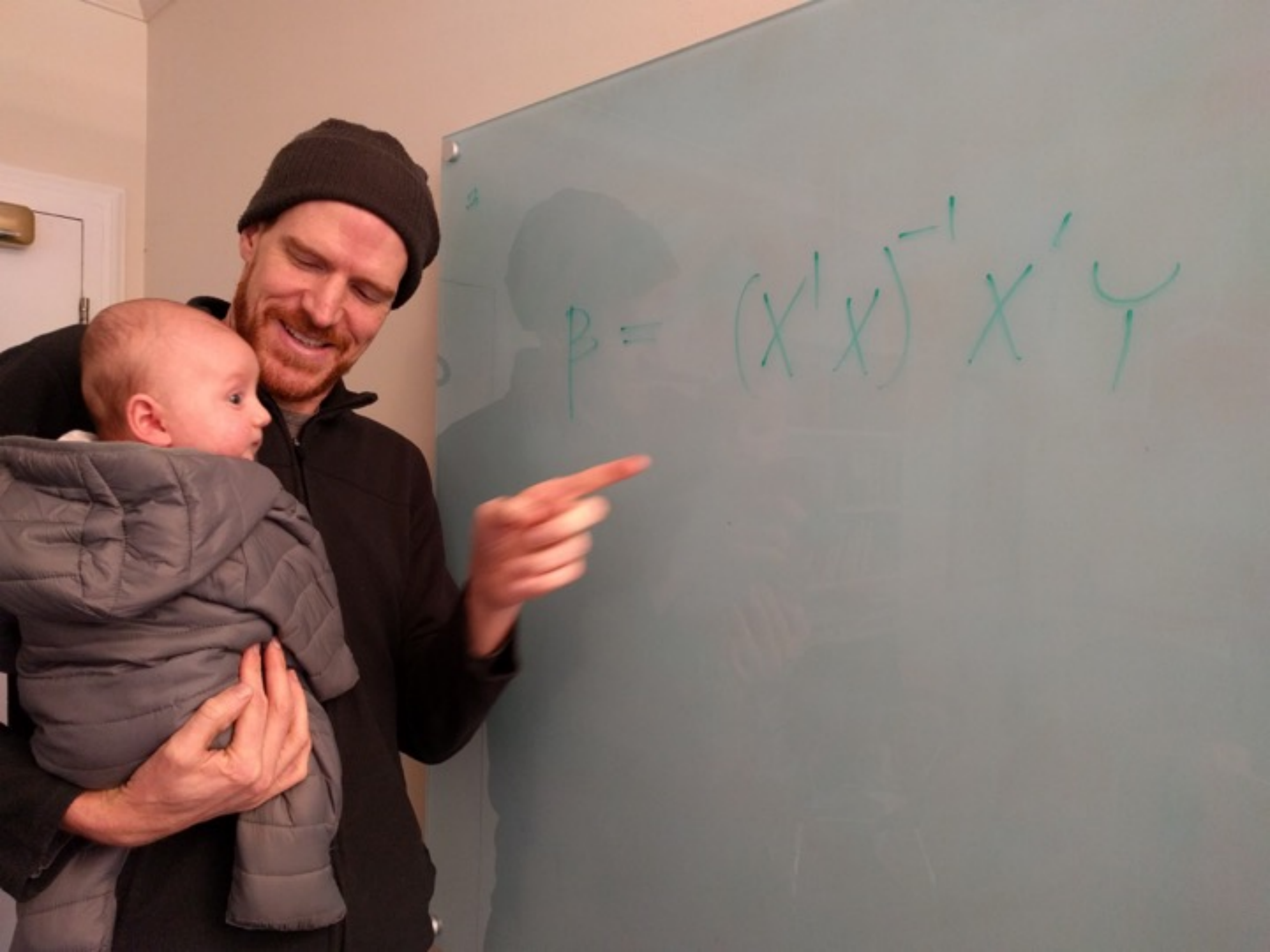
Panel data for causal inference

Intermediate Social Statistics
Week 5 (14 February 2017)
Andy Eggers

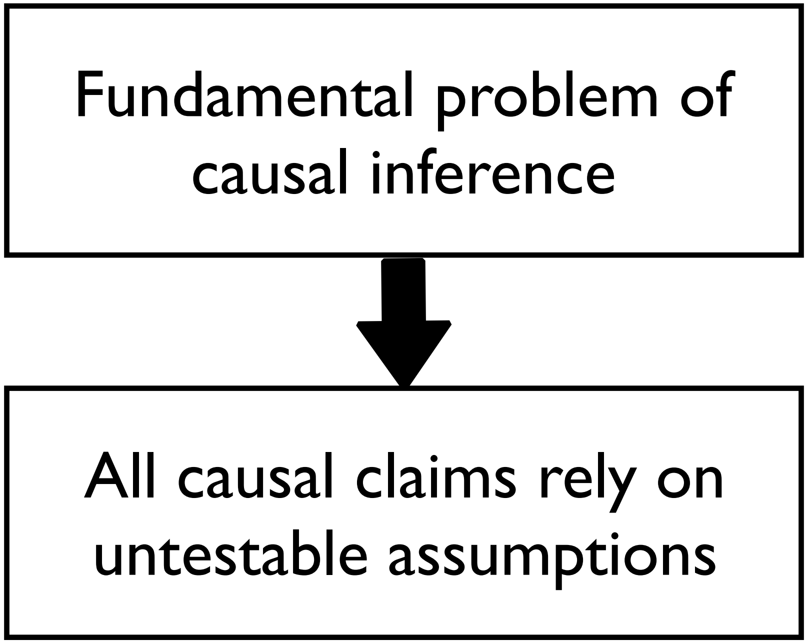

$\beta = (X'X)^{-1}X'Y$

Causal inference: the big picture

Causal inference: the big picture

Fundamental problem of causal inference

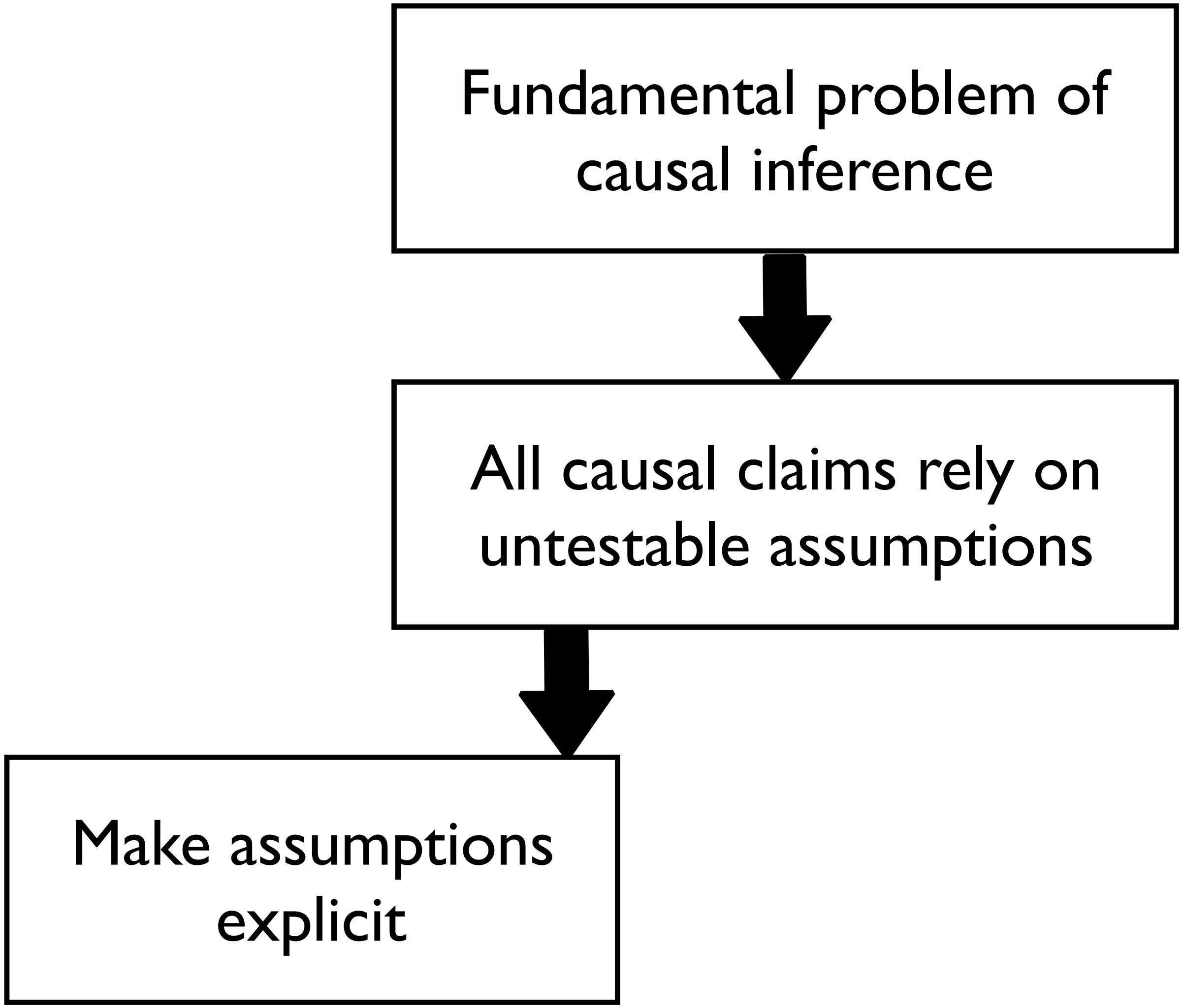
Causal inference: the big picture



Fundamental problem of causal inference

All causal claims rely on untestable assumptions

Causal inference: the big picture

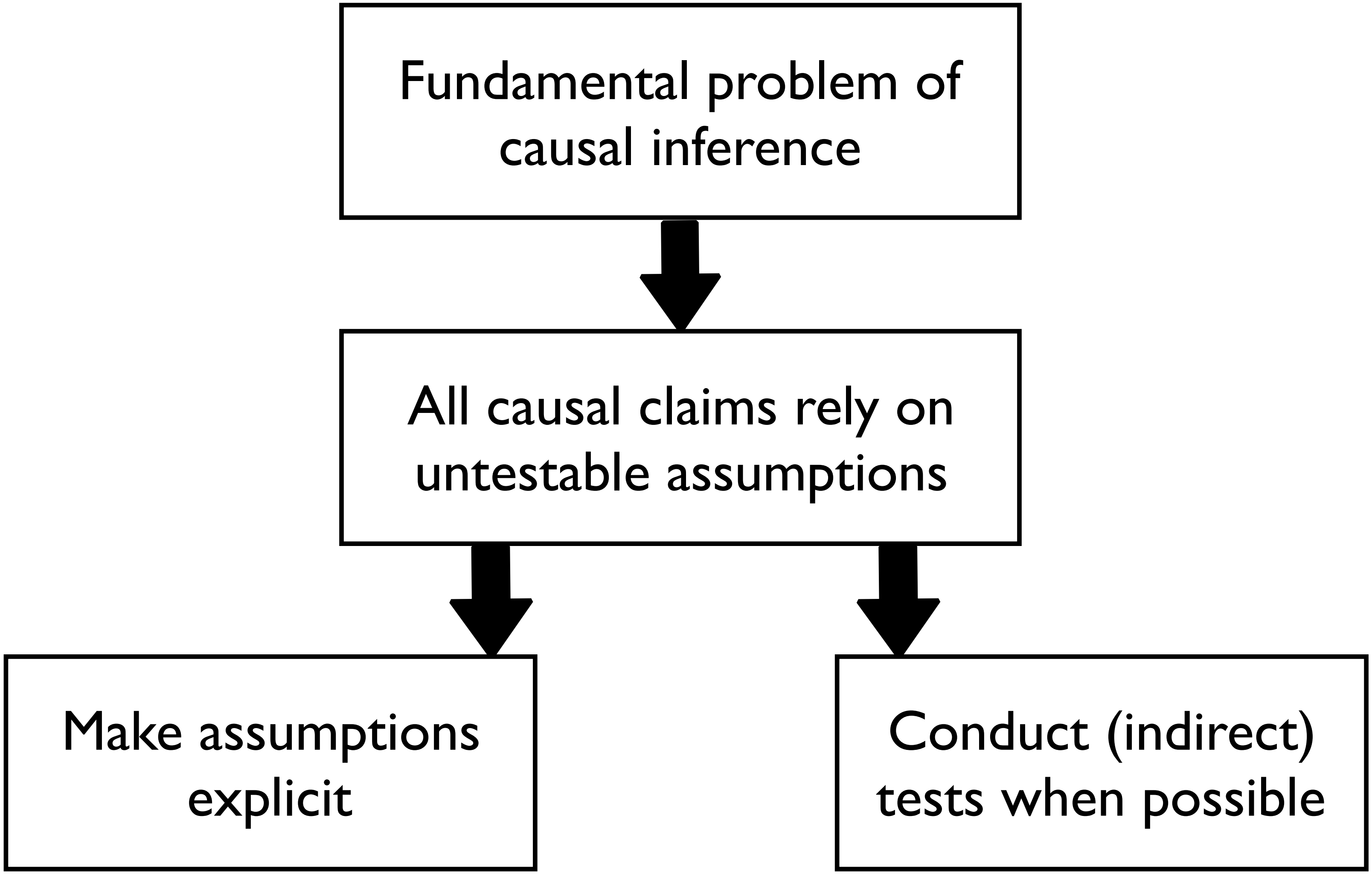


Fundamental problem of causal inference

All causal claims rely on untestable assumptions

Make assumptions explicit

Causal inference: the big picture



Fundamental problem of causal inference

All causal claims rely on untestable assumptions

Make assumptions explicit

Conduct (indirect) tests when possible

Causal inference: the big picture (2)

Causal inference: the big picture (2)

What are the assumptions? What indirect tests?

Causal inference: the big picture (2)

What are the assumptions? What indirect tests?

- Regression

Causal inference: the big picture (2)

What are the assumptions? What indirect tests?

- Regression
- Matching

Causal inference: the big picture (2)

What are the assumptions? What indirect tests?

- Regression
- Matching
- IV

Causal inference: the big picture (2)

What are the assumptions? What indirect tests?

- Regression
- Matching
- IV
- Diff-in-diff

Today's plan

- Big picture on importance of assumptions (done)
- Brief diff-in-diff review
- Generalizing in panel data
  - "First differences" approach
  - "Dummy variables" approach
  - "Fixed effects" approach ("within" regression)
- An example from the reading
- General guidelines: when is panel data useful for causal inference?

Standard diff-in-diff (two-period, binary treatment): review

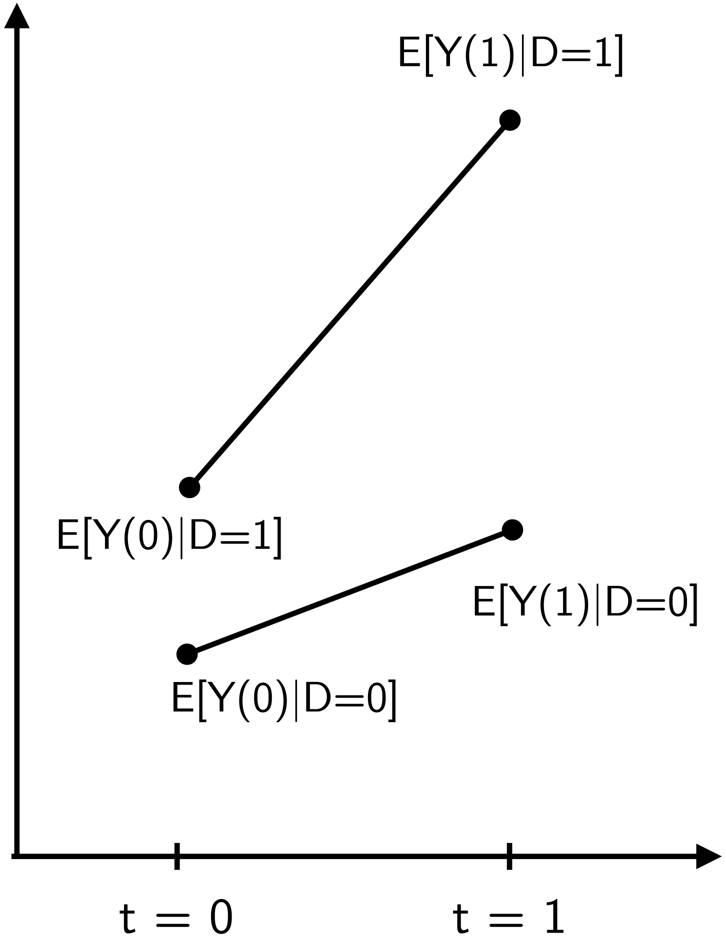
Constraints:

- Binary treatment applied to some units at a point in time
- One or more “pre-treatment” periods

Standard diff-in-diff (two-period, binary treatment): review

Constraints:

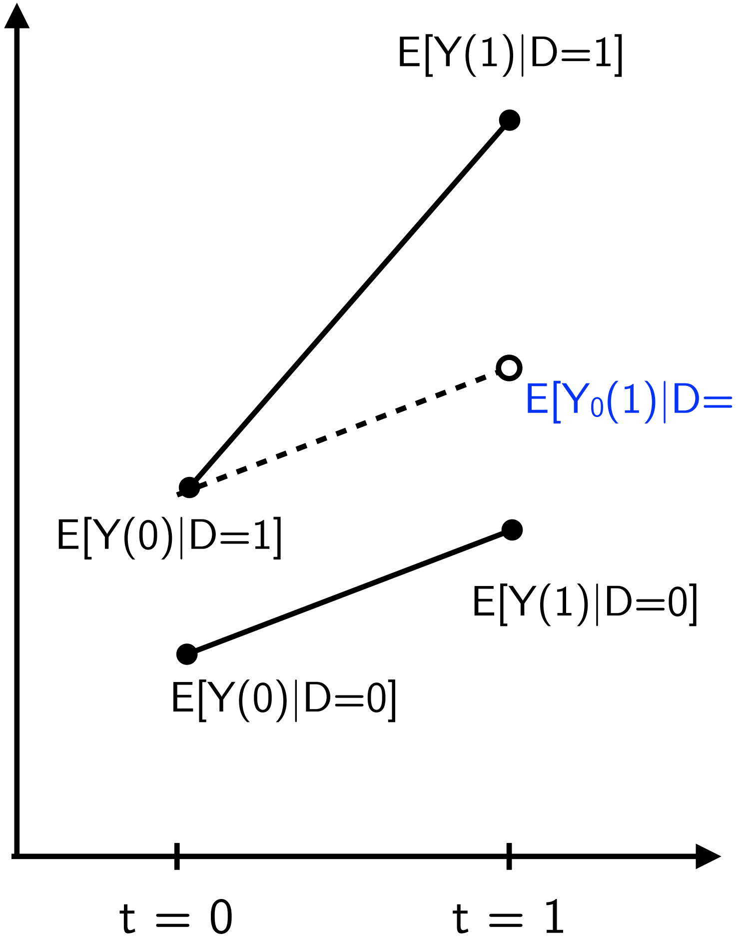
- Binary treatment applied to some units at a point in time
- One or more "pre-treatment" periods



Standard diff-in-diff (two-period, binary treatment): review

Constraints:

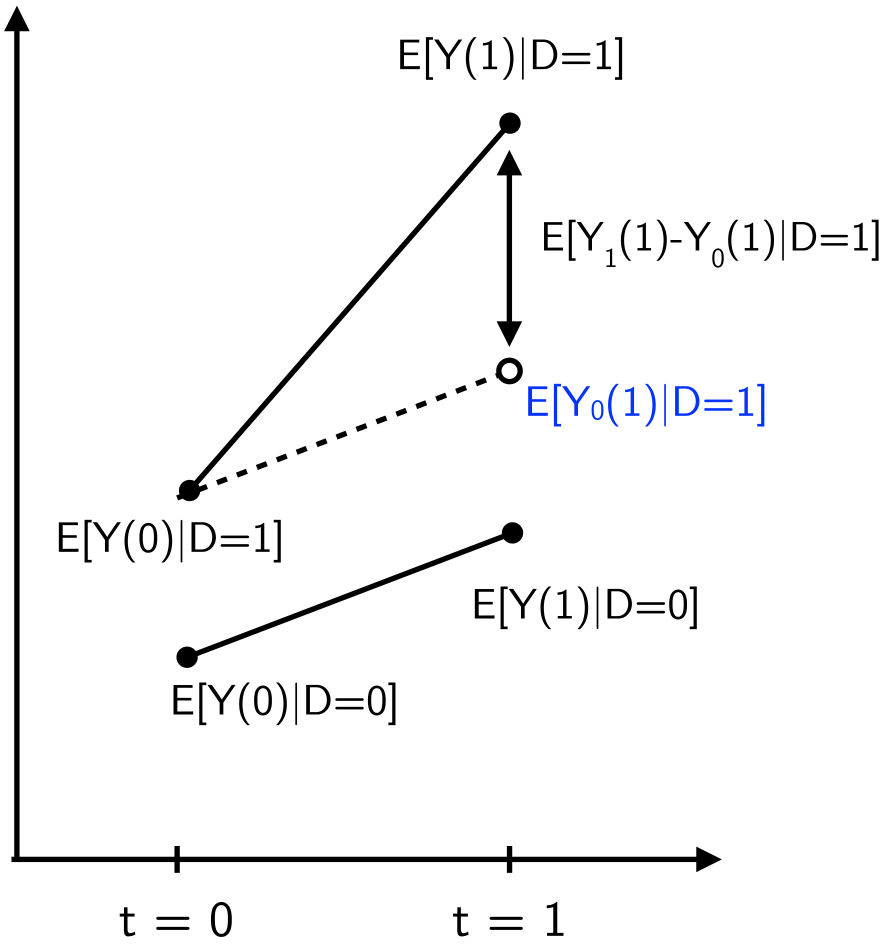
- Binary treatment applied to some units at a point in time
- One or more "pre-treatment" periods



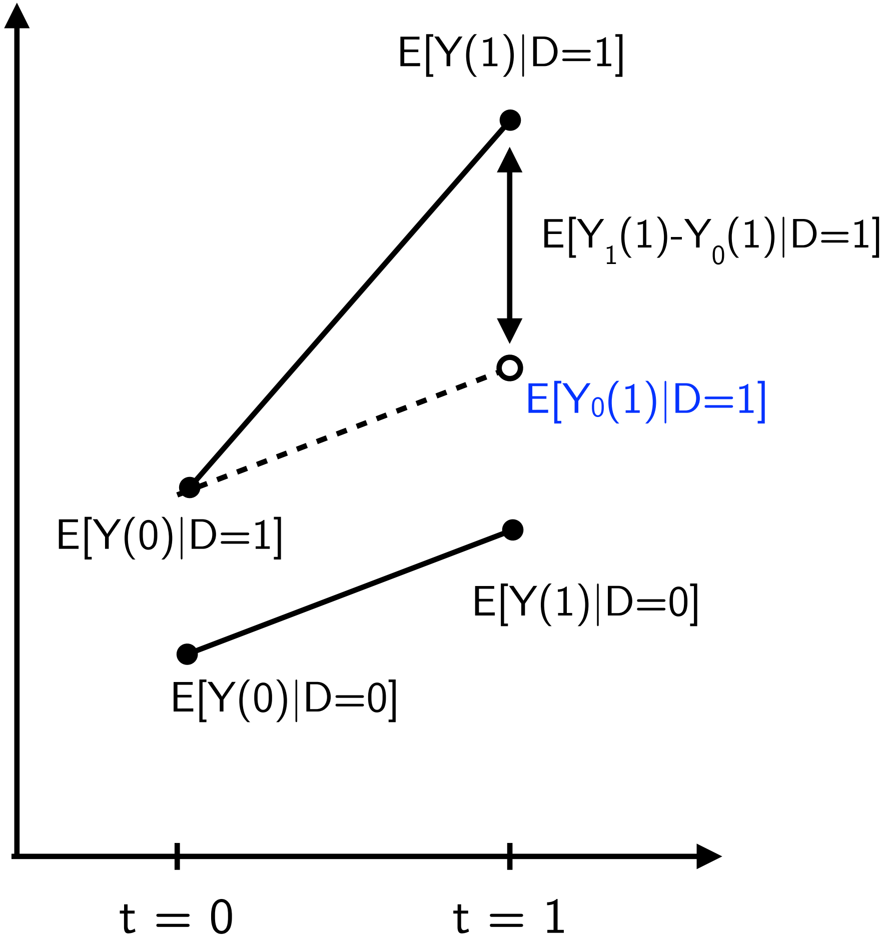
Standard diff-in-diff (two-period, binary treatment): review

Constraints:

- Binary treatment applied to some units at a point in time
- One or more "pre-treatment" periods



Standard diff-in-diff (two-period, binary treatment): review



Constraints:

- Binary treatment applied to some units at a point in time
- One or more "pre-treatment" periods

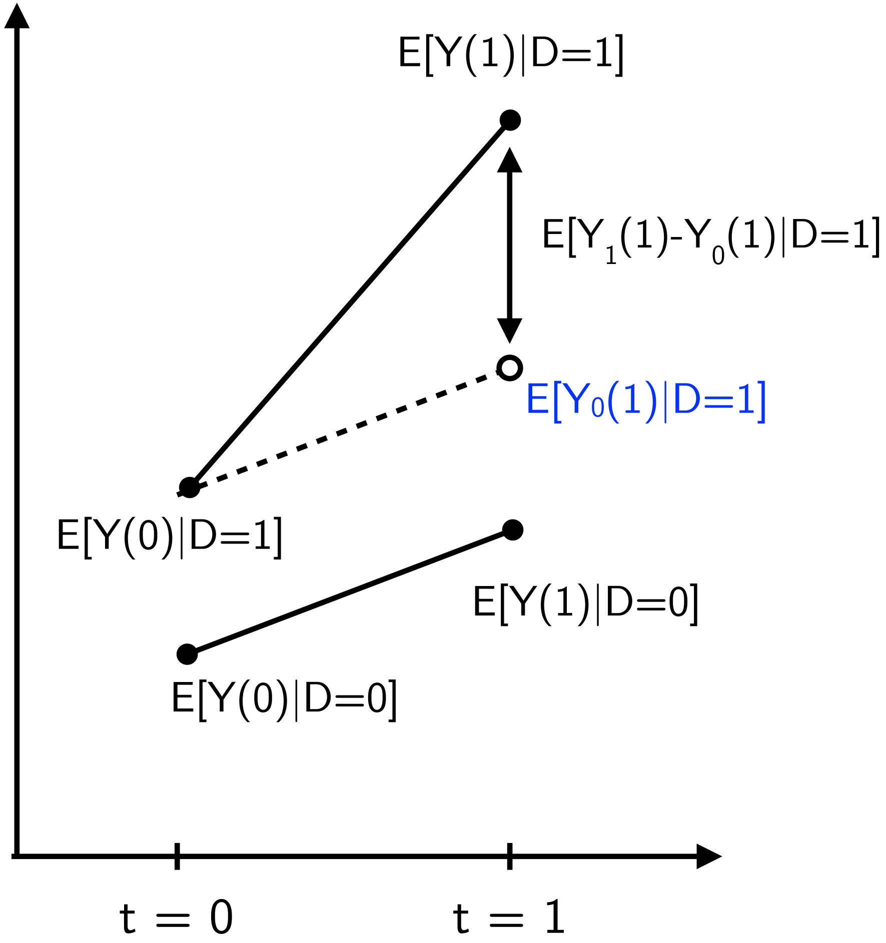
Parallel trends assumption:

$$E[Y_0(1)-Y(0)|D=1] = E[Y(1) - Y(0)|D=0]$$

Standard diff-in-diff (two-period, binary treatment): review

Constraints:

- Binary treatment applied to some units at a point in time
- One or more "pre-treatment" periods



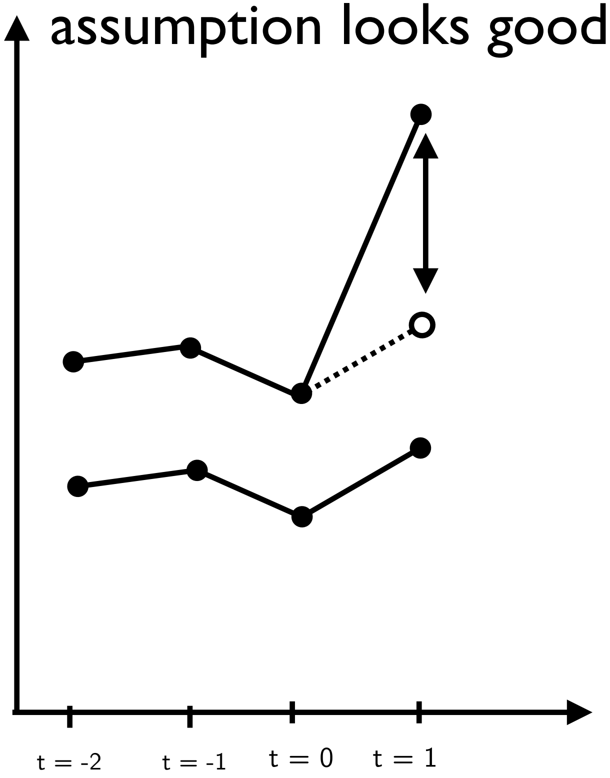
Parallel trends assumption:

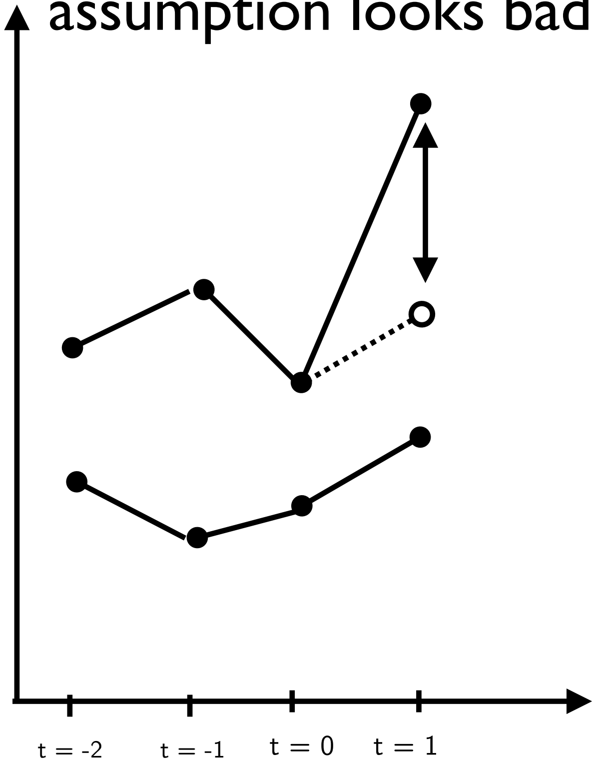
$$E[Y_0(1)-Y(0)|D=1] = E[Y(1) - Y(0)|D=0]$$

If assumption holds, ATT given by diff-in-diff, i.e.

$$E[Y(1)-Y(0)|D=1] - E[Y(1) - Y(0)|D=0]$$

Testing the parallel trends assumption


Parallel trends
assumption looks good

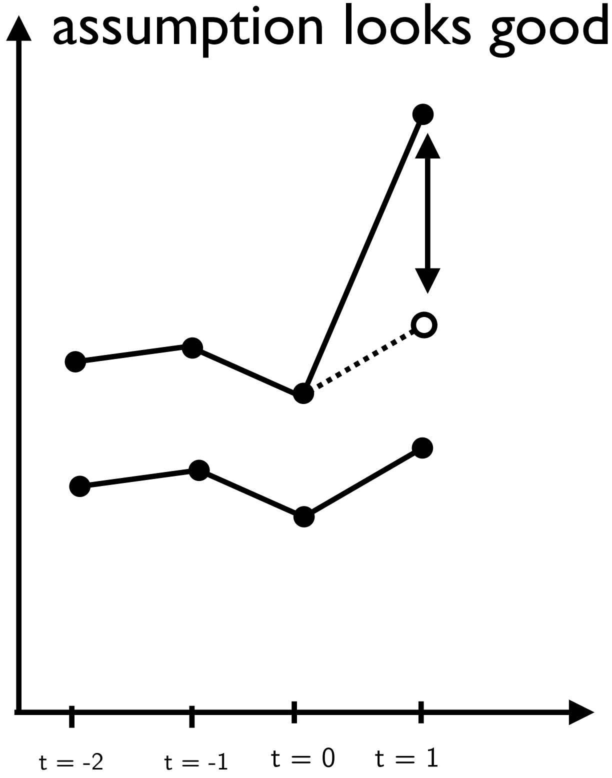

Parallel trends
assumption looks bad

Testing the parallel trends assumption

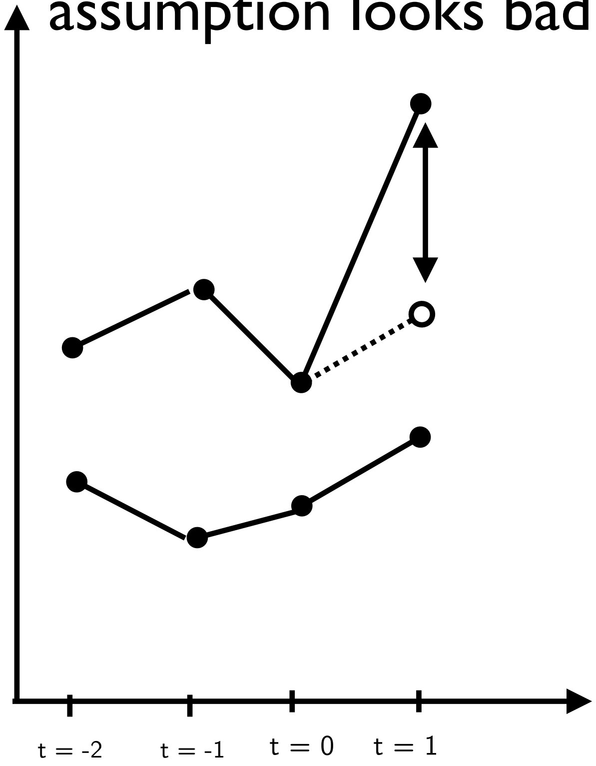
Parallel trends assumption:

$$E[Y_0(1)-Y(0)|D=1] = E[Y(1) - Y(0)|D=0]$$

Parallel trends assumption looks good



Parallel trends assumption looks bad



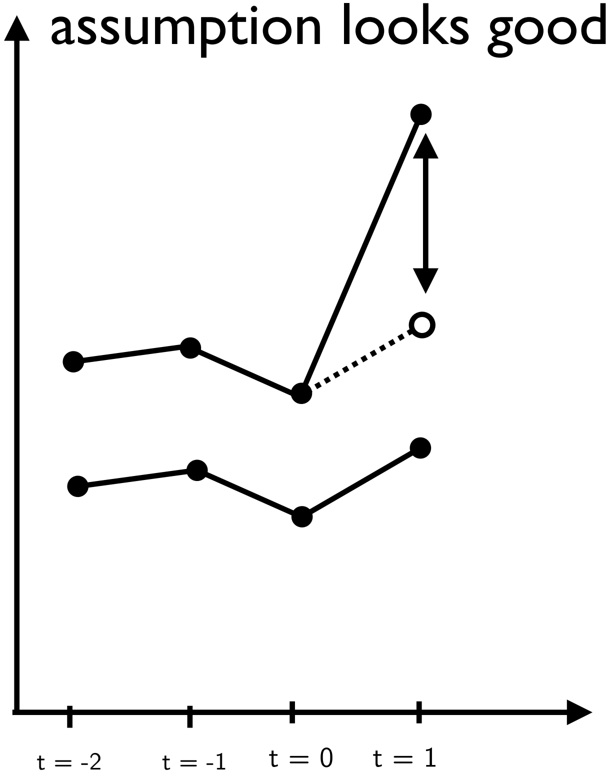
Testing the parallel trends assumption

Parallel trends assumption:

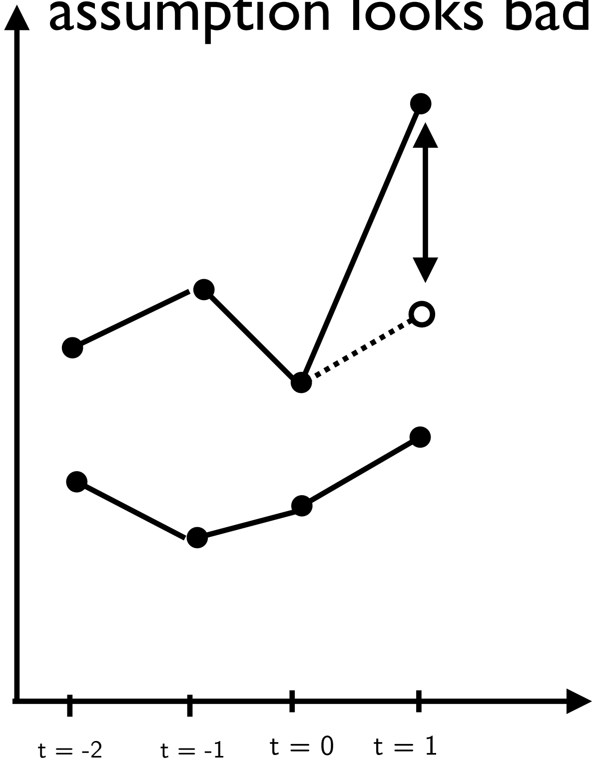
$$E[Y_0(1)-Y(0)|D=1] = E[Y(1) - Y(0)|D=0]$$

Can we test it?

Parallel trends assumption looks good



Parallel trends assumption looks bad



The beauty of the diff-in-diff: selection on unobservables

The key assumption in regression & matching can be stated as *selection on observables:*

- All covariates (factors that differ between treatment and control and affect the outcome) are observed (and properly controlled for).
- Or, *no unobserved confounding variables*.

With diff-in-diff (and today's panel methods, and IV), we can make a weaker assumption — these allow *selection on unobservables*.

Key (in diff-in-diff):

- All confounders are unchanging over time.
- Or, *no time-invariant confounding variables*.
- In other words, *unobserved okay if unchanging*.

Can we generalize?

Can we generalize?

What if:

Can we generalize?

What if:

- treatment happens to different units at different times?

Can we generalize?

What if:

- treatment happens to different units at different times?
- treatment is not binary?

Can we generalize?

What if:

- treatment happens to different units at different times?
- treatment is not binary?

We can’t do the standard diff-in-diff.

Can we generalize?

What if:

- treatment happens to different units at different times?
- treatment is not binary?

We can’t do the standard diff-in-diff.

But we can get **most** of the benefits in a more general approach, given panel data: same units (individuals, countries, classrooms) observed at several points in time.

Can we generalize?

What if:

- treatment happens to different units at different times?
- treatment is not binary?

We can't do the standard diff-in-diff.

But we can get **most** of the benefits in a more general approach, given panel data: same units (individuals, countries, classrooms) observed at several points in time.

(Note that diff-in-diff does **not** require panel data! Repeated cross-section could work.)

Three ways to get the same answer

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

- regress outcome on binary indicators for **group** (treated vs control), **time period** (pre vs post), and **interaction between them** [*group diff-in-diff formulation*]

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

- regress outcome on binary indicators for **group** (treated vs control), **time period** (pre vs post), and **interaction between them** [*group diff-in-diff formulation*]

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 Post_t + \beta_3 G_i \times Post_t$$

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

- regress outcome on binary indicators for **group** (treated vs control), **time period** (pre vs post), and **interaction between them** [*group diff-in-diff formulation*]

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 Post_t + \beta_3 G_i \times Post_t$$

- regress difference in outcome ($\Delta Y = Y_{post} - Y_{pre}$) on difference in treatment ($\Delta D = D_{post} - D_{pre}$) [*first differences formulation*]

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

- regress outcome on binary indicators for **group** (treated vs control), **time period** (pre vs post), and **interaction between them** [*group diff-in-diff formulation*]

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 Post_t + \beta_3 G_i \times Post_t$$

- regress difference in outcome ($\Delta Y = Y_{post} - Y_{pre}$) on difference in treatment ($\Delta D = D_{post} - D_{pre}$) [*first differences formulation*]

$$\Delta Y = \tau \Delta D$$

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

- regress outcome on binary indicators for **group** (treated vs control), **time period** (pre vs post), and **interaction between them** [*group diff-in-diff formulation*]

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 Post_t + \beta_3 G_i \times Post_t$$

- regress difference in outcome ($\Delta Y = Y_{post} - Y_{pre}$) on difference in treatment ($\Delta D = D_{post} - D_{pre}$) [*first differences formulation*]

$$\Delta Y = \tau \Delta D$$

- regress outcome on binary indicator for each unit (i.e. dummy for each individual, city, classroom), indicator for time period (pre vs post), and treatment [*least-squares dummy variable formulation, LSDV*]

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

- regress outcome on binary indicators for **group** (treated vs control), **time period** (pre vs post), and **interaction between them** [*group diff-in-diff formulation*]

  $Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 Post_t + \beta_3 G_i \times Post_t$

- regress difference in outcome ($\Delta Y = Y_{post} - Y_{pre}$) on difference in treatment ($\Delta D = D_{post} - D_{pre}$) [*first differences formulation*]

  $\Delta Y = \tau \Delta D$

- regress outcome on binary indicator for each unit (i.e. dummy for each individual, city, classroom), indicator for time period (pre vs post), and treatment [*least-squares dummy variable formulation, LSDV*]

  $Y_{it} = \mu_i + \beta_1 Post_t + \tau D_{it}$

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

- regress outcome on binary indicators for **group** (treated vs control), **time period** (pre vs post), and **interaction between them** [*group diff-in-diff formulation*]

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 Post_t + \beta_3 G_i \times Post_t$$

- regress difference in outcome ($\Delta Y = Y_{post} - Y_{pre}$) on difference in treatment ($\Delta D = D_{post} - D_{pre}$) [*first differences formulation*]

$$\Delta Y = \tau \Delta D$$

- regress outcome on binary indicator for each unit (i.e. dummy for each individual, city, classroom), indicator for time period (pre vs post), and treatment [*least-squares dummy variable formulation, LSDV*]

$$Y_{it} = \mu_i + \beta_l Post_t + \tau D_{it}$$

Of these, which could be used with

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

- regress outcome on binary indicators for **group** (treated vs control), **time period** (pre vs post), and **interaction between them** [*group diff-in-diff formulation*]

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 Post_t + \beta_3 G_i \times Post_t$$

- regress difference in outcome ($\Delta Y = Y_{post} - Y_{pre}$) on difference in treatment ($\Delta D = D_{post} - D_{pre}$) [*first differences formulation*]

$$\Delta Y = \tau \Delta D$$

- regress outcome on binary indicator for each unit (i.e. dummy for each individual, city, classroom), indicator for time period (pre vs post), and treatment [*least-squares dummy variable formulation, LSDV*]

$$Y_{it} = \mu_i + \beta_l Post_t + \tau D_{it}$$

Of these, which could be used with

- repeated cross-section?

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

- regress outcome on binary indicators for **group** (treated vs control), **time period** (pre vs post), and **interaction between them** [*group diff-in-diff formulation*]

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 Post_t + \beta_3 G_i \times Post_t$$

- regress difference in outcome ($\Delta Y = Y_{post} - Y_{pre}$) on difference in treatment ($\Delta D = D_{post} - D_{pre}$) [*first differences formulation*]

$$\Delta Y = \tau \Delta D$$

- regress outcome on binary indicator for each unit (i.e. dummy for each individual, city, classroom), indicator for time period (pre vs post), and treatment [*least-squares dummy variable formulation, LSDV*]

$$Y_{it} = \mu_i + \beta_l Post_t + \tau D_{it}$$

Of these, which could be used with

- repeated cross-section?
- treatment being applied at different times?

Three ways to get the same answer

Given panel data (e.g. in train bombing, Elbe flood) with two time periods, three approaches give same estimate:

- regress outcome on binary indicators for **group** (treated vs control), **time period** (pre vs post), and **interaction between them** [*group diff-in-diff formulation*]

  $$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 Post_t + \beta_3 G_i \times Post_t$$

- regress difference in outcome ($\Delta Y = Y_{post} - Y_{pre}$) on difference in treatment ($\Delta D = D_{post} - D_{pre}$) [*first differences formulation*]

  $$\Delta Y = \tau \Delta D$$

- regress outcome on binary indicator for each unit (i.e. dummy for each individual, city, classroom), indicator for time period (pre vs post), and treatment [*least-squares dummy variable formulation, LSDV*]

  $$Y_{it} = \mu_i + \beta_1 Post_t + \tau D_{it}$$

Of these, which could be used with

- repeated cross-section?
- treatment being applied at different times?
- non-binary treatment?

First differences formulation, more generally

First differences formulation, more generally

Suppose the data generating process (DGP) is

First differences formulation, more generally

Suppose the data generating process (DGP) is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

First differences formulation, more generally

Suppose the data generating process (DGP) is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise (and D_{it} and u_i are related).

First differences formulation, more generally

Suppose the data generating process (DGP) is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise (and D_{it} and u_i are related).

In words: outcome for a unit at a point in time depends on treatment, the time period, some unit-specific time-invariant characteristics u_i, and random noise.

First differences formulation, more generally

Suppose the data generating process (DGP) is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise (and D_{it} and u_i are related).

In words: outcome for a unit at a point in time depends on treatment, the time period, some unit-specific time-invariant characteristics u_i, and random noise.

In cross-section, could use regression/matching to estimate τ if u_i is observed (selection on observables).

First differences formulation, more generally

Suppose the data generating process (DGP) is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise (and D_{it} and u_i are related).

In words: outcome for a unit at a point in time depends on treatment, the time period, some unit-specific time-invariant characteristics u_i, and random noise.

In cross-section, could use regression/matching to estimate τ if u_i is observed (selection on observables).

But what if u_i is not observed (selection on unobservables)?

First differences formulation, more generally

Suppose the data generating process (DGP) is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise (and D_{it} and u_i are related).

In words: outcome for a unit at a point in time depends on treatment, the time period, some unit-specific time-invariant characteristics u_i, and random noise.

In cross-section, could use regression/matching to estimate τ if u_i is observed (selection on observables).

But what if u_i is not observed (selection on unobservables)?

With panel, take first differences:

First differences formulation, more generally

Suppose the data generating process (DGP) is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise (and D_{it} and u_i are related).

In words: outcome for a unit at a point in time depends on treatment, the time period, some unit-specific time-invariant characteristics u_i, and random noise.

In cross-section, could use regression/matching to estimate τ if u_i is observed (selection on observables).

But what if u_i is not observed (selection on unobservables)?

With panel, take first differences:

$Y_{it} - Y_{i,t-1} = \tau (D_{it}-D_{i,t-1}) + (\lambda_t-\lambda_{t-1}) + (u_i - u_i) + (\omega_{it}-\omega_{i,t-1})$

First differences formulation, more generally

Suppose the data generating process (DGP) is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it},$

where ω_{it} is random noise (and D_{it} and u_i are related).

In words: outcome for a unit at a point in time depends on treatment, the time period, some unit-specific time-invariant characteristics u_i, and random noise.

In cross-section, could use regression/matching to estimate τ if u_i is observed (selection on observables).

But what if u_i is not observed (selection on unobservables)?

With panel, take first differences:

$$Y_{it} - Y_{i,t-1} = \tau (D_{it}-D_{i,t-1}) + (\lambda_t-\lambda_{t-1}) + (u_i - u_i) + (\omega_{it}-\omega_{i,t-1})$$

$$\Delta Y_{it} = \tau \Delta D_{it} + \tilde{\lambda}_t + \tilde{\omega}_{it}$$

First differences formulation, more generally

Suppose the data generating process (DGP) is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it},$

where ω_{it} is random noise (and D_{it} and u_i are related).

In words: outcome for a unit at a point in time depends on treatment, the time period, some unit-specific time-invariant characteristics u_i, and random noise.

In cross-section, could use regression/matching to estimate τ if u_i is observed (selection on observables).

But what if u_i is not observed (selection on unobservables)?

With panel, take first differences:

$Y_{it} - Y_{i,t-1} = \tau (D_{it}-D_{i,t-1}) + (\lambda_t-\lambda_{t-1}) + (u_i - u_i) + (\omega_{it}-\omega_{i,t-1})$

$\Delta Y_{it} = \tau \Delta D_{it} + \tilde{\lambda}_t + \tilde{\omega}_{it}$

You can estimate τ even though u_i is unobserved!

LSDV formulation, more generally

LSDV formulation, more generally

Suppose again that DGP is

LSDV formulation, more generally

Suppose again that DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

LSDV formulation, more generally

Suppose again that DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise (and D_{it} and u_i are related), and suppose that u_i is unobserved.

LSDV formulation, more generally

Suppose again that DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise (and D_{it} and u_i are related), and suppose that u_i is unobserved.

Idea: include a dummy variable for every unit (individual, municipality, school) and time period.

LSDV formulation, more generally

Suppose again that DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise (and D_{it} and u_i are related), and suppose that u_i is unobserved.

Idea: include a dummy variable for every unit (individual, municipality, school) and time period.

Idea, continued: There may be many important unobserved covariates that affect outcome and treatment. Any time-invariant covariates are controlled for by the unit-specific dummy variable. Any common time trends are controlled for by the time period dummy variables.

LSDV and “fixed effects”

LSDV and “fixed effects”

Instead of LSDV, more common to specify “fixed effects” (`fe` option in Stata `xtreg`). This produces the same result by “demeaning” the variables for each unit rather than including a dummy variable for each unit.

LSDV and “fixed effects”

Instead of LSDV, more common to specify “fixed effects” (`fe` option in Stata `xtreg`). This produces the same result by “demeaning” the variables for each unit rather than including a dummy variable for each unit.

Why does this work?

LSDV and “fixed effects”

Instead of LSDV, more common to specify “fixed effects” (`fe` option in Stata `xtreg`). This produces the same result by “demeaning” the variables for each unit rather than including a dummy variable for each unit.

Why does this work?

- Logic of partial regression (see MHE on this)

LSDV and "fixed effects"

Instead of LSDV, more common to specify "fixed effects" (`fe` option in Stata `xtreg`). This produces the same result by "demeaning" the variables for each unit rather than including a dummy variable for each unit.

Why does this work?

- Logic of partial regression (see MHE on this)
- Link to first differences

LSDV and “fixed effects”

Instead of LSDV, more common to specify “fixed effects” (`fe` option in Stata `xtreg`). This produces the same result by “demeaning” the variables for each unit rather than including a dummy variable for each unit.

Why does this work?

- Logic of partial regression (see MHE on this)
- Link to first differences

Depending on the problem, may help to think about fixed effects regression as LSDV or first differences.

LSDV and “fixed effects”

Instead of LSDV, more common to specify “fixed effects” (`fe` option in Stata `xtreg`). This produces the same result by “demeaning” the variables for each unit rather than including a dummy variable for each unit.

Why does this work?

- Logic of partial regression (see MHE on this)
- Link to first differences

Depending on the problem, may help to think about fixed effects regression as LSDV or first differences.

Also referred to as “within” (vs “between”) regression.

The key assumption and when it fails

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

What about if the DGP is

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

What about if the DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + v_{it} + \omega_{it}$

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

What about if the DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + v_{it} + \omega_{it}$

and both u_i and v_{it} unobserved?

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

What about if the DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + v_{it} + \omega_{it}$

and both u_i and v_{it} unobserved?

u_i is a **time-invariant confounder**.

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

What about if the DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + v_{it} + \omega_{it}$

and both u_i and v_{it} unobserved?

u_i is a **time-invariant confounder**.

v_{it} is a **time-variant confounder**.

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

What about if the DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + v_{it} + \omega_{it}$

and both u_i and v_{it} unobserved?

u_i is a **time-invariant confounder**.

v_{it} is a **time-variant confounder**.

First-differences, LSDV, fixed effects all handle u_i but not v_{it}.

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

What about if the DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + v_{it} + \omega_{it}$

and both u_i and v_{it} unobserved?

u_i is a **time-invariant confounder**.

v_{it} is a **time-variant confounder**.

First-differences, LSDV, fixed effects all handle u_i but not v_{it}.

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

What about if the DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + v_{it} + \omega_{it}$

and both u_i and v_{it} unobserved?

u_i is a **time-invariant confounder**.

v_{it} is a **time-variant confounder**.

First-differences, LSDV, fixed effects all handle u_i but not v_{it}.

So in panel studies (generalized DiD with two-way fixed effects)

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

What about if the DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + v_{it} + \omega_{it}$

and both u_i and v_{it} unobserved?

u_i is a **time-invariant confounder**.

v_{it} is a **time-variant confounder**.

First-differences, LSDV, fixed effects all handle u_i but not v_{it}.

So in panel studies (generalized DiD with two-way fixed effects)

- key assumption: no time-variant unit-specific confounders

The key assumption and when it fails

We showed how to estimate the ATE when DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + \omega_{it}$,

where ω_{it} is random noise and u_i unobserved.

What about if the DGP is

$Y_{it} = \tau D_{it} + \lambda_t + u_i + v_{it} + \omega_{it}$

and both u_i and v_{it} unobserved?

u_i is a **time-invariant confounder**.

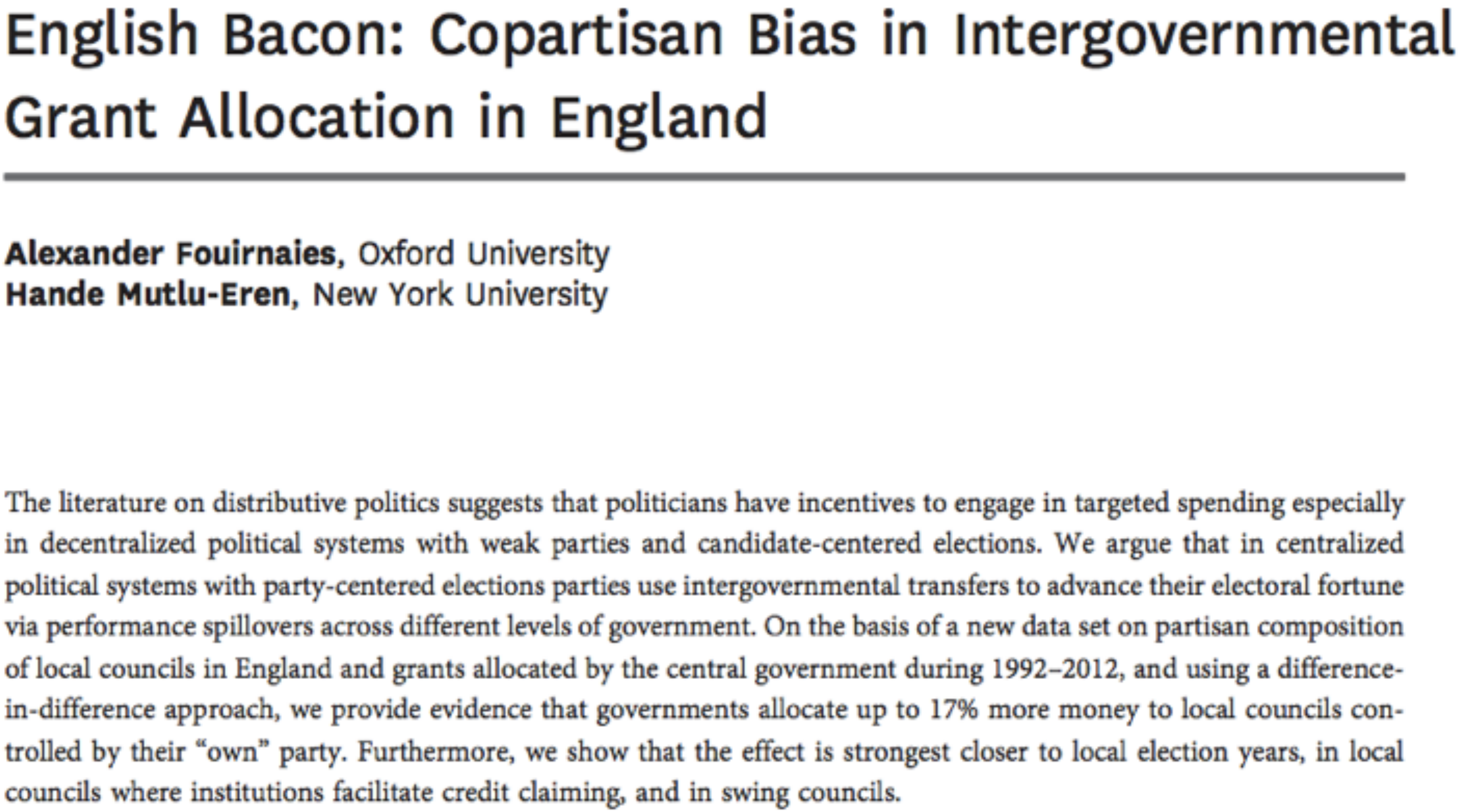
v_{it} is a **time-variant confounder**.

First-differences, LSDV, fixed effects all handle u_i but not v_{it}.

So in panel studies (generalized DiD with two-way fixed effects)

- key assumption: no time-variant unit-specific confounders
- your job: think about what could violate this assumption

Example

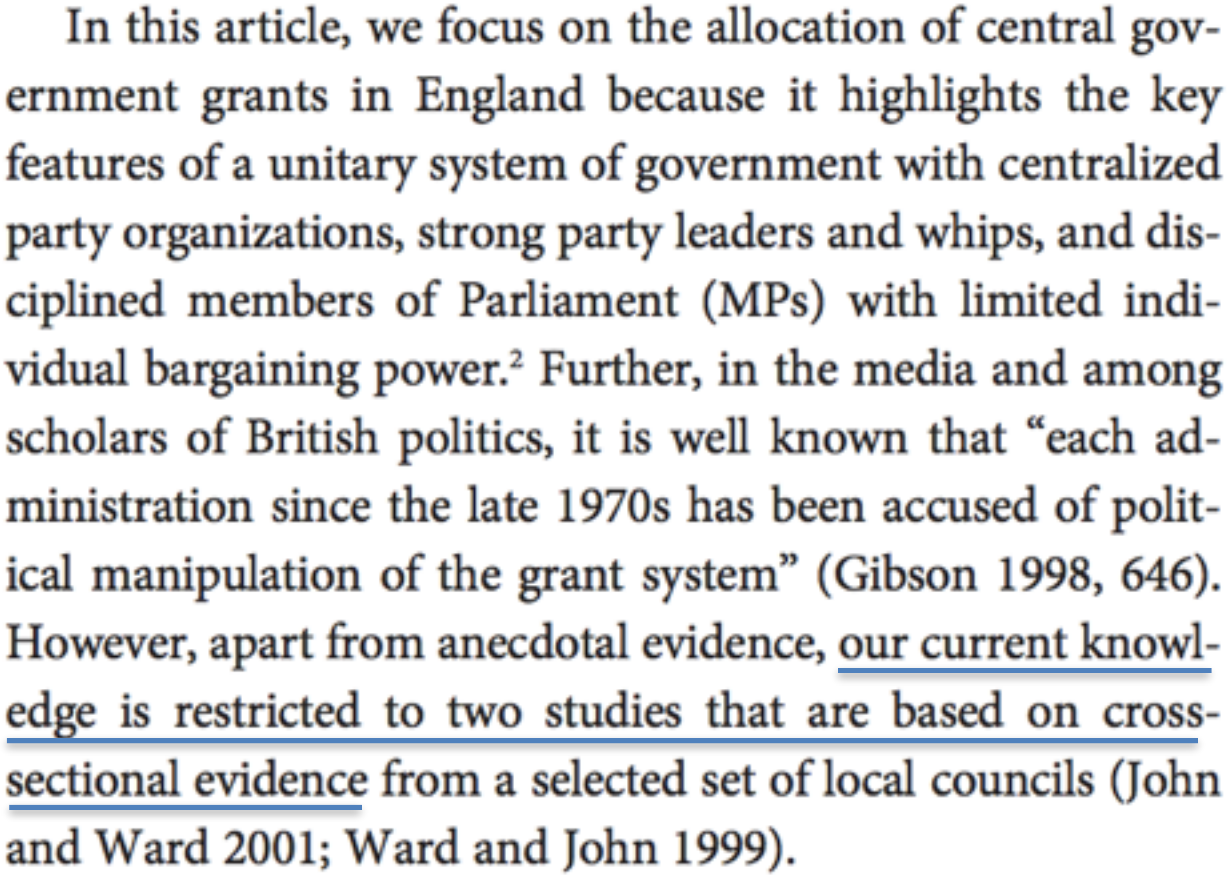


English Bacon: Copartisan Bias in Intergovernmental Grant Allocation in England

Alexander Fouirnaies, Oxford University
Hande Mutlu-Eren, New York University

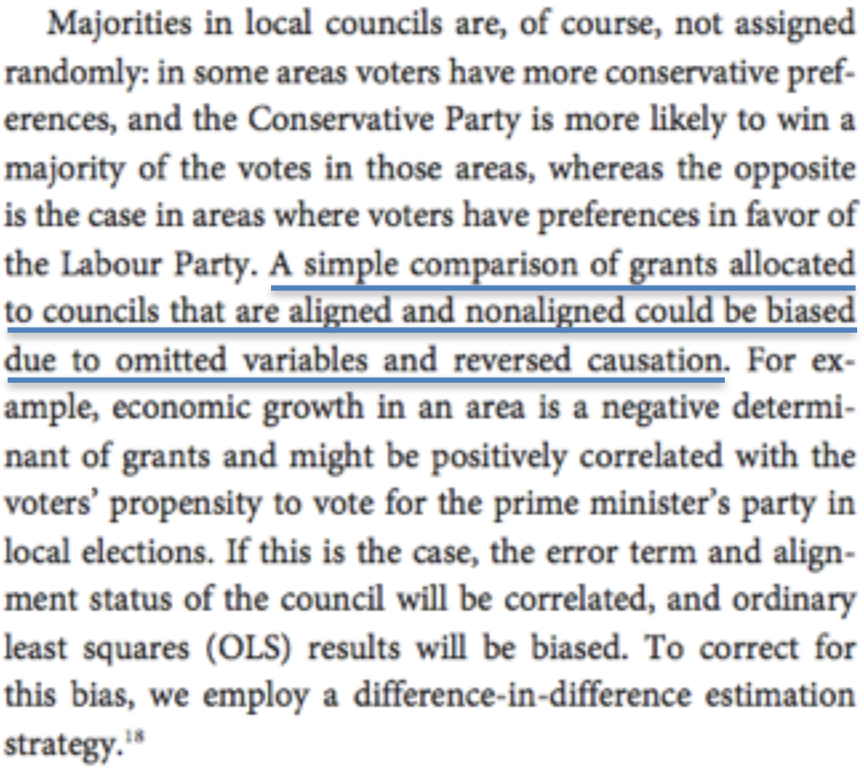
The literature on distributive politics suggests that politicians have incentives to engage in targeted spending especially in decentralized political systems with weak parties and candidate-centered elections. We argue that in centralized political systems with party-centered elections parties use intergovernmental transfers to advance their electoral fortune via performance spillovers across different levels of government. On the basis of a new data set on partisan composition of local councils in England and grants allocated by the central government during 1992–2012, and using a difference-in-difference approach, we provide evidence that governments allocate up to 17% more money to local councils controlled by their "own" party. Furthermore, we show that the effect is strongest closer to local election years, in local councils where institutions facilitate credit claiming, and in swing councils.

Example: motive

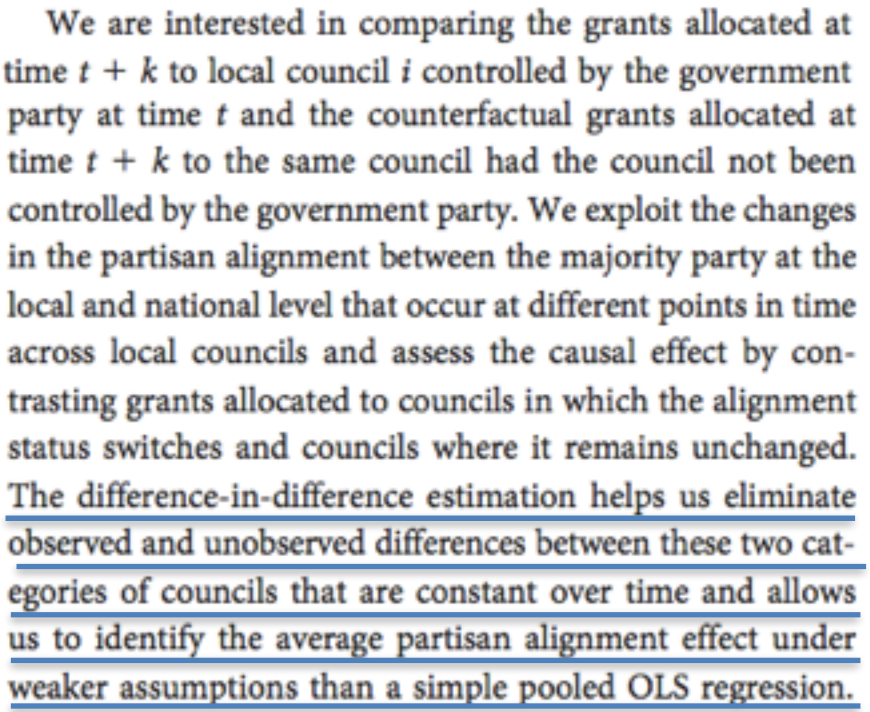


In this article, we focus on the allocation of central government grants in England because it highlights the key features of a unitary system of government with centralized party organizations, strong party leaders and whips, and disciplined members of Parliament (MPs) with limited individual bargaining power.[2] Further, in the media and among scholars of British politics, it is well known that "each administration since the late 1970s has been accused of political manipulation of the grant system" (Gibson 1998, 646). However, apart from anecdotal evidence, our current knowledge is restricted to two studies that are based on cross-sectional evidence from a selected set of local councils (John and Ward 2001; Ward and John 1999).

Example: design and assumptions

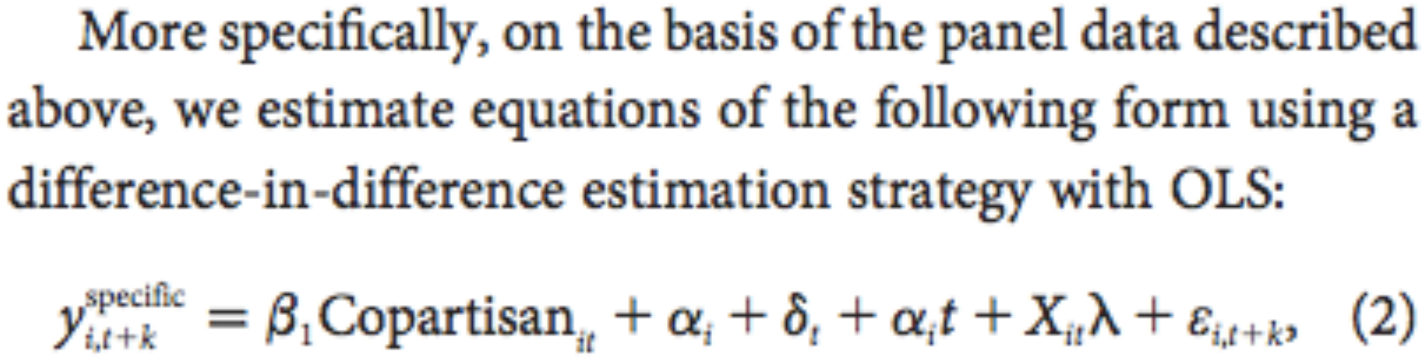


Majorities in local councils are, of course, not assigned randomly: in some areas voters have more conservative preferences, and the Conservative Party is more likely to win a majority of the votes in those areas, whereas the opposite is the case in areas where voters have preferences in favor of the Labour Party. A simple comparison of grants allocated to councils that are aligned and nonaligned could be biased due to omitted variables and reversed causation. For example, economic growth in an area is a negative determinant of grants and might be positively correlated with the voters' propensity to vote for the prime minister's party in local elections. If this is the case, the error term and alignment status of the council will be correlated, and ordinary least squares (OLS) results will be biased. To correct for this bias, we employ a difference-in-difference estimation strategy.[18]



We are interested in comparing the grants allocated at time $t + k$ to local council i controlled by the government party at time t and the counterfactual grants allocated at time $t + k$ to the same council had the council not been controlled by the government party. We exploit the changes in the partisan alignment between the majority party at the local and national level that occur at different points in time across local councils and assess the causal effect by contrasting grants allocated to councils in which the alignment status switches and councils where it remains unchanged. The difference-in-difference estimation helps us eliminate observed and unobserved differences between these two categories of councils that are constant over time and allows us to identify the average partisan alignment effect under weaker assumptions than a simple pooled OLS regression.

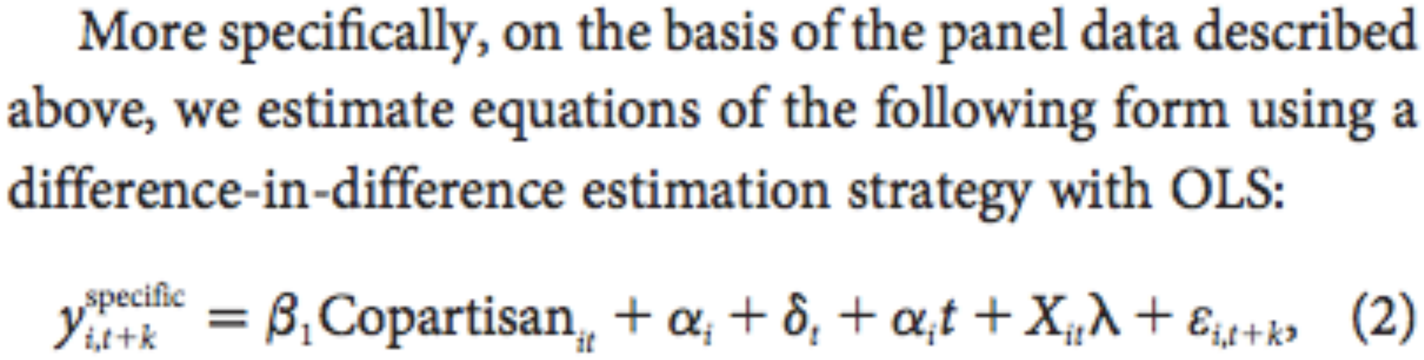
Example: specification



More specifically, on the basis of the panel data described above, we estimate equations of the following form using a difference-in-difference estimation strategy with OLS:

$$y^{\text{specific}}_{i,t+k} = \beta_1 \text{Copartisan}_{it} + \alpha_i + \delta_t + \alpha_i t + X_{it}\lambda + \varepsilon_{i,t+k}, \quad (2)$$

Example: specification

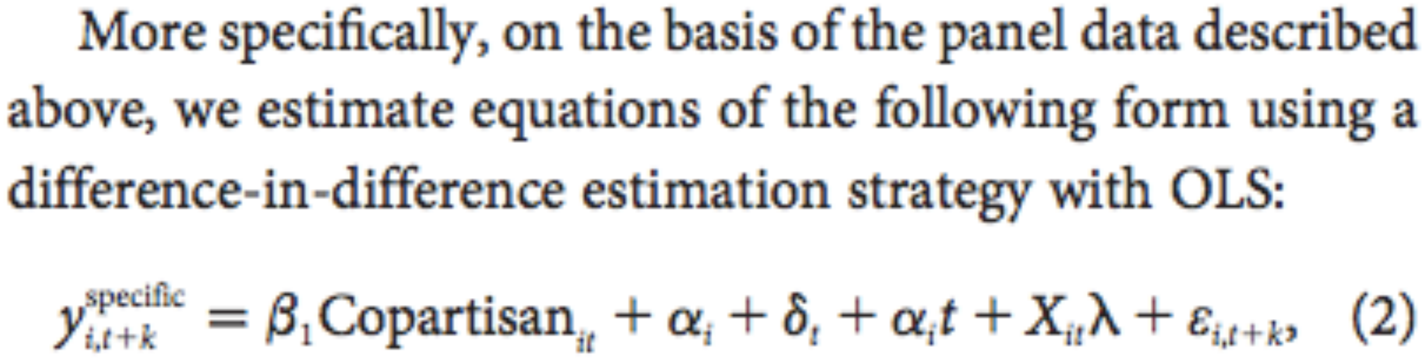


More specifically, on the basis of the panel data described above, we estimate equations of the following form using a difference-in-difference estimation strategy with OLS:

$$y_{i,t+k}^{\text{specific}} = \beta_1 \text{Copartisan}_{it} + \alpha_i + \delta_t + \alpha_i t + X_{it}\lambda + \varepsilon_{i,t+k}, \quad (2)$$

Note a few special features:

Example: specification



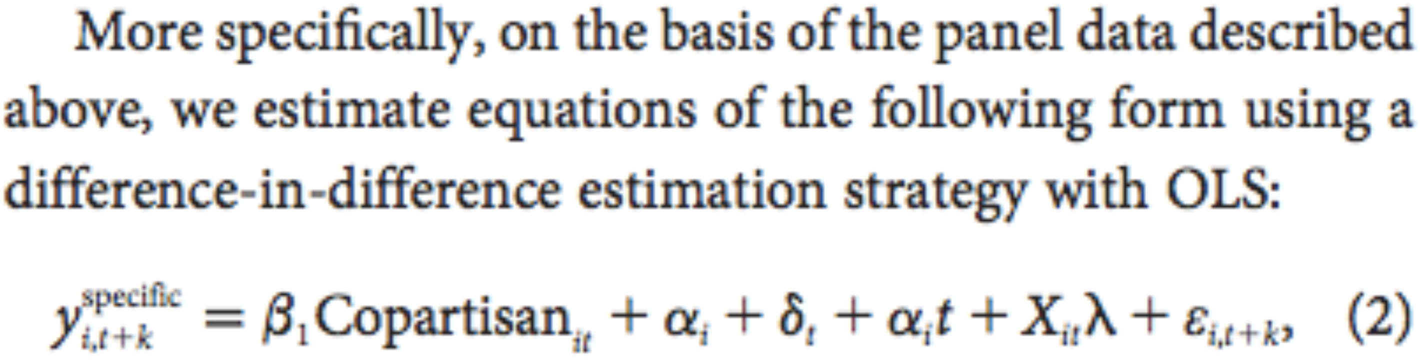
More specifically, on the basis of the panel data described above, we estimate equations of the following form using a difference-in-difference estimation strategy with OLS:

$$y_{i,t+k}^{\text{specific}} = \beta_1 \text{Copartisan}_{it} + \alpha_i + \delta_t + \alpha_i t + X_{it}\lambda + \varepsilon_{i,t+k}, \qquad (2)$$

Note a few special features:

- unit-specific time trends ($\alpha_i t$) — how does this relax the parallel trends assumption?

Example: specification



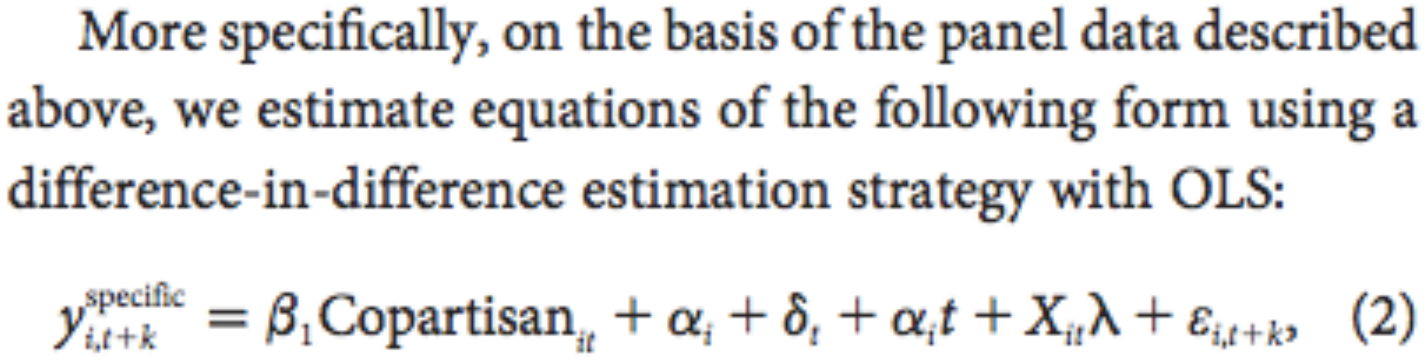
More specifically, on the basis of the panel data described above, we estimate equations of the following form using a difference-in-difference estimation strategy with OLS:

$$y_{i,t+k}^{\text{specific}} = \beta_1 \text{Copartisan}_{it} + \alpha_i + \delta_t + \alpha_i t + X_{it}\lambda + \varepsilon_{i,t+k}, \quad (2)$$

Note a few special features:

- unit-specific time trends ($\alpha_i t$) — how does this relax the parallel trends assumption?
- let k be -3, -2, . . . , 6 — what does this test? What should we expect to find if partisan alignment affects spending?

Example: specification



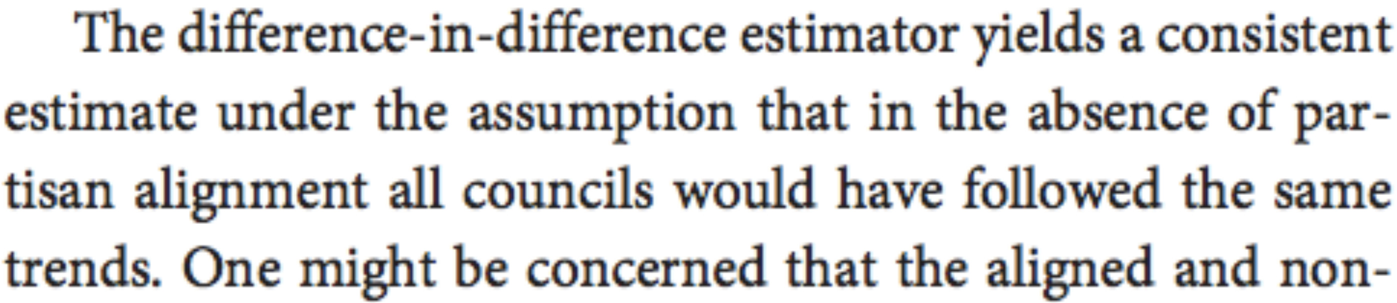
More specifically, on the basis of the panel data described above, we estimate equations of the following form using a difference-in-difference estimation strategy with OLS:

$$y^{\text{specific}}_{i,t+k} = \beta_1 \text{Copartisan}_{it} + \alpha_i + \delta_t + \alpha_i t + X_{it}\lambda + \varepsilon_{i,t+k}, \quad (2)$$

Note a few special features:

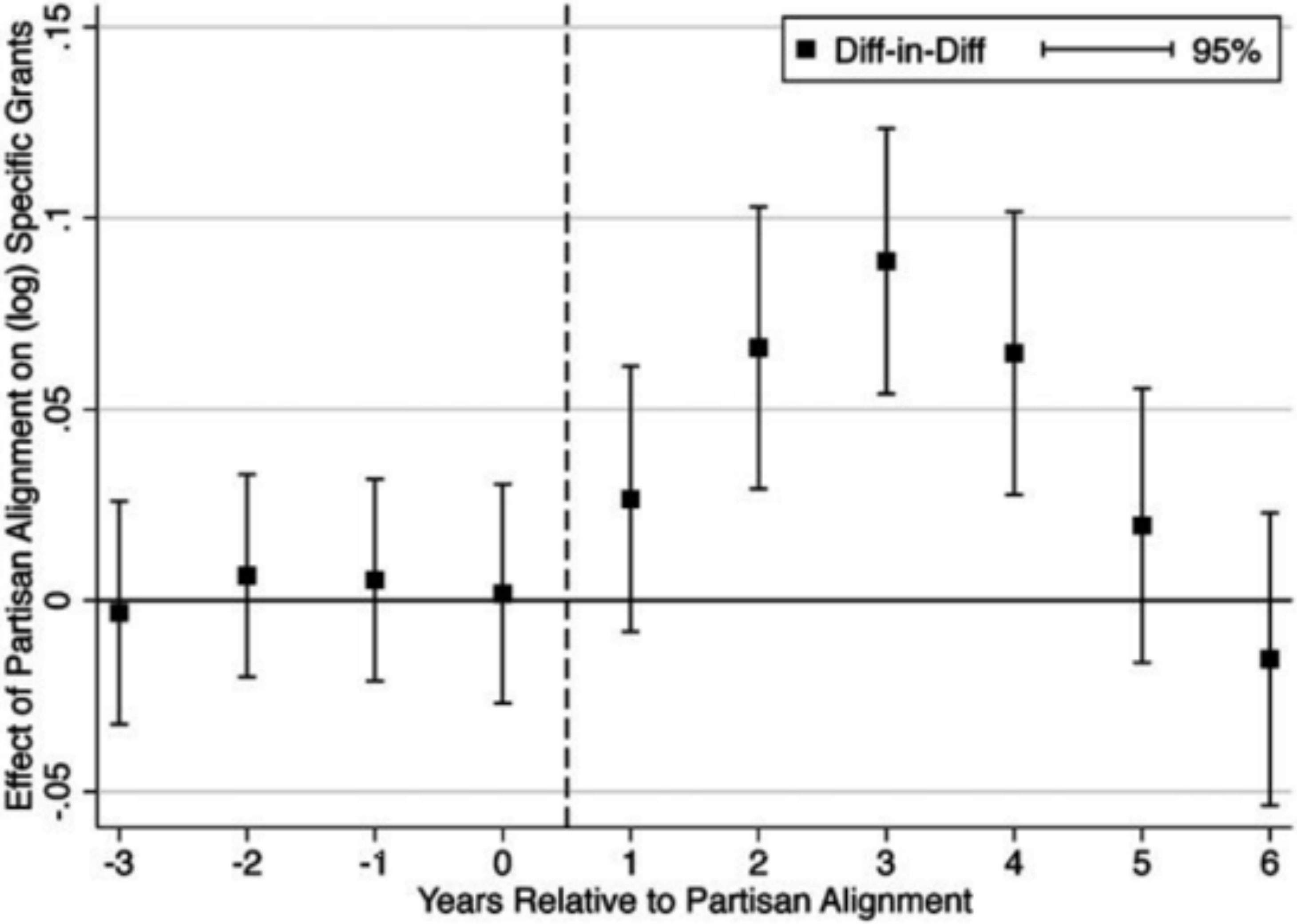
- unit-specific time trends ($\alpha_i t$) — how does this relax the parallel trends assumption?
- let k be -3, -2, . . . , 6 — what does this test? What should we expect to find if partisan alignment affects spending?

Statement of assumptions under which this gives the right answer (and possible violations of those assumptions):

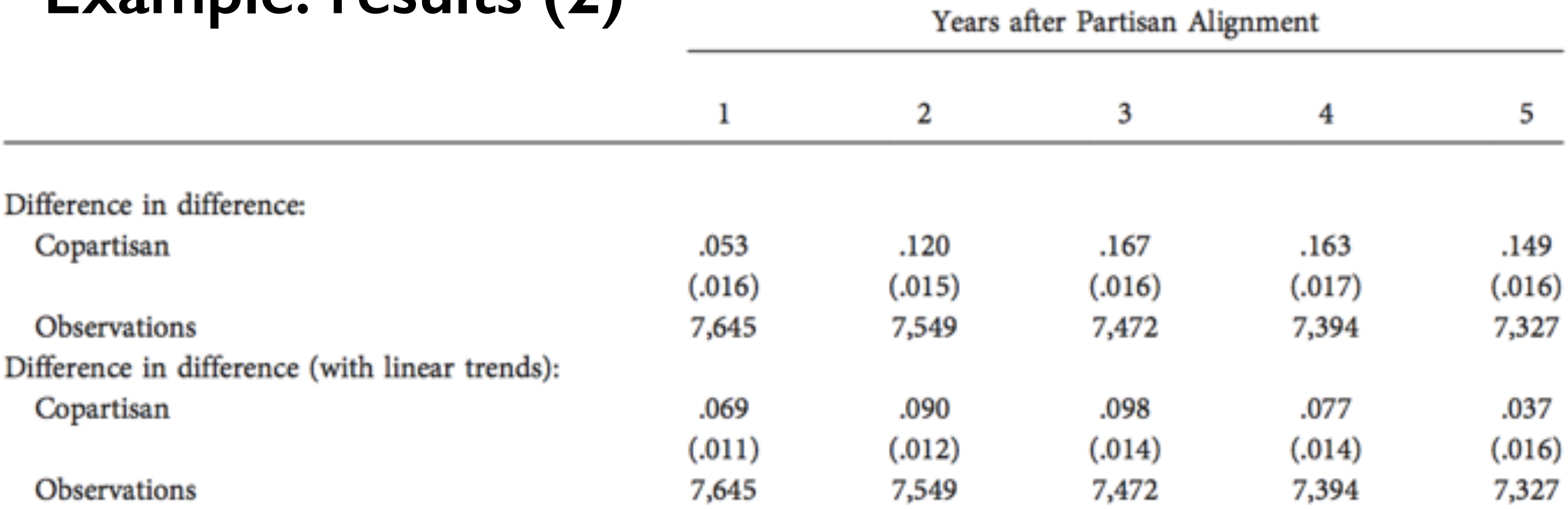


The difference-in-difference estimator yields a consistent estimate under the assumption that in the absence of partisan alignment all councils would have followed the same trends. One might be concerned that the aligned and non-

Example: results

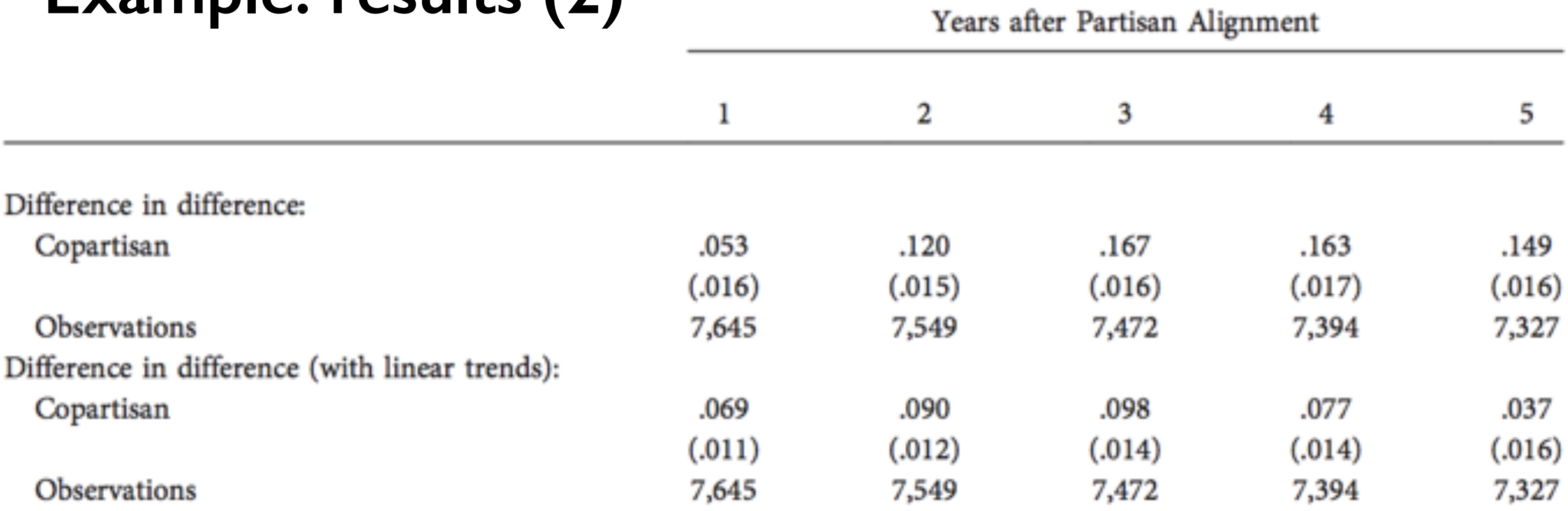


Example: results (2)



	Years after Partisan Alignment				
	1	2	3	4	5
Difference in difference:					
Copartisan	.053	.120	.167	.163	.149
	(.016)	(.015)	(.016)	(.017)	(.016)
Observations	7,645	7,549	7,472	7,394	7,327
Difference in difference (with linear trends):					
Copartisan	.069	.090	.098	.077	.037
	(.011)	(.012)	(.014)	(.014)	(.016)
Observations	7,645	7,549	7,472	7,394	7,327

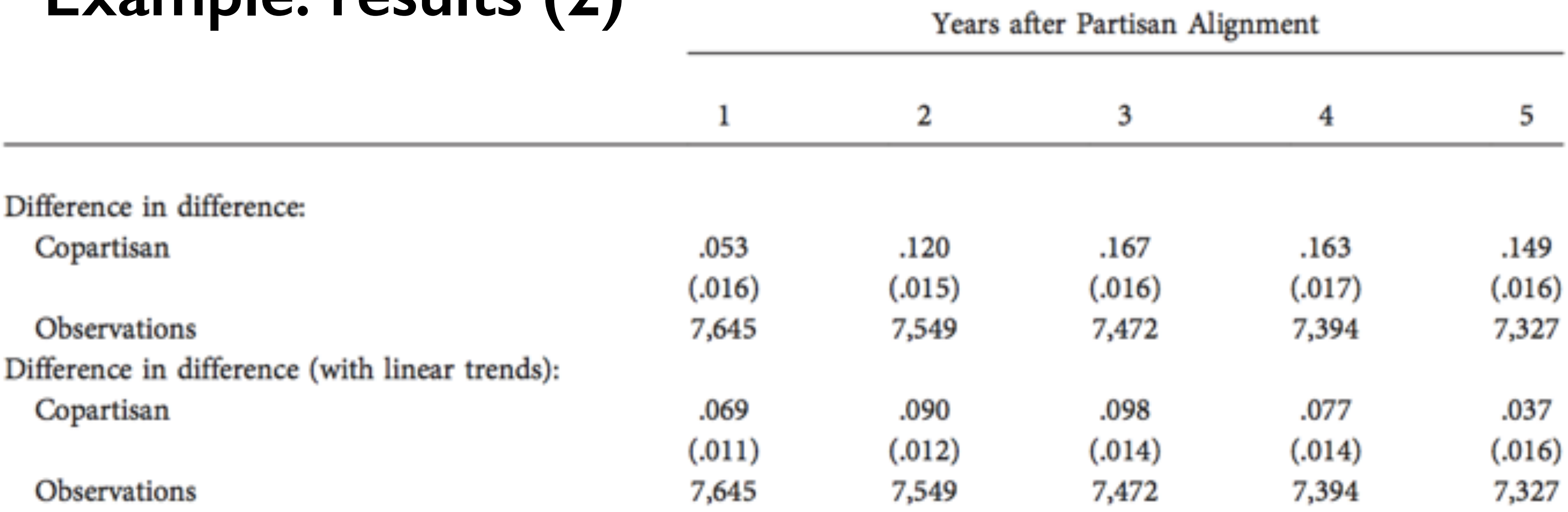
Example: results (2)



	Years after Partisan Alignment				
	1	2	3	4	5
Difference in difference:					
Copartisan	.053 (.016)	.120 (.015)	.167 (.016)	.163 (.017)	.149 (.016)
Observations	7,645	7,549	7,472	7,394	7,327
Difference in difference (with linear trends):					
Copartisan	.069 (.011)	.090 (.012)	.098 (.014)	.077 (.014)	.037 (.016)
Observations	7,645	7,549	7,472	7,394	7,327

Using interactions, they also show:

Example: results (2)

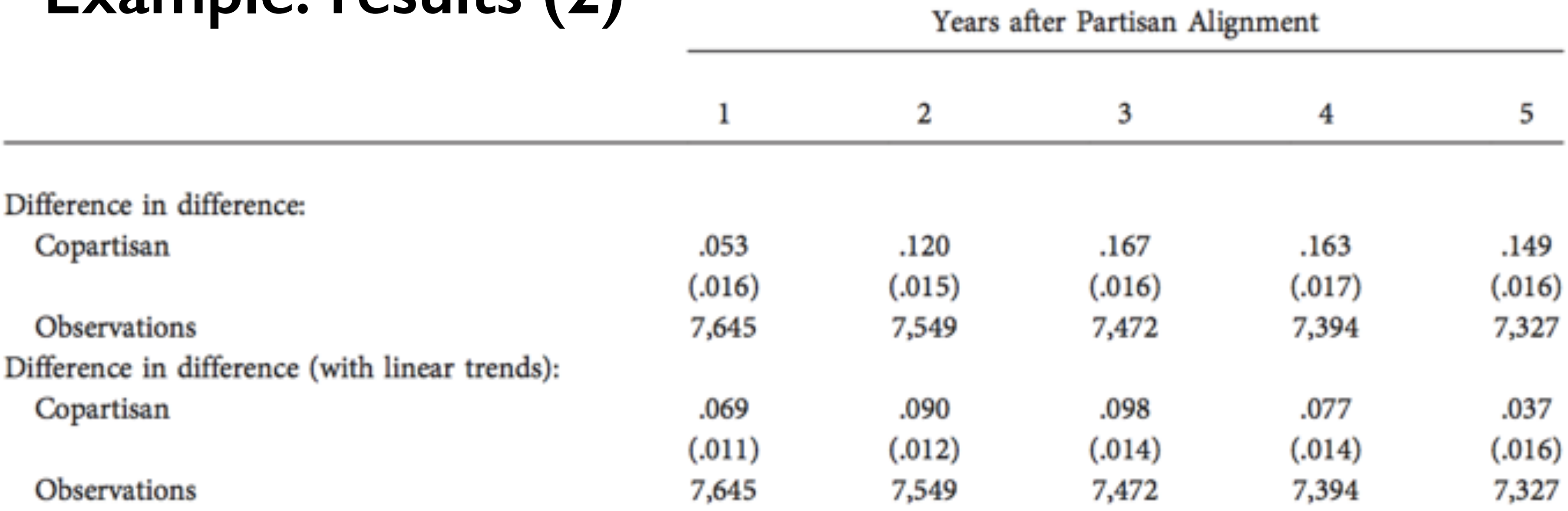


	Years after Partisan Alignment				
	1	2	3	4	5
Difference in difference:					
Copartisan	.053 (.016)	.120 (.015)	.167 (.016)	.163 (.017)	.149 (.016)
Observations	7,645	7,549	7,472	7,394	7,327
Difference in difference (with linear trends):					
Copartisan	.069 (.011)	.090 (.012)	.098 (.014)	.077 (.014)	.037 (.016)
Observations	7,645	7,549	7,472	7,394	7,327

Using interactions, they also show:

- larger effect before elections

Example: results (2)

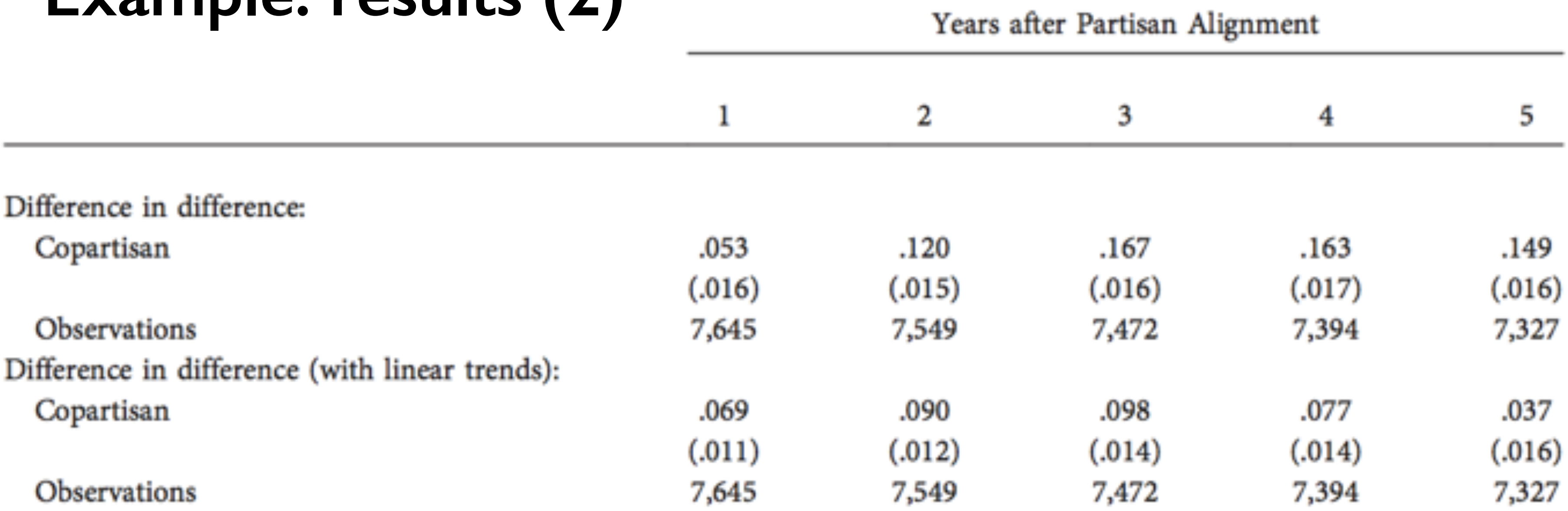


	Years after Partisan Alignment				
	1	2	3	4	5
Difference in difference:					
Copartisan	.053	.120	.167	.163	.149
	(.016)	(.015)	(.016)	(.017)	(.016)
Observations	7,645	7,549	7,472	7,394	7,327
Difference in difference (with linear trends):					
Copartisan	.069	.090	.098	.077	.037
	(.011)	(.012)	(.014)	(.014)	(.016)
Observations	7,645	7,549	7,472	7,394	7,327

Using interactions, they also show:

- larger effect before elections
- larger effect in county councils with less frequent elections

Example: results (2)

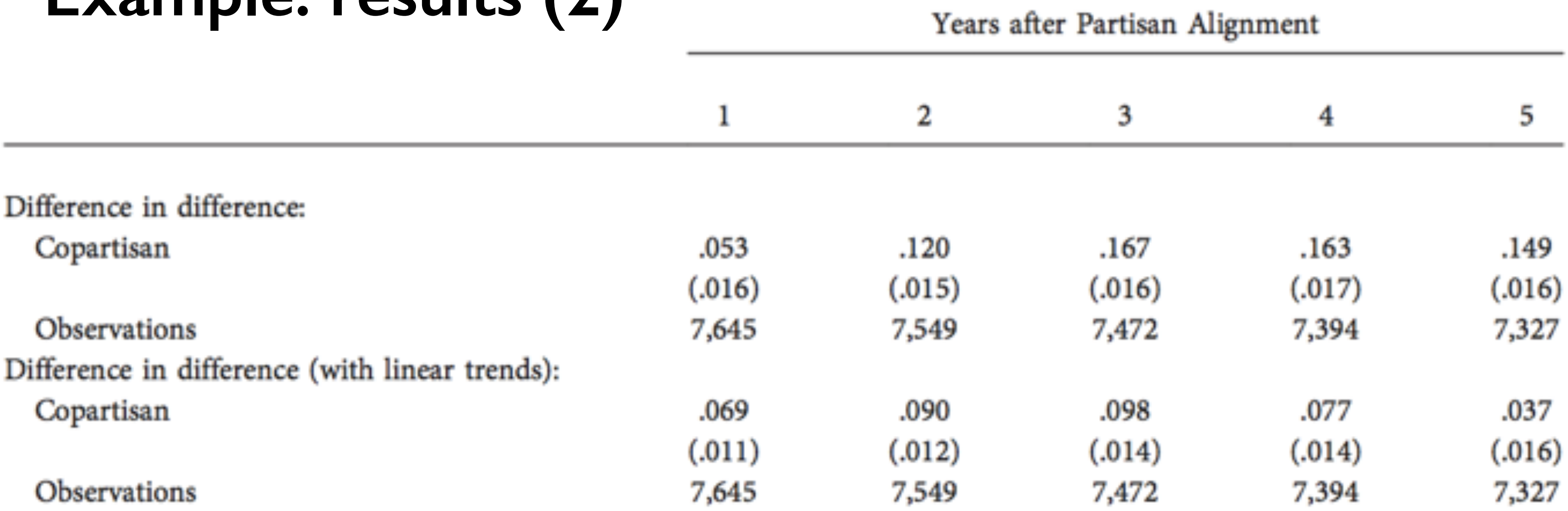


	Years after Partisan Alignment				
	1	2	3	4	5
Difference in difference:					
Copartisan	.053	.120	.167	.163	.149
	(.016)	(.015)	(.016)	(.017)	(.016)
Observations	7,645	7,549	7,472	7,394	7,327
Difference in difference (with linear trends):					
Copartisan	.069	.090	.098	.077	.037
	(.011)	(.012)	(.014)	(.014)	(.016)
Observations	7,645	7,549	7,472	7,394	7,327

Using interactions, they also show:

- larger effect before elections
- larger effect in county councils with less frequent elections
- larger effect in more competitive councils

Example: results (2)



	Years after Partisan Alignment				
	1	2	3	4	5
Difference in difference:					
Copartisan	.053	.120	.167	.163	.149
	(.016)	(.015)	(.016)	(.017)	(.016)
Observations	7,645	7,549	7,472	7,394	7,327
Difference in difference (with linear trends):					
Copartisan	.069	.090	.098	.077	.037
	(.011)	(.012)	(.014)	(.014)	(.016)
Observations	7,645	7,549	7,472	7,394	7,327

Using interactions, they also show:

- larger effect before elections
- larger effect in county councils with less frequent elections
- larger effect in more competitive councils

They also do a “triple-difference” analysis, but this is not what people usually call a “triple-difference”.

Triple differences (difference-in-differences-in-differences), briefly

Triple differences (difference-in-differences-in-differences), briefly

Suppose you do a classic diff-in-diff, but you know the parallel trends assumption probably doesn't hold.

Triple differences (difference-in-differences-in-differences), briefly

Suppose you do a classic diff-in-diff, but you know the parallel trends assumption probably doesn't hold.

Example: bicycles offered to girls in Bihar to help them get to school. What about using diff-in-diff to estimate effect of program, using boys as control group? (Muralidharan and Prakash 2017)

Triple differences (difference-in-differences-in-differences), briefly

Suppose you do a classic diff-in-diff, but you know the parallel trends assumption probably doesn't hold.

Example: bicycles offered to girls in Bihar to help them get to school. What about using diff-in-diff to estimate effect of program, using boys as control group? (Muralidharan and Prakash 2017)

To deal with likely violation of parallel trends assumption, could do diff-in-diff in neighboring state (without program) and subtract first DiD from second DiD => triple-diff.

In what circumstances is this better than using girls in neighboring state as control group instead of boys in Bihar?

General guidelines: when can panel data help with causal inferences

General guidelines: when can panel data help with causal inferences

Common to all of our designs:

General guidelines: when can panel data help with causal inferences

Common to all of our designs:

- when there is a clearly defined independent variable of interest (“treatment”)

General guidelines: when can panel data help with causal inferences

Common to all of our designs:

- when there is a clearly defined independent variable of interest (“treatment”)
- when you can imagine manipulating the treatment

General guidelines: when can panel data help with causal inferences

Common to all of our designs:

- when there is a clearly defined independent variable of interest (“treatment”)
- when you can imagine manipulating the treatment

General guidelines: when can panel data help with causal inferences

Common to all of our designs:

- when there is a clearly defined independent variable of interest (“treatment”)
- when you can imagine manipulating the treatment

Specific to panel data:

General guidelines: when can panel data help with causal inferences

Common to all of our designs:

- when there is a clearly defined independent variable of interest (“treatment”)
- when you can imagine manipulating the treatment

Specific to panel data:

- when you can get data for the same units over time

General guidelines: when can panel data help with causal inferences

Common to all of our designs:

- when there is a clearly defined independent variable of interest (“treatment”)
- when you can imagine manipulating the treatment

Specific to panel data:

- when you can get data for the same units over time
- when treatment changes over time for some units

General guidelines: when can panel data help with causal inferences

Common to all of our designs:

- when there is a clearly defined independent variable of interest (“treatment”)
- when you can imagine manipulating the treatment

Specific to panel data:

- when you can get data for the same units over time
- when treatment changes over time for some units
- when the treatment’s effects are not too delayed, or are delayed in a consistent manner