A practitioners guide by Scott Cunningham Detail Explanation

27 March 2025 14:50

Beispiel das durch das Paper hindurch gezogen wird

How did the expansion of public health insurance (Medicaid) under the Affordable Care Act (ACA) affect mortality?
ACA (Affordable Cares Act) in den USA sollte 2014 eingeführt werden per Gesetz (Mehr Leute mit geringem Einkommen haben zugang zu Midcaid) und dann aber postponed druch ein gesetz 2012. folglich haben manche Staaten erst danach diese Medicaid eingeführt und manche bis heute noch gar nicht! Man sieht auch dass die einzelnen Staaten unterschiedlich sind wie sie aufgebaut sind also wie viele counties in einem staat wie viele Menschen, entsprechend hat die schrittweise erweiterung auch immer unterschiedlich viele leute erreicht!

Table 1: Medicaid Expansion Under the Affordable Care Act

Expansion Year	States	Share of States	Share of Counties	Share of Adults (2013)
Pre-2014	DE, MA, NY, VT	0.08	0.03	0.09
2014	AR, AZ, CA, CO, CT, HI, IA, IL, KY, MD, MI, MN, ND, NH, NJ, NM, NV, OH, OR, RI, WA, WV	0.44	0.36	0.45
2015	AK, IN, PA	0.06	0.06	0.06
2016	LA, MT	0.04	0.04	0.02
2019	ME, VA	0.04	0.05	0.03
2020	ID, NE, UT	0.06	0.04	0.02
2021	MO, OK	0.04	0.06	0.03
2023	NC, SD	0.04	0.05	0.03
Non-Expansion	AL, FL, GA, KS, MS, SC, TN, TX, WI, WY	0.20	0.31	0.26

Notes: The table shows which states adopted the ACA's Medicaid expansion in each year as well as the share of all states, counties, and adults in each expansion year.

D_i Definiert den Treatment Dummy =1 für treatment und = 0 für untreated
$D_{i,t}$ entspricht dem treatment status dummy allso wenn post und treated dann=1
$Y_{i,t}(0,0) \coloneqq Y_{i,t}(0)$ denote unit i's potential outcome at time t if it remained untreated in both periods. mortality rate wenn land nicht midcaid eingeführt hat
$Y_{i,t}(0,1) \coloneqq Y_{i,t}(1)$ is potential outcome at tim t if untreated in the first period but exposed to treatmend by the second period, mortaility rate Medicaid wenn eingeführt hat in 2014

Switching Equation (da unobersvable counterfactual:
$Y_{i,t} = (1 - D_i)Y_{i,t}(0) + D_i Y_{i,t}(1)$

was gemessen wird der ATT(t) also average treatment effect of treatet zu einem gewissen Zeitpunkt:
$ATT(t) = E_\omega[Y_{i,t}(1) - Y_{i,t}(0)|D_i = 1] = E_\omega[Y_{i,t}|D_i = 1] - E_\omega[Y_{i,t}(0)|D_i = 1]$

Wichtige allgemeine Annahmen die allgemein gelten für DiD:

No Anticipation - Annahme: (see, e.g., Abbring and van den Berg, 2003, Malani and Reif, 2015, and Roth et al., 2023). For all treated units i and all pre-treatment periods t,

Yi,t(1) = Yi,t(0).
Die Behandlung darf keine Effekte haben, *bevor* sie implementiert wird. D.h., das Outcome in der Vorperiode (t=1) entspricht dem unbehandelten potenziellen Outcome Y(0) für beide Gruppen.
ensures that we observe untreated potential outcomes before
Medicaid expansion takes effect:
helps to define when treatment begins: For instance, if the announcement of Medicaid expansion affects mortality before its actual
expansion, "treatment" begins when the policy is announced rather than implemented.

Parallel Trends - Annahme (ist schwächer als mean independence nund time invariance assumption bzw verletzt beide annahmen):

absence of treatment, the average outcome evolution is the same among treated and comparison groups. Durchschnittliche Veränderung des Outcomes (abh.variable) wäre in Treatmentgruppe gleichgewesen wie in Kontrollgruppe wenn die Treatmentgruppe nicht behandelt worden wäre. Wichtig es geht um Parallele Trends nicht um Parallele Level.
The (weighted) average change of Yi,t=2(0) from Yi,t=1(0)
is the same between treated and comparison groups, i.e Formal:,

$$\mathbb{E}_\omega[Y_{i,t=2}(0)|D_i=1]-\mathbb{E}_\omega[Y_{i,t=1}(0)|D_i=1]=\mathbb{E}_\omega[Y_{i,t=2}(0)|D_i=0]-\mathbb{E}_\omega[Y_{i,t=1}(0)|D_i=0]. \quad (3.3)$$

umstellen und man sieht wie counterfactual daraus konsturiert wird

$$\mathbb{E}_\omega[Y_{i,t=2}(0)|D_i=1]=\mathbb{E}_\omega[Y_{i,t=1}|D_i=1]+\big(\mathbb{E}_\omega[Y_{i,t=2}|D_i=0]-\mathbb{E}_\omega[Y_{i,t=1}|D_i=0]\big).$$

um medicaid effect in counterfactual welt der treatet gruppe ohne treatment zu berechnen, starte also mit 2013 mortality rate für die treatet und füge dann die differenz(änderung) der kontrollgruppe hinzu.
einsetzen in ATT formel liefert

$$\begin{aligned} ATT(2014) &= \overbrace{\mathbb{E}_\omega[Y_{i,2014}|D_i=1]}^{=\mathbb{E}_\omega[Y_{i,2014}(1)|D_i=1]} - \overbrace{\big(\mathbb{E}_\omega[Y_{i,2013}|D_i=1]+\big(\mathbb{E}_\omega[Y_{i,2014}|D_i=0]-\mathbb{E}_\omega[Y_{i,2013}|D_i=0]\big)\big)}^{=\mathbb{E}_\omega[Y_{i,2014}(0)|D_i=1]} \\ &= \underbrace{\big(\mathbb{E}_\omega[Y_{i,2014}|D_i=1]-\mathbb{E}_\omega[Y_{i,2013}|D_i=1]\big)}_{\text{(weighted) average change for } D_i=1} - \underbrace{\big(\mathbb{E}_\omega[Y_{i,2014}|D_i=0]-\mathbb{E}_\omega[Y_{i,2013}|D_i=0]\big)}_{\text{(weighted) average change for } D_i=0}. \end{aligned}$$

Anders als bei RCTs ist die Parallel-Trends-Annahme bei DiD keine Folge des Studiendesigns. Da die Behandlung oft nicht zufällig erfolgt, sondern von den Akteuren selbst gewählt wird (endogene Auswahl), gibt es keinen Automatismus, dass die Trends parallel verlaufen wären. Zum Beispiel könnten Firmen, die eine neue Technologie einführen (Treatment), sich auch ohne diese Technologie systematisch anders entwickelt haben als Firmen, die sie nicht einführten. Weil die Annahme so wichtig ist für die Gültigkeit der Formel im Bild, aber nicht garantiert ist, müssen Forscher, die DiD anwenden, Argumente und Belege dafür liefern, dass die Annahme in ihrem spezifischen Kontext plausibel ist.

Plausibilität kann nicht nur empirisch sondern auch theoretisch untermauert werden:

THEORETISCHE ARGUMENTATION WANN PARALLEL TRENDS HÄLT!

Ein zentraler Trade-off: Es gibt einen wichtigen Zusammenhang (Trade-off) zwischen:

- Dem **Wissen**, das die entscheidenden Akteure haben (insbesondere über zukünftige Entwicklungen *ohne* die Behandlung = Y(0)).
- Wie sie auf dieses Wissen **reagieren** (d.h., ob sie ihre Entscheidung davon abhängig machen).
- Den **zeitlichen Eigenschaften** des potenziellen Ergebnisses *ohne* Behandlung (Y(0)), also wie es sich über die Zeit normalerweise verändert (z. B. stabil, trendend, schwankend).

Extremfall 1: Perfektes Wissen und Reaktion:

- **Szenario:** Stell dir vor, jemand wüsste *genau*, wie sich sein Ergebnis Y(0) in der Zukunft entwickeln würde (sowohl vor als auch nach dem potenziellen Behandlungszeitpunkt) und könnte basierend auf diesem Wissen entscheiden, ob er die Behandlung annimmt oder nicht.
- **Konsequenz für Parallel Trends:** In diesem extremen Fall kann die Parallel-Trends-Annahme nur unter sehr unrealistischen Bedingungen halten: Entweder müsste sich Y(0) für alle Einheiten (abgesehen von einem gemeinsamen, gleichen Anstieg/Abfall) über die Zeit **kaum verändern** (konstant sein), ODER die Akteure dürften nicht auf ihr perfektes Wissen reagieren (was dem Szenario widerspricht). Warum? Wenn Y(0) individuelle Trends hat *und* die Leute die Behandlung wählen, *weil* sie diese Trends antizipieren (z. B. nur die mit erwarteten schlechten Trends nehmen die Behandlung), dann hätten Behandlungs- und Kontrollgruppe per Definition unterschiedliche *potenzielle* Trends gehabt, was Parallel Trends verletzt.
- **Bezug zum Medicaid-Beispiel:** Es ist höchst unwahrscheinlich, dass die US-Bundesstaaten 2013 genau wussten, wie hoch ihre Sterblichkeitsrate 2014 *ohne* die Medicaid-Erweiterung

sein würde (Y(0)). Ebenso ist es unwahrscheinlich, dass sich die Sterblichkeitsraten *ohne* die Erweiterung in allen Staaten exakt parallel entwickelt hätten. Dieses Extrem-Szenario trifft also nicht zu, illustriert aber das Problem.

E**xtremfall 2: Zufällige Behandlungswahl:**

- **Szenario:** Die Entscheidung für die Behandlung erfolgt rein zufällig (wie in einem randomisierten Experiment, RCT).
- **Konsequenz für Parallel Trends:** In diesem Fall würde die Parallel-Trends-Annahme (im Durchschnitt) automatisch gelten.
- **Aber:** Wenn die Zuweisung wirklich zufällig ist, ist die (manchmal komplexere) DiD-Methode gar nicht unbedingt nötig; es gäbe einfachere oder effizientere Schätzer für den Behandlungseffekt.
- **Bezug zum Medicaid-Beispiel:** Die Entscheidung der Bundesstaaten zur Medicaid-Erweiterung war offensichtlich *nicht* zufällig, sondern eine stark politisch und ökonomisch motivierte Entscheidung. Man kann also nicht einfach annehmen, dass Parallel Trends aufgrund von Zufälligkeit gilt.

Realistischere Selektionsmechanismen und Parallel Trends

1. **Bedingungen für Parallel Trends:** In der Realität, wo weder perfektes Wissen noch reiner Zufall vorliegen, kann die Parallel-Trends-Annahme nur gelten, wenn es bestimmte **Einschränkungen** dafür gibt, *wie* die potenziellen unbehandelten Outcomes (Y(0)) die Entscheidung für die Behandlung beeinflussen. Es kommt also darauf an, *welche Informationen* über Y(0) in die Entscheidung einfließen.
2. **Beispiel: Selektion nach "Fixed Effects" (langfristigen Entscheidungen):**
    - **Szenario:** Stellen Sie sich vor, die Entscheidungsträger (z. B. die Gesetzgeber der Bundesstaaten im Medicaid-Beispiel) basieren ihre Entscheidung *nur* auf der **langfristigen, stabilen Komponente** ihres unbehandelten Outcomes (Y(0)), den sogenannten "Fixed Effects" (z. B. das generelle, durchschnittliche Mortalitätsniveau eines Staates), *nicht* aber auf kurzfristigen Schwankungen ("Schocks", z. B. ein ungewöhnlich hoher oder niedriger Wert im Vorjahr).
    - **Konsequenz:** In diesem Fall hätten Staaten, die expandieren, und solche, die nicht expandieren, zwar unterschiedliche *Niveaus* (Fixed Effects) der Mortalität. Aber die DiD-Methode (wie in der Formel im Bild, Gleichung 3.3) bereinigt diese Niveauunterschiede, indem sie die *Veränderungen über die Zeit* betrachtet (Differenzbildung).
    - **Wann Parallel Trends hält:** Wenn die Selektion nur auf diesen Fixed Effects basiert, würde die Parallel-Trends-Annahme (bezogen auf die *Veränderungen*) gelten, *vorausgesetzt*, die kurzfristigen Schocks haben im Durchschnitt ein stabiles Verhalten (oder sind für beide Gruppen im Mittel gleich).
3. **Komplikation: Selektion auch nach "Shocks"(kurzfristigen Entscheidungen):**
    - **Szenario:** Was aber, wenn die Gesetzgeber *auch* auf die kurzfristigen Schwankungen ("Shocks") reagieren? Also wenn z. B. ein Staat gerade eine besonders hohe Mortalitätsrate im Jahr 2013 hatte (Y(0) in 2013 enthielt einen großen positiven Schock) und *deshalb* eher dazu neigte, Medicaid zu erweitern?
    - **Konsequenz:** Wenn die Entscheidung *sowohl* vom langfristigen Niveau (Fixed Effect) *als auch* von kurzfristigen Schocks abhängt, wird es schwieriger, die Parallel-Trends-Annahme zu rechtfertigen. Sie würde nur unter **stärkeren Annahmen** über das Zeitreihenverhalten von Y(0) gelten (z. B. wie schnell Schocks wieder abklingen). Die einfache Differenzbildung in DiD reicht dann möglicherweise nicht mehr aus, um die Selektionseffekte zu bereinigen.
    - **Verweis auf Literatur:** Ghanem et al. (2023b) und Marx et al. (2024) diskutieren diesen Zusammenhang zwischen Selektionsmechanismen und den notwendigen Annahmen über die Zeitreiheneigenschaften genauer

Wahl der Funktionalen Form entscheidend: Level vs Logarithmus können Gültigkeit von parallel Trends beeinflussen

- **Problemstellung:** Weil DiD "nur" parallele Trends im *Durchschnitt* annimmt und nicht auf voller statistischer Unabhängigkeit wie bei RCTs beruht, gibt es keine Garantie, dass die Annahme, wenn sie für eine bestimmte Messung des Outcomes Y gilt (z. B. Mortalitäts*rate*), auch für eine Transformation davon gilt (z. B. *logarithmierte* Mortalitätsrate). Die Wahl der Skala/Funktionsform kann also wichtig sein.
- **Bedingung für Unabhängigkeit von Funktionsform:** Roth und Sant'Anna (2023b) haben gezeigt, dass die Parallel-Trends-Annahme *genau dann* unabhängig von der Funktionsform ist (also z. B. für Raten und Log-Raten gleichzeitig gilt), wenn sie nicht nur für den Durchschnitt, sondern für die **gesamte Verteilung** von Y(0) gilt und das über beide Gruppen hinweg.
- **Strikte Bedingungen:** Damit Parallel Trends für die gesamte Verteilung gilt, müsste aber (vereinfacht gesagt) entweder die Behandlungswahl (Medicaid-Erweiterung) **zufällig** sein oder die **Verteilung** der Mortalität (Form, Streuung etc.) zwischen 2013 und 2014 **konstant** bleiben. Beides wurde bereits als eher unrealistisch eingeschätzt.
- **Konsequenz für die Praxis:** Da diese Bedingungen sehr streng sind, hängt die Gültigkeit der DiD-Analyse wahrscheinlich davon ab, *wie* man das Outcome misst (hier: Raten pro 100.000). Hätten die Autoren z. B. Logarithmen verwendet, könnte die Parallel-Trends-Annahme möglicherweise nicht mehr gelten (oder umgekehrt).
- **Umgang damit:**
  - Ideal wäre eine ökonomische Theorie, die die "richtige" Funktionsform vorgibt (oft nicht verfügbar).
  - Alternativ empfehlen die Autoren **Falsifikationstests** (von Roth und Sant'Anna, 2023b), um empirisch zu prüfen, ob die Parallel-Trends-Annahme in den Daten robust gegenüber Änderungen der Funktionsform erscheint. siehe hier: https://onlinelibrary.wiley.com/doi/full/10.3982/ECTA19402
  - **Ergebnis im Paper:** Für ihre spezifische Anwendung (Medicaid-Beispiel) finden die Autoren mit diesen Tests *keine Evidenz* dafür, dass ihre Ergebnisse von der Wahl der Funktionsform (Raten) abhängen (hohe p-Werte > 0.80). Das stärkt das Vertrauen in ihre Analyse mit Mortalitätsraten.

Estimation and inference of 2x2 DiD

Estimation

population weighting oder nicht macht einen rießen Unterschied wie der effekt am Ende ist! Hat dann am Ende aber natürlich auch ganz andere Interpretation:
0.1 deaths per 100,000, implying that the average treatment effect of Medicaid expansion had on mortality in 2014 among counties that are part of an expansion state was an increase of 0.1 deaths per 100,000.
In contrast, the DiD result using population weights suggests that Medicaid expansion caused a reduction of 2.6 deaths per 100,000 for the average adult in expansion states.

Also:

- ATT für den durchschnittlichen *County*.
- ATT für die durchschnittliche *Person*.
  Der Unterschied kommt zustande, weil die (Populations-)Gewichtung dazu führt, dass bevölkerungsreiche Landkreise das Ergebnis stärker beeinflussen. Wenn die Trends der Mortalitätsraten (oder die Effekte der Expansion selbst) in bevölkerungsreichen Landkreisen anders sind als in bevölkerungsarmen, dann führt die unterschiedliche Gewichtung zu unterschiedlichen Durchschnittseffekten. Es ist also kein Widerspruch, sondern die Schätzung zweier verschiedener, aber jeweils valide definierter Durchschnittswerte. Für politische Entscheidungen argumentieren die Autoren oft, dass der populationsgewichtete Effekt relevanter sein könnte [source: 119].

Table 2: Simple 2 × 2 DiD

	Unweighted Averages			Weighted Averages		
	Expansion	No Expansion	*Gap/DiD*	Expansion	No Expansion	*Gap/DiD*
2013	419.2	474.0	*-54.8*	322.7	376.4	*-53.7*
2014	428.5	483.1	*-54.7*	326.5	382.7	*-56.2*
Trend/DiD	*9.3*	*9.1*	0.1	*3.7*	*6.3*	-2.6

Notes: This table reports average county-level mortality rates (deaths among adults aged 20-64 per 100,000 adults) in 2013 (row 1) and 2014 (row 2) in states that expanded adult Medicaid eligibility in 2014 (columns 1 and 4) and states that have not expanded by 2019 (columns 2 and 5). The first three columns present unweighted averages and the second three columns present population-weighted averages. Columns 1, 2, 4, and 5 in the third row show time trends in mortality between 2013 and 2014 for each group of states. The first two rows of columns 3 and 6 show the cross-sectional gap in mortality between expansion and non-expansion states in 2013 and 2014. The entries in bold red text in row 3 show the simple 2 × 2 difference-in-differences estimates without weights (column 3) and with them (column 6)

Table 3: im einfachen kanonischen 2x2 DiD-Fall (zwei Gruppen, zwei Zeitpunkte, balancierte Paneldaten) verschiedene gängige Regressionsmethoden(einfache OLS-Interaktion, TWFE, Regression der Differenzen)) *numerisch exakt denselben* Schätzwert für den DiD-Effekt (den ATT) liefern.

Inference(Standardfehler, Konfidenzintervalle)

- **Das Grundproblem: Woher kommt die Unsicherheit?**
  - DiD-Schätzer sind aus Daten berechnet und daher unsicher. Standardfehler sollen diese Unsicherheit quantifizieren [source: 199].
  - Die entscheidende Frage ist: Was betrachten wir als "zufällig" in unserem Datensatz oder im Prozess, der die Daten generiert hat? [source: 200] Die Antwort darauf bestimmt, wie wir die Unsicherheit modellieren und berechnen.
  - Man muss sich ein Gedankenexperiment vorstellen: Wenn wir die Studie/den Prozess wiederholen könnten, was würde sich ändern (wäre zufällig), und was würde gleich bleiben (wäre fix)? [source: 201, 203]
- **Drei grundlegende Ansätze zur Inferenz (Unsicherheitsberechnung):**
  - **a) Design-basierter Ansatz (Design-based):**
    - **Fokus:** Die spezifische Stichprobe (finite population), die man hat [source: 205].
    - **Zufallsquelle:** *Nur* die Zuweisung zum Treatment (oder zur Kontrollgruppe) wird als zufällig betrachtet [source: 205, 206]. Die potenziellen Outcomes der Einheiten und ihre Kovariaten werden als *fix* angesehen [source: 205]. Man fragt: Wie würde sich unser Schätzer ändern, wenn die (zufällige) Treatment-Zuweisung anders ausgefallen wäre?
    - **Anwendung:** Klassisch bei randomisierten Experimenten, aber es gibt Erweiterungen für quasi-experimentelle Designs (siehe Fußnote 10 [source: 213]). Oft sind hierfür aber starke Annahmen nötig, wenn die Zuweisung nicht wirklich rein zufällig war.
  - **b) Stichproben-basierter Ansatz (Sampling-based):**
    - **Fokus:** Schlussfolgerung von der Stichprobe auf eine größere Grundgesamtheit (superpopulation) [source: 207, 208].
    - **Zufallsquelle:** Der *Stichprobenziehungsprozess* selbst. Man nimmt an, die beobachteten Einheiten sind eine zufällige Ziehung aus einer größeren Population [source: 207]. Alle beobachteten Variablen (Outcomes, Kovariaten, Treatmentstatus) werden als zufällig betrachtet, weil sie vom Zufall der Stichprobenziehung abhängen [source: 208, 209].
    - **Anwendung:** Das ist der häufigste Ansatz in der angewandten Ökonometrie. Man berechnet oft Standardfehler, die für Abhängigkeiten innerhalb von Clustern korrigieren (geclusterte Standardfehler), z.B. auf Ebene der Bundesstaaten oder Landkreise, je nachdem, auf welcher Ebene die Stichprobenziehung oder die Treatment-Variation stattfindet [source: 208].
    - **Nachteil:** Die Annahme einer einfachen Zufallsstichprobe aus einer klar definierten Grundgesamtheit ist manchmal unrealistisch (z.B. wenn man alle US-Bundesstaaten betrachtet) [source: 210].
  - **c) Modell-basierter Ansatz (Model-based):**

    - **Fokus:** Ein spezifisches statistisches Modell, das den Datengenerierungsprozess beschreiben soll, insbesondere die Struktur der Fehlerterme (shocks) [source: 211].
    - **Zufallsquelle:** Die *Fehlerterme* (ε_it in einer Regression) werden als zufällige Ziehungen aus einer bestimmten Verteilung angenommen [source: 212]. Die Unsicherheit entsteht dadurch, dass bei einer Wiederholung des Prozesses andere Fehlerterme gezogen würden [source: 212].
    - **Anwendung:** Häufig in der Ökonometrie, oft beginnend mit einer linearen Regressionsgleichung [source: 213].
    - **Nachteile:** (1) Annahmen über die Fehlerstruktur können implizit restriktive Annahmen über die Heterogenität der Behandlungseffekte bedeuten [source: 215]. (2) Es ist oft schwierig, diesen Ansatz konsistent auf Methoden anzuwenden, die nicht direkt auf Regressionen basieren, wie Inverse probability weighting oder Double robust-Schätzer [source: 216].
- **Keine universelle Antwort:**
  - Die Autoren betonen, dass die Wahl des Ansatzes komplex und kontextabhängig ist [source: 217, 218]. Es gibt keine Methode, die immer die beste ist. Die Entscheidung hängt von der Forschungsfrage, der Datenstruktur, dem Stichprobenverfahren und den Annahmen ab, die man bereit ist zu treffen [source: 219].
- **Die Wahl der Autoren für ihr Beispiel:**
  - Für ihre Analyse der Medicaid-Expansion wählen die Autoren den **stichproben-basierten Ansatz** [source: 222].
  - Sie berechnen Standardfehler, die auf **County-Ebene geclustert** sind [source: 222].
  - Clustern:
    Warum überhaupt? Annahme ist ja dass Fehlerterme voneinander unabhängig (zufall der outcome von einem Landkreis beeinflusst hat nix mit Zufall zu tun der Outcome von anderem Landkreis beeinflusst) , aber das ist oft verletzt insbesondere bei Paneldaten.
    Korrelation über die Zeit innerhalb eines Landkreises: Ein landkreis hat besonders gute Verwaltung oder geografische Lage die dafür sorgt dass wen outcome in einem Jahr höher dann auch in dem darauf höher
    Korrelation zwischen den Einheiten, also Landkreisen: Landkreise liegen regional nah beieinander, wenn es zum Beispiel besondere Wetterphänomene gibt oder Grippewellen oder Ähnlich Wirtschaftspolitik weil in selben Bezirk, und das einen Landkreis beeinflusst, dann beeinlfusst das auch einen anderen Landkreis!

    Clustering sagt dann: "Achtung, die Beobachtungen innerhalb dieser Gruppe (z.B. dieses Bundesstaates) sind nicht unabhängig voneinander! Behandle sie nicht so, als ob jede einzelne eine völlig neue Information liefert." Daraufhin korrigiert das Verfahren den Standardfehler nach oben, um der reduzierten effektiven Informationsmenge Rechnung zu tragen.

  - **Begründung:**
    - Dieser Ansatz behandelt alle Variablen (Outcomes, Kovariaten, Gruppenzugehörigkeit) als zufällig [source: 223].
    - Er vermeidet starke Annahmen über Zeitreihenabhängigkeiten (da sie viele Einheiten N, aber wenige Zeitperioden T haben) [source: 224].
    - Er vermeidet Annahmen über die Struktur der Fehlerterme, was gut zu ihrem "Forward-Engineering"-Ansatz passt, der bei den potenziellen Outcomes beginnt [source: 224]. Clustern auf County-Ebene ermöglicht es ihnen, flexibel mit Korrelationen *innerhalb* eines Landkreises über die Zeit umzugehen, ohne spezifische Annahmen über die Zeitreihenstruktur der Fehler oder eine bestimmte funktionale Form des Modells treffen zu müssen [source: 224]. Dies passt gut zu ihrem "Forward-Engineering"-Ansatz, der bei den potenziellen Outcomes ansetzt und nicht bei einem Regressionsmodell.
    - Entscheidend: Clustering auf County-Ebene ist auch vereinbar mit der Idee, dass unbeobachtete Schocks auf Bundesstaatsebene die Parallel-Trends-Annahme verletzen könnten [source: 225]. Sie erkennen also an dass die Behandlung auf State ebene passiert und auch schocks auf staatenebene die parallel Trends annahme beeinflussen könnte. Allerdings sagen sie dass dieses staateneffekte fix (in dm sinne also beindgt, bzw

nicht zufällig, konstant oder gegeben) Für unsere Analyse nehmen wir die spezifischen Ereignisse und Bedingungen, die auf Staatenebene *tatsächlich passiert sind* (z.B. die Wirtschaftslage in Texas 2015, die Gesundheitspolitik in Kalifornien 2016), als **gegeben und unveränderlich** an. Sie sind der feste Rahmen, die Kulisse, vor der wir unsere Analyse durchführen.

  - **Einschränkung:** Sie merken an, dass diese Wahl nicht unumstritten ist und andere Ansätze (z.B. Clustering auf Staatenebene unter einem Design-basierten Blickwinkel) ebenfalls argumentierbar wären [source: 226, 227].

- **Schlussfolgerung des Abschnitts:**
  - Der letzte Absatz [source: 229-233] schlägt den Bogen zurück: Die anfängliche Attraktivität von einfachen Regressionen im 2x2-Fall (wie in Tabelle 3 gezeigt) lag auch daran, dass sie Standard-Inferenzwerkzeuge boten [source: 198]. Diese enge Verbindung zwischen Regression und verlässlicher Inferenz bricht jedoch in den komplexeren, in der Praxis häufigen Fällen auseinander [source: 231]. Dies unterstreicht die Notwendigkeit der sorgfältigeren Methoden und Überlegungen zur Inferenz, die im Paper diskutiert werden.

Covariates bzw Kontrollvariablen miteinbeziehen in den 2x2 DiD

Bisher wurde das DiD-Design ohne Berücksichtigung zusätzlicher Variablen betrachtet Allgemein können Covariates unterschiedliche Effekte/Sinn haben sie miteinzubeziehen:

- **Praxis:** In der Forschungspraxis werden Kovariaten jedoch häufig auf drei Arten genutzt:
  1. **Balance prüfen:** Man überprüft, ob sich Treatment- und Kontrollgruppe in Variablen unterscheiden, von denen man annimmt, dass sie das Outcome (ohne Treatment, Y(0)) beeinflussen könnten. [source: 235]
  2. **Kontrollieren:** Man bezieht diese Variablen aktiv in die Schätzung mit ein, um die Glaubwürdigkeit der Parallel-Trends-Annahme zu erhöhen. [source: 235]
  3. **Heterogenität schätzen:** Man untersucht, ob der Behandlungseffekt für Gruppen mit unterschiedlichen Kovariatenwerten verschieden ist. [source: 235]

  Beispiel (Medicaid & Mortalität):
  - Ohne die Medicaid-Expansion hätten sich die Mortalitätsraten in ärmeren Landkreisen wahrscheinlich anders entwickelt als in reicheren [source: 235]. Wenn sich nun die Expansions- und Nicht-Expansions-Landkreise systematisch in ihrer Armutsquote *unterscheiden*, dann ist die einfache Parallel-Trends-Annahme (dass die Trends ohne Expansion parallel gewesen wären) möglicherweise verletzt [source: 236]. Man könnte also versuchen, für die Armutsquote zu "kontrollieren" [source: 237].
  - Außerdem könnte die Expansion in ärmeren Landkreisen (wo mehr Menschen potenziell profitieren) einen größeren Effekt auf die Gesamtmortalität haben als in reicheren. Diese Unterschiedlichkeit (Heterogenität) der Effekte könnte interessant sein [source: 238, 239].

  Es soll gezeigt werden, wie man Kovariaten nutzt, um (1) die Parallel-Trends-Annahme zu bewerten, (2) den ATT unter potenziell schwächeren Annahmen (Conditional Parallel Trends) zu identifizieren und (3) Heterogenität zu untersuchen [source: 241]. Der rote Faden ist wieder die Betrachtung durch die "2x2-Baustein"-Brille [source: 241]. Es wird auch darauf eingegangen, warum einfache Regressionsansätze zum Kontrollieren von Kovariaten problematisch sein können

Unconditional parallel trends erfüllt oder nicht?

- **Problem:** Die zentrale Parallel-Trends-Annahme (Assumption PT [source: 129]) ist formal **nicht testbar**, da sie sich auf kontrafaktische (unbeobachtete) Entwicklungen bezieht [source: 243].
- **Indirekte Überprüfung:** Man kann sie aber indirekt auf Plausibilität prüfen, indem man sich

beobachtbare Variablen (Kovariaten) ansieht, von denen man annimmt, dass sie mit den unbeobachteten Outcome-Trends (Y(0)-Trends) zusammenhängen [source: 244]. Beispiele sind Demografie oder ökonomische Bedingungen auf regionaler Ebene [source: 245].

- **Logik der Balance-Checks:**
  - Wenn solche Kovariaten die Outcome-Trends beeinflussen, *und*
  - wenn sich die Treatment- und Kontrollgruppe in diesen Kovariaten *vor* der Behandlung systematisch unterscheiden,
  - *dann* ist es plausibel, dass die beiden Gruppen auch ohne das Treatment unterschiedliche Outcome-Trends gehabt hätten. Die einfache (unbedingte) Parallel-Trends-Annahme wäre dann verletzt [source: 246, 247].
  - Daher ist das Überprüfen der "Balance" (Ähnlichkeit) der Gruppen in relevanten Kovariaten ein wichtiger Schritt [source: 247].
- **Was wird verglichen?**
  - **Baseline Levels (Ausgangswerte):** Man vergleicht die Durchschnittswerte der Kovariaten in der Periode *vor* dem Treatment (t=1) zwischen der Treatment- (D=1) und Kontrollgruppe (D=0) [source: 248]. Genau das macht **Tabelle 4, Panel A** für das Jahr 2013 [source: 251, 260].
  - **Covariate Trends (Entwicklungen):** Man kann auch die *Veränderung* der Kovariaten zwischen der Vorher- (t=1) und Nachher-Periode (t=2) zwischen den Gruppen vergleichen [source: 248]. Das macht **Tabelle 4, Panel B** für die Veränderung von 2013 zu 2014 [source: 262, 270].
- **Messung der Unähnlichkeit:** Um Unterschiede über Variablen mit verschiedenen Einheiten hinweg vergleichen zu können, wird oft die "normalisierte Differenz" verwendet (Formel im Text) [source: 255]. Als Faustregel gilt oft: Absolute Werte über 0.25 deuten auf eine potenziell problematische Unähnlichkeit (Imbalance) hin [source: 256].

Beispiel:

- **Panel A (Levels 2013):** Es gibt deutliche Unterschiede zwischen den Gruppen bei den Ausgangswerten 2013 (z.B. bei % Weiß, Armutsquote, Medianeinkommen) [source: 264]. Das ist problematisch, wenn diese Ausgangswerte mit den *zukünftigen Trends* der Mortalität zusammenhängen [source: 266, 267]. Das Argument, DiD würde Level-Unterschiede herausrechnen, greift nicht, wenn die Levels die Trends beeinflussen [source: 266].
- **Panel B (Veränderungen 2013-2014):** Auch bei den *Veränderungen* der Kovariaten zwischen 2013 und 2014 gibt es Unterschiede (z.B. Arbeitslosenquote fiel in Expansionsstaaten schneller) [source: 271, 272]. Wenn diese Kovariaten-Änderungen die Outcome-Änderung (ΔY(0)) beeinflussen, könnte auch das auf eine Verletzung der PT-Annahme hindeuten

Table 4: Covariate Balance Statistics

	Unweighted			Weighted		
Variable	Non-Adopt	Adopt	Norm. Diff.	Non-Adopt	Adopt	Norm. Diff.
2013 Covariate Levels						
% Female	49.43	49.33	-0.03	50.48	50.07	-0.24
% White	81.64	90.48	0.59	77.91	79.54	0.11
% Hispanic	9.64	8.23	-0.10	17.01	18.86	0.11
Unemployment Rate	7.61	8.01	0.16	7.00	8.01	0.50
Poverty Rate	19.28	16.53	-0.42	17.24	15.29	-0.37
Median Income	43.04	47.97	0.43	49.31	57.86	0.68
2014 - 2013 Covariate Differences						
% Female	-0.02	-0.02	0.00	0.02	0.01	-0.09
% White	-0.21	-0.21	0.01	-0.32	-0.33	-0.04
% Hispanic	0.20	0.21	0.04	0.25	0.33	0.29
Unemployment Rate	-1.16	-1.30	-0.21	-1.08	-1.36	-0.55
Poverty Rate	-0.55	-0.28	0.14	-0.41	-0.35	0.05
Median Income	0.98	1.11	0.06	1.10	1.74	0.32

Notes: This table reports the covariate balance between adopting and non-adopting states. In the top panel, we report the averages and standardized differences of each variable, measured in 2013, by adoption status. All variables are measured in percentage values, except for median household income, which is measured in thousands of U.S. dollars. In the bottom panel we report the average and standardized differences of the county-level long differences between 2014 and 2013 of each variable. We report both weighted and unweighted measures of the averages to correspond to the different estimation methods of including covariates in a 2×2 setting.

Wichtige Einschränkung: Kovariate oder Mechanismus/Outcome? [source: 274]

- Die Interpretation der Unterschiede in den *Veränderungen* (Panel B) ist knifflig.
- Wenn eine Variable X vom Treatment **unbeeinflusst** ist (eine echte "exogene" Kovariate), dann deuten unterschiedliche Trends in X zwischen den Gruppen stark auf eine Verletzung der PT-Annahme hin (irgendetwas anderes muss die unterschiedlichen Trends verursachen) [source: 275, 276]. Beispiel: Arbeitslosenquote könnte hier als exogen angesehen werden [source: 277].

- Wenn aber das Treatment die Variable X **selbst beeinflussen kann** (z.B. wenn Medicaid Expansion die Armutsquote oder die Bevölkerungszusammensetzung ändert), dann ist X teilweise ein *Outcome* oder *Mechanismus* des Treatments [source: 278]. In diesem Fall spiegeln unterschiedliche Trends in X nach dem Treatment nicht notwendigerweise eine Verletzung der PT-Annahme für das ursprüngliche Outcome Y(0) wider, sondern können Teil des *kausalen Effekts* sein [source: 279, 280]. Man spricht hier von "Bad Controls".
- **Fazit:** Ob eine Variable eine echte Kovariate (geeignet für Balance Checks oder Kontrolle) oder ein Mechanismus/Outcome ist, kann man nicht allein aus den Daten entscheiden. Man braucht **kontextspezifisches Wissen oder Annahmen** darüber, wie das Treatment wirkt [source: 281].

Wie identifiziere ich Did mit kontrollvariablen? weitere Annahmen SO (Strong Overlapp)

Entwickle also nun eine alternative Identifkationsstrategie wenn die simple PT Annahme verletzt ist!

Die Kernidee: Bedingte Parallele Trends (Conditional Parallel Trends - CPT)

1. **Das Problem:** Die einfache Parallel-Trends-Annahme (PT) sagt: Die Trends wären *insgesamt* parallel gewesen. Wenn sich aber z.B. arme und reiche Landkreise unterschiedlich entwickeln und die Gruppen unterschiedlich viele arme/reiche Landkreise haben, stimmt das *insgesamt* vielleicht nicht.
2. **Die Lösungsidee:** Wir schwächen die Annahme ab. Statt paralleler Trends *insgesamt*, nehmen wir nur an, dass die Trends parallel sind *innerhalb von Gruppen, die sich in beobachtbaren Merkmalen (Kovariaten X) ähneln* [source: 283, 290].
3. **Definition von X_i:** X_i ist ein Vektor von Variablen, von denen angenommen wird, dass sie die Trends des unbehandelten Outcomes Y(0) beeinflussen [source: 284]. Das können zeitkonstante Merkmale sein oder Werte aus der Vorher-Periode (t=1) [source: 285]. (Der Text erwähnt auch die Möglichkeit, Werte aus t=2 einzubeziehen, aber der Fokus hier liegt auf Variablen, die die Vergleichbarkeit *vor* dem Treatment herstellen).
4. **Assumption CPT (Conditional Parallel Trends):**
    - **Formal (Gleichung 4.1) [source: 289]:** $E[Y_2(0) - Y_1(0) \mid X_i, D=1] = E[Y_2(0) - Y_1(0) \mid X_i, D=0]$
    - **In Worten:** Die durchschnittliche Veränderung des unbehandelten Outcomes zwischen Periode 1 und 2 ist für behandelte (D=1) und unbehandelte (D=0) Einheiten **gleich**, *vorausgesetzt, sie haben denselben Wert für die Kovariaten X_i*.
    - **Intuition:** Die Annahme muss nicht mehr für alle gelten, sondern nur noch für *vergleichbare* Einheiten. Beispiel: Der Trend für arme Expansions-Counties wäre derselbe gewesen wie für arme Nicht-Expansions-Counties; der Trend für reiche Expansions-Counties wäre derselbe gewesen wie für reiche Nicht-Expansions-Counties usw. für alle Kovariatenkombinationen [source: 290]. Die *Gesamttrends* der Gruppen dürfen sich unterscheiden, wenn ihre Zusammensetzung bezüglich X unterschiedlich ist, aber *innerhalb* der durch X definierten Untergruppen wird Parallelität angenommen [source: 291].

Eine notwendige Zusatzannahme: Starker Overlap (Strong Overlap - SO)

5. **Problem:** Damit die CPT-Annahme überhaupt Sinn ergibt und die bedingten Erwartungswerte auf beiden Seiten von Gleichung 4.1 definiert sind, muss es für jeden möglichen Wert der Kovariaten X_i *sowohl* behandelte (D=1) als auch unbehandelte (D=0) Einheiten geben [source: 292, 293]. Gäbe es z.B. einen Typ von County (definiert durch X), der nur in der Treatmentgruppe vorkommt, hätten wir keinen Vergleichspartner in der Kontrollgruppe mit denselben Eigenschaften [source: 293].
6. **Assumption SO (Strong Overlap):**
    - **Formal [source: 296]:** $\varepsilon < P(D=1 \mid X_i) < 1 - \varepsilon$ für ein kleines $\varepsilon > 0$.
    - **In Worten:** Für jeden Wert der Kovariaten X_i ist die (bedingte) Wahrscheinlichkeit, behandelt zu werden, immer echt größer als Null und echt kleiner als Eins. Es gibt also

keine Merkmalskombination X_i, die *deterministisch* vorhersagt, ob eine Einheit behandelt wird oder nicht. Man spricht auch von "Common Support".

Identifikation des ATT(2) unter CPT und SO

Unter diesen beiden Annahmen (CPT und SO) sowie der No-Anticipation-Annahme können die Autoren nun zeigen, wie man den ATT(2) identifizieren kann (d.h., wie man ihn aus beobachtbaren Daten berechnen kann). Die Herleitung erfolgt in Gleichung 4.2 [source: 301]:

- **Zeile 1:** Startet mit der Definition des ATT(2): E[Y_t=2(1) | D=1] - E[Y_t=2(0) | D=1]. Der erste Teil ist E[Y_t=2 | D=1] (beobachtbar), der zweite Teil ist das unbeobachtete kontrafaktische Outcome.
- **Zeile 2:** Wendet das Gesetz der iterierten Erwartungen an. Der Durchschnitt des kontrafaktischen Outcomes E[Y_t=2(0) | D=1] kann geschrieben werden als der Durchschnitt der *bedingten* kontrafaktischen Outcomes E[Y_t=2(0) | X_i, D=1], gemittelt über die Verteilung von X_i in der Treatmentgruppe.
- **Zeile 3: Hier wird die CPT-Annahme (4.1) angewendet!** Sie erlaubt es, den unbeobachtbaren *bedingten* kontrafaktischen Trend für die Behandelten (E[ΔY_t=2(0) | X_i, D=1]) durch den beobachtbaren *bedingten* Trend der Unbehandelten (E[ΔY_t=2(0) | X_i, D=0] = E[ΔY_t=2 | X_i, D=0]) zu ersetzen. Man nutzt auch die No-Anticipation-Annahme (Y_t=1(0) = Y_t=1).
- **Zeile 4:** Fasst die Terme neu zusammen und kommt zur finalen Identifikationsformel: ATT(2) = E[ΔY_t=2 | D=1] - E[E[ΔY_t=2 | X_i, D=0] | D=1]

Intuition der finalen Formel:

- E[ΔY_t=2 | D=1]: Das ist einfach die beobachtete durchschnittliche Veränderung des Outcomes in der Treatmentgruppe.
- E[ΔY_t=2 | X_i, D=0]: Das ist die beobachtete durchschnittliche Veränderung des Outcomes für die Einheiten in der **Kontrollgruppe**, die die **gleichen Merkmale X_i** haben wie eine bestimmte Einheit in der Treatmentgruppe. Dies ist unsere Schätzung des kontrafaktischen Trends für vergleichbare Einheiten.
- E[... | D=1]: Das ist der entscheidende äußere Erwartungswert. Er bedeutet: Wir nehmen den Trend der Kontrollgruppe, der von X_i abhängt (E[ΔY_t=2 | X_i, D=0]), und mitteln diesen Wert über alle Einheiten der **Treatmentgruppe (D=1)**, entsprechend ihrer Verteilung der Merkmale X_i.

Was heißt das praktisch? Man berechnet den Trend für jeden "Typ" X_i in der Kontrollgruppe und bildet dann einen gewichteten Durchschnitt dieser Trends, wobei die Gewichte der Häufigkeit entsprechen, mit der jeder Typ X_i in der *Treatmentgruppe* vorkommt. Man passt also den durchschnittlichen Kontrollgruppentrend so an, dass er die Zusammensetzung der Treatmentgruppe widerspiegelt.

Fazit des Abschnitts: Wenn die einfache PT-Annahme wegen Kovariatenunterschieden nicht hält, kann man stattdessen die CPT- und SO-Annahmen treffen. Unter diesen Annahmen ist der ATT(2) immer noch identifiziert und kann durch die Formel in Gleichung 4.2 [source: 301] ausgedrückt werden. Diese Formel berücksichtigt die Kovariatenunterschiede, indem sie den Kontrollgruppentrend effektiv an die Zusammensetzung der Treatmentgruppe anpasst.

Estimation mit Covariates :TWFE

Ausgangspunkt und Problemstellung:

1. **Herausforderung bei der Schätzung:** Abschnitt 4.2 hat gezeigt, wie der ATT(2) unter der Annahme bedingter paralleler Trends (CPT) theoretisch identifiziert wird [source: 301]. Die praktische Schätzung dieses Ergebnisses (Formel 4.2) ist jedoch schwierig, wenn die Kovariaten X_i kontinuierlich sind oder es sehr viele diskrete Kovariaten gibt. Man kann dann den benötigten bedingten Erwartungswert E[ΔY | X_i, D=0] nicht einfach durch simple Mittelwertbildung für jede Kovariatenkombination schätzen [source: 303, 304]. Man braucht also fortgeschrittenere ökonometrische Techniken [source: 304].
2. **Die gängige Praxis: TWFE mit Kovariaten:** Bevor die Autoren die empfohlenen Techniken vorstellen, diskutieren sie den *häufigsten* Ansatz in der Praxis: Man erweitert einfach die TWFE-Regression (die im 2x2-Fall *ohne* Kovariaten funktionierte) um Kovariaten [source: 306, 307]. Typische Spezifikationen sind:
    - **Modell 4.3:** Y_it = θ_t + η_i + β_treat*D_it + X_it'*β_covs + e_it. Hier werden die Kovariaten X_it direkt als zeitveränderliche Variablen aufgenommen [source: 308].

- **Modell 4.4:** Y_it = θ_t + η_i + β_treat,2*D_it + (1{t=2}*X_{i,t=1})'*β_covs,2 + e_it. Hier werden die *Baseline*-Kovariaten (X_{i,t=1}) interagiert mit einem Post-Treatment-Dummy aufgenommen [source: 309].

3. **Empirisches Problem (Tabelle 5):** Die Autoren verweisen auf Tabelle 5 [source: 310], die zeigt, dass diese unterschiedlichen TWFE-Spezifikationen (mit Baseline-Kovariaten vs. zeitveränderlichen Kovariaten) zu *unterschiedlichen Schätzwerten* für den Behandlungseffekt führen können [source: 314, 315, 316, 317]. Das wirft die Frage auf, was diese Koeffizienten (β_treat, β_treat,2) eigentlich messen.

A typical regression specification is

$$Y_{i,t} = \theta_t + \eta_i + \beta_{treat} D_{i,t} + X'_{i,t}\beta_{covs} + e_{i,t}, \tag{4.3}$$

where the unit and time fixed effects, treatment status, and covariates have already been defined, $e_{i,t}$ is an error term, and β_{treat} is interpreted as the parameter of interest. A related specification explicitly controls for baseline covariates by replacing $X_{i,t}$ with interactions of the pre-treatment covariates and a post-treatment dummy,

$$Y_{i,t} = \theta_t + \eta_i + \beta_{treat,2} D_{i,t} + (\mathbf{1}\{t = 2\} X_{i,t-1})\beta_{covs,2} + e_{i,t}, \tag{4.4}$$

In Table 5, we report the OLS and weighted least squares estimates of the unconditional 2 × 2 DiD estimate, β_{covs} from (4.4), $\beta_{covs,2}$, from (4.3), and their cluster-robust standard errors using the covariates from Table 4.

Table 5: Regression 2 × 2 DiD with Covariates

	Unweighted			Weighted		
	No Covs	$X_{i,t=2013}$	$X_{i,t}$	No Covs	$X_{i,t=2013}$	$X_{i,t}$
	(1)	(2)	(3)	(4)	(5)	(6)
Medicaid Expansion	0.12	-2.35	-0.49	-2.56*	-2.56	-1.37
	(3.75)	(4.29)	(3.83)	(1.49)	(1.78)	(1.62)

Notes: This table reports the regression 2 × 2 DiD estimate comparing counties that expand Medicaid in 2014 to counties that do not expand Medicaid, adjusting for the inclusion of covariates (percent female, percent white, percent hispanic, the unemployment rate, the poverty rate, and median household income). Columns 1-3 report unweighted regression results, while columns 4-6 weight by county population aged 20-64 in 2013. Columns 1 and 4 report results for expansion states without covariates, columns 2 and 5 adjust for the baseline levels of the covariates in 2013, and columns 3 and 6 control for the time-varying covariate values in 2014 and 2013. Standard errors (in parentheses) are clustered at the county level.

Kritik am TWFE-Ansatz mit Kovariaten:

1. **Kernfrage:** Entspricht der Koeffizient β_treat aus diesen Regressionen dem Zielparameter ATT(2) unter der CPT-Annahme? [source: 318] **Antwort: Im Allgemeinen nein.** Die enge Verbindung zwischen TWFE-Regression und ATT, die im 2x2-Fall ohne Kovariaten bestand, geht hier verloren [source: 319].
2. **Gründe für das Versagen von TWFE:**
   - **Was wird kontrolliert?** Eine Standard-TWFE-Regression (Modell 4.3) kontrolliert aufgrund der Einheits-Fixed-Effects (η_i) effektiv nur für die *Veränderungen* der Kovariaten (ΔX_it), nicht für deren Levels [source: 322]. Um für Baseline-Levels zu kontrollieren, braucht man Interaktionsterme wie in Modell 4.4 [source: 322]. Die Wahl der Spezifikation impliziert also unterschiedliche Annahmen darüber, *worauf* die CPT-Annahme eigentlich konditioniert [source: 323]. "Kontrollieren für jährliche Armutsquote" bedeutet in TWFE oft "Kontrollieren für die *Änderung* der Armutsquote" [source: 324].
   - **"Bad Controls":** Wenn zeitveränderliche Kovariaten (X_it in t=2) selbst vom Treatment beeinflusst werden (z.B. Armutsrate sinkt durch Medicaid), dann führt das Kontrollieren für diese Variablen (wie in Modell 4.3 über ΔX) zu verzerrten Schätzungen des ATT(2) [source: 325, 326]. Dies knüpft an die Diskussion in 4.1 an: Man darf nur für echte Kovariaten kontrollieren, nicht für Outcomes/Mechanismen [source: 327].
   - **Heterogene Effekte & Gewichtung:** Selbst wenn CPT gilt und die Kovariaten "gut" sind, ist β_treat problematisch, wenn der Behandlungseffekt über die Kovariatenwerte hinweg *heterogen* ist (d.h., ATT_X(2) ist nicht für alle X gleich) [source: 331]. Neuere Forschung (z.B. Caetano & Callaway 2024 [source: 331], Sloczynski 2022 [source: 334]) zeigt: β_treat ist dann zwar ein gewichteter Durchschnitt der bedingten Effekte ATT_X(2), aber die Gewichte sind oft unintuitiv, können negativ sein (!), und geben z.B. Gruppen, die unter den Behandelten relativ selten sind (verglichen mit den

Unbehandelten), unverhältnismäßig viel Gewicht [source: 336]. Das kann sogar dazu führen, dass β_treat ein anderes Vorzeichen hat als alle tatsächlichen bedingten Effekte ATT_X(2) [source: 335].

- **Implizite Annahme konstanter Effekte:** Zusammengenommen bedeutet das: Die TWFE-Koeffizienten β_treat aus Modellen wie 4.3 oder 4.4 identifizieren den ATT(2) unter CPT eigentlich nur dann korrekt, wenn man zusätzlich annimmt, dass der **Behandlungseffekt über alle Kovariatengruppen hinweg konstant** ist [source: 337, 338].
  In other words, (4.3) implicitly rules out that treatment effects can vary across covariate-strata,
  which makes the weighting issues identified by Caetano and Callaway (2024) irrelevant to the
  interpretation of βtreat Diese Annahme **über konstante Behandlungseffekte** ist aber oft unrealistisch (wie Studien zur Medicaid Expansion zeigen:
  Research on the Medicaid expansion using data on mortality rates by
  income shows clear evidence of heterogeneous effects (Miller et al., 2021; Wyse and Meyer, 2024),
-) [source: 339].

Ausblick und Empfehlung:

1. **Mögliche Auswege:** Man könnte versuchen, die TWFE-Modelle durch Hinzufügen vieler Interaktionsterme sehr flexibel zu gestalten [
   One way to avoid these limitations would be to make (4.4) (or (4.3)) more flexible by including interactions of the covariates with treatment group, time, and treatment-group-by-time] ODER man wählt den im Paper empfohlenen **"Forward-Engineering"-Ansatz** [source: 341].
2. **Forward-Engineering:** Dieser Ansatz leitet Schätzer (wie RA, IPW, DR, die in 4.4 kommen) direkt aus den gewünschten Annahmen (NA, CPT, SO) und dem Zielparameter (ATT(2)) ab, anstatt ein Standard-Regressionstool zu nehmen und zu hoffen, dass es das Richtige tut [source: 342, 344]. Dies führt zu mehr Transparenz darüber, was geschätzt wird und unter welchen Annahmen [source: 345, 347]. Der Nachteil ist, dass man eventuell neue Methoden lernen muss [source: 348].

Fazit des Abschnitts 4.3: Der gängige und einfache Ansatz, Kovariaten durch Hinzufügen zu einer Standard-TWFE-Regression zu kontrollieren, ist bei heterogenen Behandlungseffekten und potenziellen "Bad Controls" mit erheblichen Problemen behaftet. Der Regressionskoeffizient für das Treatment identifiziert in der Regel nicht den gewünschten ATT(2), sondern einen komplex gewichteten Durchschnitt, der irreführend sein kann. Das Paper plädiert daher für alternative Methoden, die direkt auf den zugrundeliegenden Annahmen aufbauen.

4.4 Wie beschreibt man ATT(2) mit covariates alternativ zu TWFE

Erinnerung ATT(2): ATT(2) = E[Y_i,2(1) - Y_i,2(0) | D_i=1] TT(2) ist der **durchschnittliche kausale Effekt der Behandlung für die tatsächlich behandelten Einheiten in der Periode nach der Behandlungs-Einführung**. Er beantwortet die Frage: "Um wie viel hat sich das Outcome für die Behandelten in Periode 2 durch die Behandlung im Durchschnitt verändert, verglichen mit dem hypothetischen Szenario, dass sie keine Behandlung erhalten hätten?"

in diesem Abschnitt nun die Schätzer vorgestellt, die diesem Forward-Engineering-Ansatz (aus vorherigen Abschnitt gerecht werden.

Dieser Abschnitt stellt die **Alternativen zur problematischen TWFE-Regression** vor, wenn man Kovariaten in die DiD-Analyse einbeziehen möchte. Diese Methoden folgen dem "Forward-Engineering"-Ansatz, d.h., sie bauen direkt auf der Identifikationsformel für ATT(2) unter Conditional Parallel Trends (CPT) und Strong Overlap (SO) auf, die in Abschnitt 4.2 hergeleitet wurde (Formel 4.2

[source: 301]).

Ausgangspunkt: Die Identifikationsformel (4.2)

Zur Erinnerung, unter CPT und SO gilt [source: 351]: ATT(2) = E[ΔY | D=1] - E[E[ΔY | X_i, D=0] | D=1]

- Der erste Term E[ΔY | D=1] ist einfach der durchschnittliche Outcome-Trend in der Treatmentgruppe (leicht zu schätzen).
- Der zweite, kompliziertere Term ist der "angepasste" Kontrollgruppentrend. Wir brauchen eine Methode, um E[ΔY | X_i, D=0] (den Trend der Kontrollgruppe für einen bestimmten Kovariaten-Typ X_i) zu schätzen und diesen dann über die Kovariatenverteilung der Treatmentgruppe zu mitteln (E[... | D=1]).

Abschnitt 4.4 stellt drei Hauptmethoden vor, um dies zu erreichen:

1. Regression Adjustment (RA) / Outcome Regression [source: 359]

- **Grundidee:** Man schätzt direkt die Beziehung zwischen den Kovariaten X_i und dem Outcome-Trend ΔY *nur für die Kontrollgruppe (D=0)* [source: 353].
- **Vorgehen:**
  1. **Modellieren:** Man spezifiziert ein Modell (oft eine lineare Regression: ΔY = X'β) dafür, wie ΔY von X in der Kontrollgruppe abhängt (μ(X) = X'β) [source: 354, 355].
  2. **Schätzen:** Man schätzt die Parameter dieses Modells (β_D=0) nur mit den Daten der Kontrollgruppe (siehe Tabelle 6, Spalten 1 & 3 für die geschätzten Koeffizienten im Beispiel) [source: 355, 360].
  3. **Vorhersagen:** Man benutzt das geschätzte Modell (μ_hat(X)), um für *jede Einheit* (auch die aus der Treatmentgruppe) vorherzusagen, welchen Trend ΔY sie gehabt hätte, wenn sie zur Kontrollgruppe gehört hätte und nur ihre Kovariaten X_i den Trend bestimmen würden [source: 357].
  4. **Mitteln & ATT berechnen:** Man mittelt diese vorhergesagten Trends μ_hat(X_i) über alle Einheiten der **Treatmentgruppe**. Das ergibt die Schätzung für den zweiten Term der ATT(2)-Formel. Der ATT(2) wird dann nach Formel 4.5 [source: 359] als durchschnittliche Differenz zwischen dem tatsächlichen Trend ΔY_i und dem vorhergesagten kontrafaktischen Trend μ_hat(X_i) für die behandelten Einheiten berechnet. (Im Beispiel ergibt das -1.62 ungewichtet und -3.46 gewichtet, siehe Tabelle 7, Spalte "Regression" [source: 365, 366]).
- **Vorteile/Nachteile:** Relativ intuitiv. Die Güte des Ergebnisses hängt **entscheidend** davon ab, ob das Modell für den Outcome-Trend der Kontrollgruppe **korrekt spezifiziert** ist [source: 383, 384]. Ist das Modell falsch (z.B. weil wichtige Variablen fehlen oder die Beziehung nicht-linear ist), ist der RA-Schätzer verzerrt [source: 384]. Man kann flexiblere Modelle (nicht nur linear) verwenden, wenn die Daten es erlauben [source: 374].

2. Inverse Probability Weighting (IPW) [source: 387]

- **Grundidee:** Statt den Outcome-Trend zu modellieren, modelliert man die **Wahrscheinlichkeit, behandelt zu werden**, abhängig von den Kovariaten (p(X) = P(D=1 | X), der sogenannte Propensity Score) [source: 390]. Man gewichtet dann die Kontrollgruppe so um, dass sie hinsichtlich der Kovariaten X der Treatmentgruppe "ähnlicher" wird [source: 388, 389]. Die Idee ist, der Kovariaten-Imbalance direkt entgegenzuwirken.
- **Vorgehen:**
  1. **Modellieren:** Man spezifiziert ein Modell (oft Logit oder Probit) für den Propensity Score p(X) [source: 407].
  2. **Schätzen:** Man schätzt dieses Modell mit *allen* Daten (Treatment- und Kontrollgruppe) (siehe Tabelle 6, Spalten 2 & 4 für Logit-Koeffizienten im Beispiel) [source: 408]. Man erhält für jede Einheit einen geschätzten Propensity Score p_hat(X_i).
  3. **Gewichte berechnen:** Man berechnet spezielle IPW-Gewichte für jede Einheit basierend auf p_hat(X_i) (Formeln in 4.7 [source: 395]). Grob gesagt: Behandelte Einheiten erhalten ein Gewicht, das ihrem Anteil entspricht. Unbehandelte Einheiten (D=0) erhalten ein Gewicht, das umso größer ist, je *höher* ihr p_hat(X_i) ist (d.h., je mehr sie den Behandelten ähneln) [source: 399].
  4. **ATT berechnen:** Man berechnet den gewichteten Durchschnitt des Outcome-Trends ΔY für die behandelte Gruppe und subtrahiert den gewichteten Durchschnitt von ΔY für die

(umgewichtete) Kontrollgruppe (Formel 4.8 [source: 411]). (Im Beispiel ergibt das -0.86 ungewichtet und 0.18 gewichtet, siehe Tabelle 7, Spalte "IPW" [source: 412, 413]).

- **Vorteile/Nachteile:** Man muss den Outcome-Prozess nicht modellieren [source: 404]. Funktioniert nur gut, wenn die **Overlap-Annahme (SO)** gilt, d.h., wenn es für alle Kovariaten-Typen sowohl Behandelte als auch Unbehandelte gibt [source: 416]. Wenn p_hat(X_i) für manche Unbehandelte sehr nahe bei 1 liegt (oder bei 0 für Behandelte), werden die Gewichte extrem groß/instabil, was zu unpräzisen Schätzungen führt [source: 416, 417]. Man sollte die Verteilung der Propensity Scores prüfen (wie in Figure 1
Few untreated units have very high estimated propensity scores, so extreme weighting is not a significant concern. In addition, propensity scores of non-expansion counties seem to lie within the support of the expansion counties' propensity scores, supporting strong overlap. Trimming high- or low-propensity score observations from the sample may be warranted
when overlap is weak; for a discussion, see, e.g., Crump, Hotz, Imbens and Mitnik (2009), Sasaki
and Ura (2022), and Ma, Sant'Anna, Sasaki and Ura (2023)
).

3. Doubly Robust (DR) / Augmented IPW (AIPW) [source: 426]

- **Grundidee:** Kombiniert die Stärken von RA und IPW, um einen Schätzer zu erhalten, der robuster gegenüber Modellfehlspezifikationen ist [source: 425, 426].
- **Vorgehen:**
  1. **Modellieren & Schätzen:** Man braucht **beides**: ein Modell für den Outcome-Trend der Kontrollgruppe (μ_hat(X), wie bei RA) *und* ein Modell für den Propensity Score (p_hat(X), wie bei IPW) [source: 429, 435].
  2. **ATT berechnen:** Man verwendet eine spezielle Formel (4.10 / 4.11 [source: 432, 435]), die beide geschätzten Modelle und die IPW-Gewichte kombiniert. Die Formel ähnelt der IPW-Formel, enthält aber einen zusätzlichen Term, der auf dem Outcome-Modell basiert.
- **Vorteile/Nachteile:** Die namensgebende **"doppelte Robustheit"**: Der DR-Schätzer ist konsistent (im Durchschnitt korrekt bei großen Stichproben), wenn **entweder** das Outcome-Modell **oder** das Propensity-Score-Modell korrekt spezifiziert ist – man braucht nicht beide! [source: 429, 430] Wenn beide Modelle leicht falsch sind, ist DR oft immer noch besser als RA oder IPW allein [source: 431]. **Dies ist die von den Autoren generell empfohlene Methode** [source: 444]. (Im Beispiel ergibt das -1.23 ungewichtet und 0.49 gewichtet, siehe Tabelle 7, Spalte "Doubly Robust" [source: 441]). Es gibt auch Anpassungen für Fälle mit schlechtem Overlap [source: 447].

Fazit des Abschnitts 4.4: Statt auf problematische TWFE-Regressionen zurückzugreifen, sollte man bei DiD mit Kovariaten Methoden wie Regression Adjustment, Inverse Probability Weighting oder (bevorzugt) Doubly Robust Estimation verwenden. Diese bauen direkt auf der Identifikation unter CPT auf und haben klarere Eigenschaften, insbesondere der DR-Ansatz durch seine Robustheit gegenüber Fehlspezifikationen.

4.5 Weitere Heterogenitätsanalysen mit covariaten und DiD

Grundidee: Über den Durchschnitt hinausblicken

1. **ATT als Durchschnitt:** Das bisherige Hauptziel war die Schätzung des *durchschnittlichen* Behandlungseffekts für die Behandelten, ATT(2). Dieser Durchschnitt kann jedoch wichtige Unterschiede im Effekt über verschiedene Untergruppen hinweg verdecken [source: 477].
2. **Heterogenität:** Oft ist es interessant oder wichtig zu wissen, ob der Behandlungseffekt für verschiedene Gruppen, die durch ihre Kovariaten X definiert sind, unterschiedlich ausfällt [source: 450]. Beispiel: Ist der Effekt von Medicaid auf die Mortalität in armen Landkreisen

anders als in reichen? [source: 238]

3. **Ziel des Abschnitts:** Dieser Abschnitt zeigt, wie man solche **bedingten Durchschnittseffekte** (Conditional Average Treatment Effects on the Treated, kurz ATT_X(2)) identifizieren und schätzen kann, also den ATT *spezifisch für Einheiten mit bestimmten Kovariatenwerten X*.

Identifikation von bedingten ATTs (ATT_X(2))

1. **Baustein-Struktur:** Die Autoren zeigen zuerst, dass der Gesamt-ATT(2) (aus Formel 4.2) mathematisch als ein Durchschnitt von bedingten DiD-Vergleichen geschrieben werden kann [source: 449].
2. **Implikation:** Wenn die Annahmen CPT (Conditional Parallel Trends) und SO (Strong Overlap) gelten (die ja für die Identifikation des Gesamt-ATT(2) mit Kovariaten nötig waren), dann erlauben sie auch die Identifikation der **bedingten ATTs** für jede Kovariaten-Kombination X_i [source: 449]: ATT_X(2) = E[ΔY | X_i, D=1] - E[ΔY | X_i, D=0] Das heißt, für jede spezifische Gruppe X_i ist der ATT einfach die Differenz der durchschnittlichen Outcome-Trends zwischen Behandelten und Unbehandelten *innerhalb dieser spezifischen Gruppe*.

Schätzung von bedingten ATTs

Die Schätzung hängt davon ab, ob die Kovariaten diskret oder kontinuierlich sind:

1. **Fall 1: Diskrete Kovariaten:**
   - **Vorgehen:** Wenn alle Kovariaten diskret sind (oder diskretisiert wurden), kann man für jede mögliche Kombination der Kovariatenwerte eine eigene Untergruppe (Partition x_k) bilden [source: 452].
   - **Schätzung:** Innerhalb jeder dieser Untergruppen k schätzt man ATT_{x_k}(2) einfach durch einen **standardmäßigen 2x2 DiD-Vergleich** (Vergleich der Trenddifferenzen zwischen Behandelten und Unbehandelten nur für diese Untergruppe) [source: 452, 454]. Man kann dafür einfache Mittelwertdifferenzen oder eine einfache DiD-Regression (wie in Gleichung 3.7) verwenden, aber eben nur mit den Daten der jeweiligen Untergruppe [source: 455].
   - **Beispiel:** Man könnte die USA in Zensusregionen aufteilen und für jede Region einen separaten (unbedingten) DiD-Effekt schätzen [source: 457, 458].
   - **Voraussetzung:** Jede Untergruppe muss genügend Beobachtungen (behandelte und unbehandelte) enthalten, damit die Schätzung und Inferenz zuverlässig sind [source: 455, 458].
2. **Fall 2: Kontinuierliche oder viele Kovariaten:**
   - **Problem:** Wenn Kovariaten kontinuierlich sind oder es sehr viele diskrete gibt, kann man nicht mehr für jede einzelne Kovariatenkombination X_i einen separaten Effekt schätzen (zu wenige Beobachtungen pro "Gruppe") [source: 459].
   - **Lösung 1: Aggregierte bedingte ATTs (ATT_k(2)):** Man bildet **gröbere Partitionen** (PART(X) = k), die auf den Kovariaten basieren [source: 461, 462]. Beispiel: Landkreise mit Arbeitslosenquote über dem Median (k=1) vs. unter dem Median (k=2). Man definiert dann ATT_k(2) als den ATT für die behandelten Einheiten *innerhalb der Partition k* [source: 462].
   - **Identifikation/Schätzung von ATT_k(2):** Innerhalb jeder Partition k wird ATT_k(2) analog zum Gesamt-ATT(2) unter CPT identifiziert und geschätzt [source: 462, 463]. Man kann die Methoden aus Abschnitt 4.4 (RA, IPW, DR) verwenden, aber **lokalisiert** für jede Partition k [source: 463, 464].
   - **Zusammenhang zum Gesamt-ATT:** Der Gesamt-ATT(2) ist dann der gewichtete Durchschnitt der partitions-spezifischen ATT_k(2), wobei die Gewichte dem Anteil der jeweiligen Partition an der Treatmentgruppe entsprechen (P(PART(X)=k | D=1)) [source: 462].
   - **Lösung 2: Granularere Analysen:** Alternativ zu groben Partitionen kann man auch untersuchen, wie der ATT über den Wertebereich einer *einzelnen* (oft kontinuierlichen) Kovariate variiert oder die beste lineare Annäherung an diesen Zusammenhang schätzen (vgl. Abadie 2005 [source: 467]) [source: 469, 470, 471]. Das erfordert keine Diskretisierung.

Herausforderungen bei der Heterogenitätsanalyse:

1. **Wahl der Untergruppen:** Wie definiert man sinnvolle Untergruppen oder Partitionen? [source: 474] Es gibt oft keine klare theoretische Vorgabe [source: 476]. Zu viele Untergruppen führen zu unpräzisen Schätzungen, zu wenige können wichtige Unterschiede verdecken [source: 477,

478]. Hier ist ein Mittelweg gefragt [source: 478].

2. **Bedarf an Forschung:** Die Autoren merken an, dass es in der Praxis noch an etablierten Methoden und klarer Anleitung für Heterogenitätsanalysen im DiD-Kontext mangelt [source: 482, 483]. Sie verweisen auf mögliche zukünftige Richtungen wie die Adaption von "Sorted Effects"-Methoden [source: 480].

Fazit des Abschnitts 4.5: Die Annahmen, die die Identifikation des ATT unter Kovatenkontrolle ermöglichen (CPT, SO), erlauben auch die Identifikation und Schätzung von bedingten ATTs (ATT_X(2)). Dies ermöglicht die Untersuchung von Behandlungseffekt-Heterogenität. Die konkrete Umsetzung hängt von der Art der Kovariaten ab (diskret vs. kontinuierlich), wobei oft Partitionen gebildet oder Effekte entlang einzelner Kovariaten geschätzt werden. Die Wahl der relevanten Untergruppen bleibt jedoch eine praktische Herausforderung.

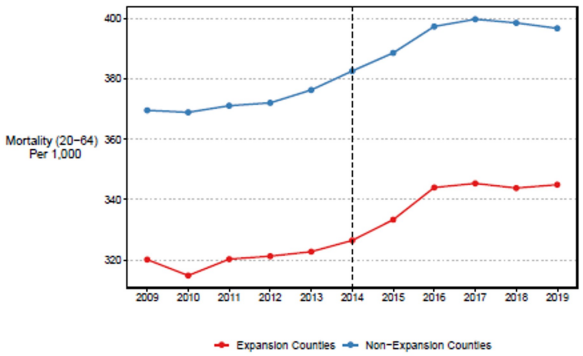
Dif in Dif mit mutiplen time periods aber zwei gruppen, also 2xT und GxT Estimate

Kapitel 5: DiD-Designs mit multiplen Zeitperioden [source: 484]

- **Ausgangspunkt:** Die vorherigen Abschnitte konzentrierten sich auf das 2x2 DiD-Grundmodell. Nun wird der Blick auf Designs erweitert, bei denen Daten für **mehr als zwei Zeitperioden** (T > 2) vorliegen [source: 484].
- **Neue Möglichkeiten:** Mehrere Zeitperioden eröffnen zwei wichtige neue Analysemöglichkeiten:
  1. **Dynamik des Effekts:** Man kann den ATT(t) für *jede* Periode *nach* dem Treatmentbeginn (t ≥ g) schätzen und so untersuchen, wie sich der Behandlungseffekt über die Zeit entwickelt (z.B. Anstieg, Abfall, Konstanz) [source: 485, 501].
  2. **Plausibilität der Parallel-Trends-Annahme prüfen:** Man kann die Outcome-Trends der Behandlungs- und Kontrollgruppe *vor* dem Treatmentbeginn vergleichen ("Pre-Trends"), um die Glaubwürdigkeit der Parallel-Trends-Annahme (die sich formal auf die Post-Treatment-Periode bezieht) zu beurteilen [source: 485, 515].
- **Fokus des Kapitels:** Das Kapitel wird sich später auf komplexere Fälle mit gestaffeltem Treatmentbeginn konzentrieren, beginnt aber mit dem einfacheren Fall von nur zwei Gruppen (eine Treatment-, eine Kontrollgruppe), aber vielen Zeitperioden (2xT) [source: 486, 493].
- **Notation:** Die Notation wird leicht erweitert. Die Zeit läuft nun von t=1 bis T [source: 490]. Die Zuordnung zu potenziellen Outcomes bleibt für diesen einfachen Fall gleich: Y_it = D_i*Y_it(1) + (1-D_i)*Y_it(0) [source: 490]. g ist der Zeitpunkt des Treatmentbeginns für die Treatmentgruppe (D_i=1).

5.1 Simple Event Study 2xT

Figure 2: **County Mortality Trends by Expansion Decision**



Notes: This figure shows county population-weighted average mortality rates for adults ages 20-64 by expansion category from 2009 to 2019.

Abschnitt 5.1: Einfache Event Studies (2xT) [source: 492]

- **Definition "Event Study":** Dieser Begriff bezeichnet hier die Schätzung und Darstellung von Behandlungseffekten über einen Zeitverlauf hinweg, relativ zum Zeitpunkt des "Events" (hier: Beginn der Behandlung) [source: 492].
- **Setup:** Man betrachtet den Fall mit nur einer Treatmentgruppe (beginnt Treatment zur Zeit g) und einer Kontrollgruppe über T Perioden [source: 493]. Im Beispiel wird die Medicaid-Expansion von 2014 (g=2014) analysiert, verglichen mit Staaten, die bis 2019 nicht expandierten, über den Zeitraum 2009-2019 (T=11) [source: 494]. Die Ergebnisse werden populationsgewichtet dargestellt [source: 494].
- **Abbildung 2 (Figure 2): County Mortality Trends**
  - **Inhalt:** Die Abbildung zeigt die durchschnittlichen (populationsgewichteten) Mortalitätsraten (Outcome Y) für die Gruppe der 2014-Expander (rote Linie, beschriftet als "Expansion Counties") und die der Nicht-Expander (blaue Linie, "Non-Expansion Counties") für jedes Jahr von 2009 bis 2019 [source: 498].
  - **Zweck:** Sie visualisiert die Rohdaten-Trends, die der DiD-Event-Study-Analyse zugrunde liegen [source: 495]. Man sieht die Entwicklung der beiden Gruppen vor und nach dem Treatmentjahr 2014 (das die x-Achse teilt) [source: 496].

Abschnitt 5.1.1: Event-Study-Schätzer in den Post-Treatment-Perioden (t >= g) [source: 499]

- **Zielparameter:** Man möchte den ATT(t) für jede einzelne Periode *nach* Beginn der Behandlung ($t \geq g$) schätzen [source: 500].
- **Interpretation:** Die Abfolge dieser ATT(t)-Schätzungen zeigt die **Dynamik des Behandlungseffekts** über die Zeit [source: 501]. Nimmt der Effekt zu, ab, oder bleibt er konstant? [source: 503, 504]
- **Identifikationsannahme (Assumption PT-ES):** Um diese ATT(t)s zu identifizieren, braucht man eine erweiterte Parallel-Trends-Annahme [source: 505]. Sie besagt, dass der durchschnittliche Trend des unbehandelten Outcomes Y(0) zwischen der letzten Vorher-Periode (g-1) und *jeder einzelnen* Nachher-Periode t für die Treatmentgruppe derselbe gewesen wäre wie für die Kontrollgruppe (siehe Formel 5.1 [source: 505]). Wichtig: Um den Effekt z.B. im Jahr 2019 (t= 2019) zu schätzen, muss die Annahme für alle Jahre von 2014 bis 2019 halten. Langfristige Effekte erfordern also stärkere Annahmen als kurzfristige [source: 506, 507].
- **Schätzung:** Unter PT-ES (und No Anticipation) wird jeder einzelne ATT(t) durch einen **simplen 2x2 DiD-Vergleich** geschätzt: Man vergleicht die Veränderung von g-1 zu t in der Treatmentgruppe mit der Veränderung von g-1 zu t in der Kontrollgruppe (Formel 5.2 [source: 509]).
- **Abbildung 3 (Figure 3): 2xT Event Study - Post-Treatment Teil**
  - **Inhalt:** Die Punkte **rechts** von der vertikalen gestrichelten Linie (bei Event Time = 0, was dem Jahr t=g=2014 entspricht) zeigen die geschätzten ATT(t)-Werte für die Jahre 2014, 2015, ..., 2019 [source: 511]. Die x-Achse ist hier als "Event Time" (e = t-g) skaliert, also relativ zum Behandlungsbeginn.
  - **Interpretation:** Der Punkt bei e=0 ist der 2x2 DiD-Effekt für das erste Jahr (-2.7 laut Text, sehr nah am gewichteten Wert -2.6 aus Tabelle 2/3) [source: 511]. Die Punkte bei e=1, 2, ... zeigen die geschätzten Effekte für die Folgejahre, immer relativ zum Niveau im Jahr 2013 (g-1) [source: 512]. Im Beispiel deuten die Punktschätzer auf keine großen Effekte hin [source: 513]. Die vertikalen Linien zeigen Konfidenzintervalle (schwarz: punktweise, rot: simultan für den gesamten Pfad).

Abschnitt 5.1.2: Event-Study-Schätzer in den Pre-Treatment-Perioden (t < g) [source: 515]

- **Zweck:** Die Daten aus der Zeit *vor* der Behandlung ($t < g$) ermöglichen sogenannte **Falsifizierungs- oder Placebo-Tests** [source: 515].
- **Logik:** Die No-Anticipation-Annahme besagt, dass der *wahre* Behandlungseffekt ATT(t) vor dem Treatmentbeginn ($t < g$) gleich Null sein muss [source: 515]. Wenn man nun für diese Vorher-Perioden ebenfalls DiD-Schätzungen berechnet (z.B. vergleicht man die Veränderung von t=g-k zu t=g-1 zwischen den Gruppen, Formel für τ_{-k} [source: 520]), sollten diese Schätzungen idealerweise nahe Null sein.

- **Was wird geschätzt (τ_{-k})?:** Diese Schätzer für die Vorher-Perioden messen die **Differenz in den Outcome-Trends** zwischen Treatment- und Kontrollgruppe *bevor* das Treatment begann [source: 520]. Man nennt sie oft "Pre-Trends" oder "differenzielle Vorher-Trends".
- **Abbildung 3 (Figure 3): 2xT Event Study - Pre-Treatment Teil**
    - **Inhalt:** Die Punkte **links** von der vertikalen gestrichelten Linie (negative Event Time) zeigen diese geschätzten Pre-Trends (τ_{-k}) für die Jahre vor 2014, relativ zum Jahr 2013 (g-1) [source: 520].
- **Interpretation der Pre-Trends (Herausforderungen):**

    Lektion 1: Pre-Trends sind KEIN direkter Test der benötigten Parallel-Trends-Annahme (PT-ES)! [source: 528, 529]
        - Das Fundamentale: Die eigentliche Parallel-Trends-Annahme (PT-ES, Assumption 5.1 [source: 505]) bezieht sich auf die *kontrafaktische* Entwicklung der Post-Treatment-Perioden (t ≥ g) [source: 528]. Da sie kontrafaktisch ist, kann man sie grundsätzlich nicht testen.
        - Was Pre-Trends messen: Die Schätzungen für die Pre-Treatment-Perioden (τ_{-k} für t < g) messen die Differenz der Outcome-Trends zwischen den Gruppen *vorher* [source: 520, 529]. Sie sind nur dann gleich Null, wenn (a) es keine antizipatorischen Effekte gab (No Anticipation) UND (b) die Trends tatsächlich auch *schon vorher* parallel waren.
        - Informativ, aber nicht dasselbe: Flache Pre-Trends (nahe Null) können die PT-ES Annahme *unterstützen* und plausibler machen (besonders wenn man annimmt, dass die Trend-Determinanten stabil sind) [source: 530]. Große Pre-Trends werfen Zweifel auf [source: 531]. Aber: Nur weil die Trends vorher parallel waren, heißt das nicht zwingend, dass sie es auch nachher (im kontrafaktischen Szenario ohne Treatment) gewesen wären [source: 529]. Manchmal sind Pre-Trends auch wenig informativ, wenn die Zeit vor dem Treatment ökonomisch ganz anders war als die Zeit danach [source: 533].
        - Praktische Konsequenz: Sei vorsichtig mit der Formulierung. Sage nicht "Die Pre-Trends zeigen, dass die Parallel-Trends-Annahme erfüllt ist", sondern eher "Die Pre-Trends liefern keine Evidenz gegen die Parallel-Trends-Annahme" oder "Die Pre-Trends werfen Zweifel an der Parallel-Trends-Annahme auf". Verstehe, dass du etwas anderes misst als die eigentliche Annahme.
    - **Statistische Signifikanz vs. Relevanz:** Wie in Abbildung 3 sind Pre-Trends oft statistisch nicht signifikant von Null verschieden [source: 521]. Das kann aber einfach an geringer statistischer Power liegen und schließt problematische Pre-Trends nicht aus [source: 542, 543]. Die *Größe* der geschätzten Pre-Trends ist oft wichtiger als die Signifikanz [source: 549].

    Lektion 2: Die statistische Präzision der Pre-Trend-Schätzer ist entscheidend! [source: 542]
    - **Das Problem:** Tests, die prüfen, ob Pre-Trends signifikant von Null abweichen, haben oft **wenig statistische Power** (geringe Trennschärfe) [source: 542]. Das bedeutet, sie können echte, relevante Unterschiede in den Pre-Trends möglicherweise nicht aufdecken, einfach weil die Schätzungen zu ungenau sind (zu große Standardfehler/weite Konfidenzintervalle).
    - **Fehlinterpretation:** Wenn ein Pre-Trend-Test nicht signifikant ist (wie in Figure 3 [source: 521]), heißt das **nicht** unbedingt, dass die Pre-Trends flach *waren*. Es kann auch einfach heißen, dass die Daten zu "verrauscht" sind, um eine klare Aussage zu treffen [source: 543]. Sich auf solche insignifikanten Ergebnisse zu verlassen und daraus zu schließen, die PT-Annahme sei unproblematisch, kann irreführend sein und sogar zu stärker verzerrten Ergebnissen führen (Roth, 2022 [source: 544]).
    - **Praktische Konsequenz:** Schau nicht nur auf die Signifikanzsterne der Pre-Trends! Bewerte die **Größe der Standardfehler bzw. die Breite der Konfidenzintervalle** [source: 543]. Wenn die Intervalle sehr breit sind (wie in Figure 3), sagen die Pre-Trend-Schätzungen wenig aus – sie schließen weder flache Trends aus noch schließen sie große, problematische Trends aus.

Lektion 3: Nutze Pre-Trends für quantitative Robustheits-/Sensitivitätsanalysen! [source: 545]

- **Der bessere Ansatz:** Statt Pre-Trends nur als einfachen Ja/Nein-Test ("sind sie null?") zu sehen, sollte man sie **quantitativ nutzen**, um die Robustheit der eigentlichen Effektschätzungen (ATT(t) für t ≥ g) zu bewerten [source: 545]. Die Idee ist: Wie stark würden sich meine Ergebnisse ändern, wenn die Parallel-Trends-Annahme in der Post-Periode ähnlich stark verletzt wäre wie in der Pre-Periode?
- **Methoden:**

  Methode 1: Testen gegen alternative Nullhypothesen z.B. eine die gesamten geschätzten Effekt erklären könnte (Bilinski & Hatfield, 2018 [source: 546])Dette und Schumann 2024

  - Das Problem mit dem Standard-Test: Normalerweise testet man die Nullhypothese H0: "Der Pre-Trend ist gleich Null". Wenn der p-Wert groß ist (>0.05), sagt man oft: "Okay, keine signifikanten Pre-Trends, die Parallel-Trends-Annahme scheint zu halten". Das Problem ist aber: Vielleicht ist der Pre-Trend gar nicht Null, aber unsere Daten sind zu ungenau (zu große Standardfehler), um das statistisch nachzuweisen (geringe "Power"). Wir könnten also fälschlicherweise annehmen, alles sei in Ordnung.
  - Die Idee von Bilinski & Hatfield: Statt gegen Null zu testen, testen wir gegen eine spezifische, "problematische" Abweichung. Was wäre ein problematischer Pre-Trend? Einer, der groß genug ist, um unseren geschätzten Behandlungseffekt (ATT_hat) komplett zu erklären.
  - Neuer Test:
    1. Bestimme den "problematischen" Wert: Nimm deinen geschätzten Behandlungseffekt ATT_hat (z.B. -5 Einheiten). Das ist der Wert, den ein Pre-Trend haben müsste, um deinen Effekt wegzuerklären. Nennen wir ihn τ_problem.
    2. Neue Nullhypothese (H0'): Der *wahre* Pre-Trend ist genau gleich diesem problematischen Wert (τ = τ_problem = -5).
    3. Teste H0': Prüfe nun statistisch, ob sich dein *geschätzter* Pre-Trend (τ_hat) signifikant von diesem τ_problem unterscheidet. Du berechnest also, wie wahrscheinlich es ist, τ_hat zu beobachten, *wenn* der wahre Pre-Trend tatsächlich -5 wäre.
  - Interpretation:
    - Fall A: Test verwirft H0' (p < 0.05): Das ist ein gutes Zeichen! Es bedeutet, deine Daten sprechen *gegen* die Annahme, dass der Pre-Trend so groß ist wie dein geschätzter Effekt. Du hast Evidenz dafür gefunden, dass dein ATT_hat wahrscheinlich *nicht* nur eine Fortsetzung eines schon vorher bestehenden Trends ist. Deine Zuversicht in den ATT_hat steigt.
    - Fall B: Test verwirft H0' NICHT (p > 0.05): Das ist ein Warnsignal! Es bedeutet, deine Daten sind *vereinbar* mit einem Pre-Trend, der groß genug ist, um deinen ATT_hat zu erklären. Du kannst also nicht ausschließen, dass dein geschätzter Effekt nur auf diesem Pre-Trend beruht. Deine Ergebnisse sind mit Vorsicht zu genießen.
  - Kern: Diese Methode dreht die Fragestellung um: Statt "Können wir ausschließen, dass der Pre-Trend Null ist?", fragt sie: "Können wir ausschließen, dass der Pre-Trend *so schlimm* ist, dass er unseren Effekt erklärt?".

  Methode 2: Bounding / Identifizierte Mengen (Rambachan & Roth, 2023 [source: 547, 552, 553])

  - **Die Grundidee:** Wir wissen, dass die Parallel-Trends-Annahme nach dem Treatment (PT-ES) verletzt sein *könnte*. Wir wissen nicht *wie stark*, aber wir können die Pre-Trends nutzen, um eine *plausible Obergrenze* für diese Verletzung anzunehmen. Statt eines einzelnen Schätzwertes für den ATT erhalten wir dann einen *Bereich* möglicher ATT-Werte.

- **Schritt 1: Annahme über die maximale Verletzung (M)**
  - Wir nehmen an, dass die Verletzung der Parallel-Trends-Annahme in der Zukunft (in der Post-Periode) nicht schlimmer sein wird als das, was wir in der Vergangenheit (in der Pre-Periode) beobachtet haben.
  - Eine gängige Wahl ist: Die zukünftige Verletzung (Bias) ist betragsmäßig höchstens so groß wie der größte betragsmäßige Pre-Trend (τ_{-k}), den wir geschätzt haben [source: 551]. Nennen wir diesen größten Pre-Trend-Betrag M. Unsere Annahme ist also: |Bias| ≤ M. (Man kann M auch anders wählen, z.B. basierend auf Fachwissen oder als Vielfaches des größten Pre-Trends [source: 551]).
- **Schritt 2: Berechnung der "Identifizierten Menge"**
  - Unser normaler DiD-Schätzer (ATT_hat) schätzt eigentlich: Wahrer ATT + Bias.
  - Also gilt: Wahrer ATT = ATT_hat - Bias.
  - Da wir angenommen haben, dass der Bias zwischen -M und +M liegt (-M ≤ Bias ≤ M), können wir die Grenzen für den wahren ATT bestimmen:
    - Minimaler wahrer ATT = ATT_hat - M (wenn der Bias maximal positiv war, +M)
    - Maximaler wahrer ATT = ATT_hat + M (wenn der Bias maximal negativ war, -M)
  - Der Bereich [ATT_hat - M, ATT_hat + M] ist die **Identifizierte Menge**: Alle Werte für den wahren ATT, die mit unseren Daten UND unserer Annahme über den maximalen Bias M vereinbar sind [source: 552].
- **Schritt 3: Konfidenzintervall für die Menge**
  - Sowohl ATT_hat als auch die Pre-Trends (und damit M) sind nur Schätzungen und haben statistische Unsicherheit.
  - Die Methode von Rambachan & Roth berechnet ein **Konfidenzintervall, das diese gesamte Identifizierte Menge umschließt** und dabei die Unsicherheit *beider* Schätzungen (ATT_hat und Pre-Trends) berücksichtigt [source: 553]. Dieses Konfidenzintervall ist natürlich breiter als die Identifizierte Menge selbst.
- **Interpretation:**
  - Man schaut sich das Konfidenzintervall für die Identifizierte Menge an. Beispiel aus dem Paper: [-11.1, 5.1] [source: 557].
  - **Frage:** Ist meine ursprüngliche Schlussfolgerung (z.B. "Medicaid hatte einen negativen Effekt") robust gegenüber plausiblen Verletzungen der Parallel-Trends-Annahme?
  - **Antwort Ja (Robust):** Wenn das gesamte Konfidenzintervall z.B. immer noch nur negative Werte enthält.
  - **Antwort Nein (Nicht Robust):** Wenn das Konfidenzintervall Null oder sogar positive Werte einschließt (wie im Beispiel). Das bedeutet, die potenziellen Verzerrungen (basierend auf den Pre-Trends) sind so groß (oder die Pre-Trends so ungenau geschätzt), dass wir unsere ursprüngliche Schlussfolgerung nicht aufrechterhalten können.
  - **Fazit:** Diese Methode quantifiziert, wie sehr wir uns auf unser Ergebnis verlassen können, wenn wir zulassen, dass die Parallel-Trends-Annahme verletzt sein könnte, wobei die Pre-Trends uns eine Vorstellung über das Ausmaß der möglichen Verletzung geben.

- **Anwendung im Beispiel:** Die Autoren wenden die Rambachan & Roth Methode auf ihre Ergebnisse an und zeigen, dass selbst wenn man annimmt, die zukünftigen PT-Verletzungen seien nicht größer als die beobachteten Pre-Trends, das resultierende (Konfidenz-)Intervall für den ATT(2014) extrem breit ist ([-11.1, 5.1]) [source: 557]. Das

unterstreicht, wie wenig Information die unpräzisen Pre-Trends hier liefern [source: 556].
    - **Praktische Konsequenz:** Gehe über das "visuelle Beurteilen" ("eye-balling") von Pre-Trend-Graphen hinaus [source: 554]. Verwende formale Methoden der Sensitivitätsanalyse, die die Information (und die Unsicherheit) aus den Pre-Trends nutzen, um die Glaubwürdigkeit und Robustheit deiner Hauptergebnisse zu bewerten [source: 559].

- **Was tun bei problematischen Pre-Trends?:** Wenn Pre-Trends auf eine Verletzung der PT-Annahme hindeuten, sollte man Kovariaten einbeziehen (siehe Abschnitt 5.1.4 [source: 560]) oder alternative Methoden (wie Modelle mit einheitsspezifischen Trends, obwohl diese von Standard-DiD abweichen [source: 561, 562, 563]) in Betracht ziehen. siehe dazu mehr e.g., Mora and Reggio (2019), Wooldridge (2021, Section 7), Lee and Wooldridge (2023), and Freyaldenhoven et al. (2024).

Abschnitt 5.1.3: Schätzung und Aggregation in Event Studies [source: 566]

Dieser Abschnitt beschäftigt sich damit, wie man die ATT(t)-Werte (die Effekte für jede Periode nach Treatmentbeginn t>=g) und die Pre-Trends (τ_{-k} für t<g) praktisch schätzt und welche Fallstricke es bei der Inferenz und bei der Zusammenfassung (Aggregation) gibt.

1. **Einfache Schätzung:**
    - Da jeder ATT(t) (und jeder Pre-Trend) einem 2x2 DiD-Vergleich entspricht (Periode t vs. Periode g-1), ist die Schätzung simpel: Man nimmt einfach die Formel (5.2) [source: 509] und ersetzt die Populationserwartungswerte durch Stichprobenmittelwerte [source: 566]. Die Punkte in Figure 3 [source: 516] sind genau diese Schätzwerte [source: 566].
    - **Alternative: Regression:** Man kann *alle* ATT(t)- und Pre-Trend-Koeffizienten (β_k) auch in *einer einzigen* Regression schätzen (Formel 5.3 [source: 570]). Das ist eine TWFE-Regression (mit Einheits- und Zeit-Fixed-Effects), bei der man zusätzlich Interaktionsterme zwischen dem Treatmentgruppen-Dummy (1{G_i=g}) und Dummy-Variablen für jede Zeitperiode t (außer der Baseline-Periode t=g-1) hinzufügt [source: 568, 569]. Diese Regression liefert **numerisch identische** Schätzwerte (β_t = ATT_hat(t)) wie die "manuelle" Berechnung über Formel 5.2 [source: 571]. Sie liefert auch direkt Standardfehler (z.B. geclustert) [source: 571].
2. **Problem bei der Inferenz: Multiples Testen**
    - Eine Event Study schätzt *viele* Parameter gleichzeitig (für jede Periode einen) [source: 572]. Standard-Konfidenzintervalle (die schwarzen Balken in Figure 3 [source: 575]) sind "punktweise". Das heißt, jedes einzelne Intervall für sich hat z.B. eine 95%ige Wahrscheinlichkeit, den wahren Wert zu überdecken [source: 571].
    - Wenn man aber den *gesamten Verlauf* betrachtet oder Koeffizienten miteinander vergleicht, führt man implizit viele Tests durch [source: 573]. Die Wahrscheinlichkeit, dass *mindestens eines* der vielen Intervalle den wahren Wert verfehlt, ist dann höher als 5%.
    - **Lösung:** Man braucht Konfidenz*bänder*, die das multiple Testen berücksichtigen ("uniform confidence bands", wie das rote Band in Figure 3 [source: 575]) [source: 574]. Diese Bänder sind breiter, aber sie sind so konstruiert, dass sie den *gesamten wahren Zeitpfad* der Effekte mit z.B. 95% Wahrscheinlichkeit überdecken [source: 575]. Methoden zur Berechnung sind z.B. Bootstrap-Verfahren mit sup-t Statistiken oder Simulationen basierend auf der Kovarianzmatrix der Schätzer [source: 576, 577, 578].
3. **Problem bei der Aggregation: Einfache TWFE für Durchschnittseffekt**
    - Oft möchte man einen *einzigen* zusammenfassenden Effekt für die gesamte Nachher-Periode, z.B. den Durchschnitt aller ATT(t) für t ≥ g (genannt ATT_avg) [source: 580]. Man kann diesen einfach berechnen, indem man die einzelnen ATT_hat(t) mittelt [source: 581, 582].
    - **Der problematische Shortcut:** Ein häufiger Weg ist, eine einfache TWFE-Regression zu schätzen, die nur *einen* Dummy für "behandelt und in der Nachher-Periode" enthält (D_it = 1 wenn D_i=1 und t ≥ g, Formel 5.4 [source: 584]).

- **Warum es falsch ist:** Der Koeffizient β_OLS aus dieser einfachen Regression ist **nicht** dasselbe wie ATT_avg [source: 588]. β_OLS vergleicht implizit den Durchschnitt der Post-Periode mit dem Durchschnitt der *gesamten* Pre-Periode (also inklusive der τ_{-k}-Werte vor g-1) [source: 589, 590]. ATT_avg basiert aber auf Vergleichen nur relativ zu g-1. Die beiden Werte sind nur gleich, wenn die durchschnittlichen Pre-Trends (τ_{-k}) exakt Null sind [source: 591]. Im Beispiel unterscheiden sie sich stark: ATT_avg = -0.78 vs. β_OLS = -2.60 [source: 592].
- **Konsequenz:** Man sollte ATT_avg direkt aus den einzelnen ATT_hat(t) berechnen und nicht die einfache TWFE-Regression (5.4) für diesen Zweck verwenden.

Abschnitt 5.1.4: Kovariaten in Event Studies [source: 593]

Dieser Abschnitt erklärt, wie man die in Kapitel 4 vorgestellten Methoden zur Einbeziehung von Kovariaten (RA, IPW, DR) auf Event Studies anwendet.

1. **Grundidee:** Die "Baustein"-Perspektive macht es einfach. Da jede Schätzung eines ATT(t) (oder τ_{-k}) einem 2x2 DiD entspricht, kann man die Methoden aus Abschnitt 4.4 (RA, IPW, DR) auf **jeden einzelnen dieser 2x2-Vergleiche** anwenden [source: 593].
2. **Unterschied zu 4.4:** Anstatt der "kurzen Differenz" (Y_2 - Y_1) verwendet man nun die "lange Differenz" relativ zur Baseline-Periode g-1, also Y_t - Y_{g-1} [source: 594].
3. **Anwendung der Methoden:**
   - **RA:** Man muss für **jede Zielperiode t** ein eigenes Outcome-Modell schätzen, das E[Y_t - Y_{g-1} | X, D=0] (den bedingten Trend von g-1 zu t in der Kontrollgruppe) modelliert [source: 594].
   - **IPW:** Man braucht nur **ein einziges** Propensity-Score-Modell (P(D=1 | X)), da sich die Gruppenzugehörigkeit (D=1 vs D=0) hier nicht über die Zeit ändert [source: 594]. Dieselben IPW-Gewichte werden dann verwendet, um den gewichteten Durchschnitt der langen Differenz Y_t - Y_{g-1} für jedes t zu berechnen.
   - **DR:** Man benötigt das eine Propensity-Score-Modell und zusätzlich für jede Zielperiode t ein eigenes Outcome-Modell (wie bei RA) [source: 595].
4. **Formeln:** Die Autoren listen die angepassten Formeln für ATT_ra(t), ATT_ipw(t) und ATT_dr(t) auf, die jetzt die lange Differenz Y_t - Y_{g-1} verwenden [source: 597].
5. For completeness and ease of access, we list the RA, IPW, and DR estimands for $ATT(t)$,

$$ATT_{\text{ra}}(t) = \mathbb{E}_\omega[Y_{t,t} - Y_{t,t=g-1}|D_t = 1] - \mathbb{E}_\omega\Big[\mathbb{E}_\omega[Y_{t,t} - Y_{t,t=g-1}|X_t, D_t = 0]\Big|D_t = 1\Big],$$
$$ATT_{\text{ipw}}(t) = \mathbb{E}\Big[\big(w_{\omega,D=1}(D_t) - w_{\omega,D=0}(D_t, X_t)\big)(Y_{t,t} - Y_{t,t=g-1})\Big],$$
$$ATT_{\text{dr}}(t) = \mathbb{E}\Big[\big(w_{\omega,D=1}(D_t) - w_{\omega,D=0}(D_t, X_t)\big)\big(Y_{t,t} - Y_{t,t=g-1} - \mathbb{E}_\omega[Y_{t,t} - Y_{t,t=g-1}|X_t, D_t = 0]\big)\Big],$$

where $w_{\omega,D=1}(D_t)$ and $w_{\omega,D=0}(D_t, X_t)$ are as defined in (4.7).

6. **Abbildung 4 (Figure 4): 2xT Event Study mit Kovariaten**
   - **Inhalt:** Zeigt die Event-Study-Plots (analog zu Figure 3), aber diesmal berechnet mit den drei Methoden (RA, IPW, DR) unter Einbeziehung der im Text genannten Kovariaten [source: 604, 606].
   - **Ergebnis im Beispiel:** In dieser spezifischen Anwendung ändern die Kovariaten (mit diesen Methoden einbezogen) die Ergebnisse nur geringfügig [source: 599]. Die Autoren weisen aber darauf hin, dass die Wahl der Kovariaten wichtig ist und andere Kovariaten (wie bei Borgschulte & Vogler 2020 zitiert) durchaus zu anderen Ergebnissen führen können [source: 600, 604].
7. **Warnung:** Die Probleme der einfachen TWFE-Regression mit Kovariaten (aus Abschnitt 4.3) gelten hier weiterhin und können sich sogar verschärfen [source: 596]. Man sollte also auch hier die RA-, IPW- oder DR-Methoden bevorzugen, wenn man Kovariaten einbezieht.

Zusammenfassend: Abschnitt 5.1.3 behandelt Schätzung, Inferenzprobleme (multiples Testen -> Uniform Confidence Bands) und Aggregationsprobleme (einfache TWFE vs. Durchschnitt der ATTs) in 2xT Event Studies ohne Kovariaten. Abschnitt 5.1.4 zeigt, wie man die robusteren Methoden (RA, IPW, DR) aus Kapitel 4 anwendet, um Kovariaten korrekt in die Schätzung der einzelnen ATT(t)-Punkte einer Event Study einzubeziehen.

Detailliert Abschnitt 5.1.4:

Der Trick: Jeder Punkt ist ein eigener 2x2 DiD (mit Kovariaten)

1. **Fokus auf EINEN Punkt:** Nehmen wir an, wir wollen nur den Effekt 3 Jahre nach dem Treatment berechnen, also ATT(g+3). Vergiss für einen Moment alle anderen Jahre.
2. **Der relevante Vergleich:** Wie haben wir ATT(t) definiert? Als den Effekt in Periode t relativ zur Baseline-Periode g-1 (siehe Assumption PT-ES [source: 505] und Formel 5.2 [source: 509]). Für ATT(g+3) ist der relevante Vergleich also der zwischen den Zeitpunkten t = g+3 (die "Nachher"-Periode für diesen spezifischen Effekt) und t = g-1 (die "Vorher"-Periode).
3. **Es ist wie 2x2:** Wenn wir nur diese beiden Zeitpunkte (g-1 und g+3) betrachten, haben wir wieder genau das Setup eines **2x2 DiD**:
   - Zwei Gruppen: Treatment (D=1) und Kontrolle (D=0).
   - Zwei "Zeitpunkte": "Vorher" (g-1) und "Nachher" (g+3).
4. **Kovariaten sind nötig:** Wir gehen davon aus, dass wir Kovariaten X kontrollieren müssen (wir nehmen CPT an, nicht PT).
5. **Lösung aus Kapitel 4:** Wie schätzt man einen 2x2 DiD mit Kovariaten unter CPT? Genau dafür haben wir in Abschnitt 4.4 die Methoden RA, IPW und DR gelernt!

Die Anwendung von RA, IPW, DR auf ATT(g+3)

Wir wenden jetzt also die Logik aus 4.4 an, um unseren spezifischen Wert ATT(g+3) zu schätzen:

1. **Was ist das "Outcome", dessen Veränderung wir betrachten?** Im 2x2-Fall war es Y_2 - Y_1. Hier, für den Vergleich zwischen g+3 und g-1, ist es die **"lange Differenz" Y_{g+3} - Y_{g-1}** [source: 594]. Das ist die Variable, die wir in die RA/IPW/DR-Formeln einsetzen.
2. **RA für ATT(g+3):**
   - Wir brauchen ein Modell dafür, wie die *spezifische Veränderung* Y_{g+3} - Y_{g-1} von den Kovariaten X in der Kontrollgruppe abhängt (E[Y_{g+3} - Y_{g-1} | X, D=0]).
   - Wir schätzen dieses Modell (z.B. Regression von Y_{g+3} - Y_{g-1} auf X für D=0).
   - Wir benutzen es, um den kontrafaktischen Trend für die Behandelten vorherzusagen und wenden Formel 4.5 (angepasst mit der langen Differenz) an.
3. **IPW für ATT(g+3):**
   - Wir brauchen ein Modell für die Behandlungswahrscheinlichkeit P(D=1|X). **Wichtig:** Diese Wahrscheinlichkeit hängt *nicht* davon ab, welches Jahr t wir gerade betrachten. Sie beschreibt nur, welche Gruppe (Treatment oder Kontrolle) eine Einheit aufgrund ihrer Merkmale X wahrscheinlicher ist. Deshalb müssen wir dieses **Propensity-Score-Modell nur einmal schätzen!** [source: 594]
   - Wir berechnen die IPW-Gewichte einmal.
   - Um ATT(g+3) zu schätzen, wenden wir diese Gewichte auf die lange Differenz Y_{g+3} - Y_{g-1} an (Formel 4.8 angepasst).
4. **DR für ATT(g+3):**
   - Wir brauchen das (einmal geschätzte) Propensity-Score-Modell.
   - Wir brauchen (wie bei RA) das Outcome-Modell für die *spezifische* lange Differenz Y_{g+3} - Y_{g-1} in der Kontrollgruppe.
   - Wir kombinieren beides in der DR-Formel (Formel 4.11 angepasst).

Von EINEM Punkt zur GANZEN Event Study (Figure 4)

Die Event Study Grafik (Figure 4 [source: 604]) zeigt nicht nur ATT(g+3), sondern ATT(t) für viele verschiedene t (z.B. t=g, t=g+1, t=g+2, t=g+3,... sowie die Pre-Trends).

- Um die **ganze Grafik** zu bekommen, **wiederholen** wir einfach den oben beschriebenen Prozess für **jeden einzelnen Punkt t**:
  - Für ATT(g): Wende RA/IPW/DR auf die lange Differenz Y_g - Y_{g-1} an.
  - Für ATT(g+1): Wende RA/IPW/DR auf die lange Differenz Y_{g+1} - Y_{g-1} an.
  - Für ATT(g+2): Wende RA/IPW/DR auf die lange Differenz Y_{g+2} - Y_{g-1} an.
  - Und so weiter...

Warum braucht RA/DR separate Modelle für jedes t, aber IPW nicht?

- **IPW:** Hier geht es darum, die Gruppen vergleichbar zu machen. Die *Vergleichbarkeit* (basierend auf P(D=1|X)) ist für alle Zeitpunkte t dieselbe. Daher reicht ein Propensity-Score-Modell [source: 594].
- **RA/DR:** Hier muss der *kontrafaktische Trend* der Kontrollgruppe modelliert werden. Der Trend von g-1 nach t (E[Y_t - Y_{g-1} | X, D=0]) kann aber für jedes t anders von X abhängen [source: 594, 595]. Deshalb braucht man für jedes t, das man schätzen will, ein eigenes Outcome-Modell (bei DR zusätzlich zum einen PS-Modell) [source: 594, 595].

5.2 staggered treatment adoption GxT (erstmal nicht genauer anschauen -> nur falls carla explizit fordert)

5.3 Limitationen der TWFE regression

6 Conclusion: Step by step forward engeingeering philisophy to follow:

Step 1. Define target parameters. Adopt a potential outcomes notation that fits the study's specific setting and use it to define causal target parameters that answer the study's motivating question. Building block causal parameters usually aggregate across units using (conditional) weighted averages, and summary target parameters aggregate across the building blocks. This step fixes the study's goals in terms of causal quantities and facilitates comparisons with related studies.

Step 2. State (formally) the identification assumptions. DiD studies leverage parallel trends assumptions,but they also rely on no-anticipation and, in some cases, overlap conditions, or more. Be explicit about which form of these assumptions is required for identification in the study. Engage with the theoretical arguments necessary for them to hold and generate appropriate empirical evidence, such as pre-treatment differential trends, that can falsify or (indirectly) support their plausibility.

Step 3. Justify the estimation method. In some DiD designs, estimation is as simple as replacing population expectations with sample means. In others, such as conditional DiD designs, estimation involves choosing econometric techniques (e.g., a regression adjustment, inverse probability weighting, or doubly robust procedure) to map theoretical quantities to estimable sample quantities. Each of these strategies relies on additional modeling restrictions that should be stated clearly and respect the identification assumptions.

Step 4. Discuss sources of uncertainty. Statistical inference procedures for DiD designs stem from basic assumptions about where randomness comes from in a given design. Some researchers may adopt a sampling approach to inference, whereas others may be more comfortable witha design-based perspective.
It is important to discuss what variables of the model are being treated as fixed and what variables are considered random, as well as to use inference techniques that are compatible with the model structure and assumptions.

Step 5. Estimate. Steps 1-4 provide a specific structure for using data to estimate the causal parameters of interest.

Step 6. Conduct sensitivity analysis. A clear statement of the identification and estimation assumptions
also facilitates a clear statement of what violations of those assumptions might mean.
No study is robust to all the ways its assumptions may fail, but a good study should be robust against likely violations of plausible magnitudes.
Combine context-specific knowledge
about how the assumptions from Step 2 might be violated and by how much with the structure of the estimator from Step 3 to evaluate how much the DiD estimates vary if the key identification assumptions are not exactly true.

Step 7. Conduct heterogeneity analysis. Sometimes aggregate parameters defined in Step 1 mask important heterogeneity, in which case the forward-engineering approach simply suggests targeting sub-group parameters as well. This can include variation in parameters over time, between groups of units with different characteristics, or across different sources of treatment variation. Be clear about which types of heterogeneous effects are relevant and how they are identified and estimated.

Step 8. Keep learning. DiD is not the only or even the most plausible research design in all settings; it is just one of many causal inference techniques. If the assumptions required for a DiD analysis appear implausible ex ante or are refuted by evidence or non-robustness in practice, then explore different designs whose assumptions may be more plausible. If existing DiD methods do not provide enough guidance, then use a forward engineering approach to deduce what advances would help.

List of Acronyms

Acronym	Definition
2 ^ 2	Two-Group Two-Time-Periods DiD
2 ^ T	Two-Group T-Time-Periods DiD
ACA	Affordable Care Act
ATT	Average Treatment Effect on the Treated
CPT	Conditional Parallel Trends
CPT-GT-NYT	Conditional Parallel Trends Based on Not-Yet-Treated Groups
DiD	Difference-in-Differences
DR	Doubly Robust
ETWFE	Extended Two-Way Fixed Effects
IPW	Inverse Probability Weighted
NA	No Anticipation
NA-S	No Anticipation with Staggered Treatment Timing
OLS	Ordinary Least Squares
PT	Parallel Trends
PT-ES	Parallel Trends Event Study
PT-GT-all	Parallel Trends for Every Period and Group
PT-GT-Nev	Parallel Trends Based on Never-Treated Groups
PT-GT-NYT	Parallel Trends Based on Not-Yet-Treated Groups
RA	Regression Adjustment
SO	Strong Overlap
SO-GT	Strong Overlap With Staggered Adoption
TWFE	Two-Way Fixed Effects