$ |
|---|---|---|---|---|
| 0.0105 | 0.0265 | C-LF | RM | 0.00 |
| -0.0425 | 0.0796 | C-LF | RM | 0.50 |
| -0.1006 | 0.1376 | C-LF | RM | 1.00 |
| -0.1597 | 0.1967 | C-LF | RM | 1.50 |
| -0.2197 | 0.2568 | C-LF | RM | 2.00 |

I hereby confirm that the work presented has been performed and interpreted solely by myself except for where I explicitly identified the contrary. I assure that this work has not been presented in any other form for the fulfillment of any other degree or qualification. Ideas taken from other works in letter and in spirit are identified in every single case.