Casuality and Programme Evaluation

Lecture V: Difference-in-Differences II

Dr Martin Karlsson

University of Duisburg-Essen

Summer Semester 2017

Outline

1. Recap of Last Lecture
2. Introduction
3. Test Size
    - Within-Group Correlation
    - Serial Correlation
    - Brewer et al (2013)
4. Low Power
5. Feasible GLS: A Combined Approach
6. Summary and Conclusions

Recap of Last Lecture

- DID is a well-established, powerful and simple technique.
- Simplest case: **common time trend** is sufficient to achieve consistency.
- The basic 2×2 model can be extended in various directions:
  - Multiple groups, multiple periods
  - Models with covariates
  - Multiple dimensions (triple difference etc).
- Extensions for panel data and limited dependent variables exist, but can be more tricky.
- **Synthetic control methods** are a convenient way to define credible control groups at the aggregate level.
- The **changes-in-changes** model relaxes assumptions from the standard DID model; bases identification on monotonicity and invariance in the distribution of unobservables.

Introduction

- Recent literature on inference in DID designs focus on the problem of **incorrect test size**.
- In fact, such designs give rise to potential sources of correlation between observations.
- Two main issues:
  - Treatment status varies only at the group level ('**clustering problem**').
  - Treatment status typically highly correlated over time ('**policy autocorrelation**').
- If these issues are ignored, inference may be misleading.
- Most recent literature shifts the focus to **low power** issues.
- How to address the power-size trade-off?

Type I and Type II Errors

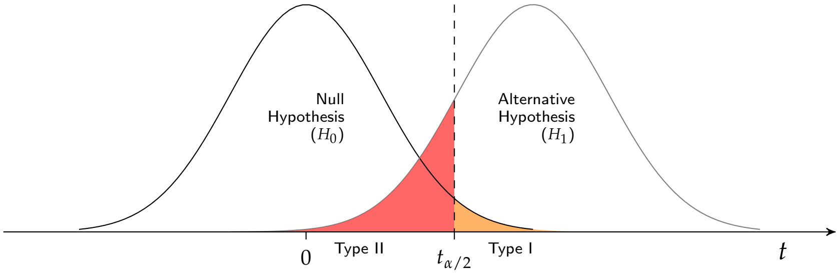

Figure 1. Type I and Type II Errors.

Problems with Standard Errors

- Recall from lecture 2: grouped residuals inflate standard errors.
- Consider the simple bivariate case

$$Y_{ig} = \alpha + \beta x_{ig} + e_{ig}$$

- where there are G groups and common group errors:

$$e_{ig} = v_g + \eta_{ig}$$

- Component v_g captures that group members are exposed to the same **environment**: classroom, teacher, weather...
- The **intraclass correlation coefficient** thus given by

$$\rho_e = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}$$

- ...but the OLS estimator assumes iid residuals ($v_g = 0$).

The Moulton Factor

- The Moulton factor: ratio between correct sampling variance and OLS variance.

$$\frac{\mathbb{V}\left(\hat{\beta}\right)}{\mathbb{V}_c\left(\hat{\beta}\right)} = 1 + \left[\frac{\mathbb{V}\left(n_g\right)}{\bar{n}} + \bar{n} - 1\right]\rho_x\rho_e \qquad (1)$$

- where

$$\rho_x = \frac{\sum_g \sum_j \sum_{i \neq j} \left(x_{ig} - \bar{x}\right)\left(x_{jg} - \bar{x}\right)}{\mathbb{V}\left(x_{ig}\right) \sum_n n_g \left(n_g - 1\right)} \qquad (2)$$

- Hence, the standard errors get inflated whenever
  - Intraclass correlation is high (ρ_e).
  - Group size varies considerably ($\mathbb{V}\left(n_g\right)$).
  - High intraclass correlation also in x_{ig} (ρ_x)
- At least two of these apply by design in a DID setting.

Two Dimensions of the Problem

- The general conclusion: OLS **underestimates** standard errors ⇒ correction needed.
- Two dimensions:
  - (A) **Within-group correlation**. Shared environment leads to correlated shocks.
  - (B) **Serial correlation**. Outcomes typically exhibit persistence (earnings, employment, health...).
- **Additional complication**: Number of groups and time periods typically **small**.
- Inference based on G or T approaching infinity.

A. Within-Group Correlation

- Donald and Lang (*Rev. Econ. Statist.* 2007) discuss inference in DID and related models.
- Focus on **within-group correlation** of outcomes.
- Problem: some explanatory variables (like the treatment indicator) are **constant** among all members of a group.
- Three traditional solutions:
  1. RE FGLS estimation. Estimate covariance matrix, reweight.
  2. Correcting standard errors using covariance matrix with common group errors (Moulton 1990).
  3. Cluster (Liang and Zeger 1986).
- D&L: These procedures based on $G \to \infty$.
- Consider instead aggregating and drawing inference using T_{G-2}.

B. Serial Correlation, an Example

Bertrand et al (2004) utilise a standard dataset (the Current Population Survey - CPS):

$$Y_{ist} = A_s + B_t + \tau D_{st} + X_{ist}\beta + \epsilon_{ist} \tag{3}$$

where

Y_{ist} **Log weekly earnings** of females between 25-50 at t 1979 to 2000 in state s.

D_{st} Treatment indicator $= 1$ if state s is affected in year t.

A_s State fixed effects.

B_t Year fixed effects.

X_{ist} Individual-level control variables.

ϵ_{ist} Residual variation.

- Y_{is} exhibits strong **positive serial correlation**: $\rho_1 = 0.51$, $\rho_2 = 0.44$ and $\rho_3 = 0.33$.
- In total $50 \times 21 = 1,050$ state-year cells.

The Problem

- OLS gives an **unbiased** and **consistent** estimate $\hat{\tau}$ of effect.
- Bertrand et al run Monte Carlo simulations using **placebo law changes**.
- With consistent standard errors, false treatment effect should be observed in roughly $\mathbf{5}\%$ of cases.
- But standard errors are often **inconsistent**.
- H_0 is rejected in $\mathbf{67.5}\%$ of cases when neither within-group correlation nor serial correlation are taken into account.
- Taking **within-group** correlation into account (cluster or aggregate): H_0 is rejected in $\mathbf{44}\%$ of cases.
- **Serial** correlation can matter a lot!
- Many approaches to address the problem; none is uniformly better.

Serial Correlation: Solutions

To evaluate possible solutions to the serial correlation problem, Bertrand et compare the simulated performance of five different techniques:

1. **Parametric methods** $(AR(p))$: perform poorly.
2. **Block bootstrap**: (sample clusters and calculate t statistic) performs well when the no. of groups is **large**.
3. **Aggregate (collapse) time series information**: reliable also when the no. of groups is small, on the other hand power is relatively low.
4. **Empirical variance-covariance matrix**: performs well in panels with high no. of groups, but assumes cross-sectional homoskedasticity (cf. Hausman & Kuersteiner, 2008).
5. **Arbitrary (clustered) variance-covariance matrix**: allows for an arbitrary correlation patterns over time. Performs well for moderate no. of groups; for small no. of groups d.o.f. adjustment needed.

The Clustered Covariance Matrix Estimator

- The empirical VCV estimator is consistent only under **homoskedasticity**.
- A robust alternative is the *Clustered Covariance Matrix* estimator (CCM; cf. Arellano, 1987):

$$\Sigma = (\mathrm{Z}'\mathrm{Z})^{-1}\left(\sum_{s=1}^{N} e_s' e_s\right)(\mathrm{Z}'\mathrm{Z})^{-1}.$$

where

Z Matrix of independent variables (i.e. A_s, B_t and D_{st}) with NT vectors z_{st}.

e_s $\sum_{t=1}^{T} \upsilon_{st} z_{st}$.

υ_{st} Estimated residuals for state s at time t.

CCM: Properties

- The estimation procedure that uses SEs computed according to the CCM performs quite well in finite samples.
- Approximately correct size regardless of relationship btw. N and T.
- However, there is still **overrejection** to some extent when the number of states is **small**: Bertrand et al reject H_0 in 8% (11%) of cases using a sample from 10 (6) states.
- Much better than before, but still twice nominal test size.

CCM: Properties II

- Asymptotic properties of CCM estimator for $N \to \infty$ are well known.
- Even without restrictions on the serial dependence, $\hat{\Sigma}$ is $\sqrt{N}$-consistent and asymptotically normal.
- But in DID studies, we often have **small samples**, in which robust standard errors are **downwards biased**.

Hansen Correction

Hansen (2007a) derives properties of $\hat{\Sigma}$ for $T \rightarrow \infty$, N fixed:

- Even if $\{z_{st}, v_{st}\}$ is a **strong mixing sequence** (i.e. temporal dependence decreases in distance), $\hat{\Sigma}$ is no longer consistent.
- If $\text{Var}\left(z_s\right)$ and Σ_s are the same for all s, standard t-statistics will be scaled by a factor of $\frac{(N-1)}{N}$.
- Thus, using $\left(\frac{N}{N-1}\right)\hat{\Sigma}$ and a t_{N-1} distribution will provide **asymptotically unbiased** inference – irrespective of dimension approaching infinity.

Brewer et al (2013): Test Size

- Brewer et al (2013): correct size can be obtained quite easily – **even when G is low!**.
- Consider Model 3. The ‘benchmark’ is the OLS estimator of $\hat{\beta}$’s standard error, *i.e. assuming that errors are i.i.d.*
- To get cluster-robust standard errors (CRSE), they use Liang and Zeger’s (1986) formula to compute a cluster-robust variance matrix.

Brewer et al (2013): Test Size II

- The estimator is consistent and Wald statistics are asymptotically normal as the no. of groups $G \to \infty$.
- But it is **biased** (SE downward biased).
- The bias can be substantial when G is small.
- One way to *reduce* such bias is to **scale up** the residuals by $\sqrt{\frac{G}{G-1}}$ before plugging them into the CRSE estimator.
- An alternative is to recover empirically the distribution of the t-statistic using a bootstrap procedure.
- The **wild cluster bootstrap-t procedure** by Cameron et al (2008) outperformed other bootstrap-based approaches and works well also with small G.

Wild Cluster Bootstrap-t

Cf. Cameron & Miller (2013) *A Practitioner's Guide to Cluster-Robust Inference*.

1. Estimate with OLS, imposing $H_0 : \beta_1 = \beta_1^0$ and recover residual vectors $\{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_G\}$.
2. Generate pseudo-residuals as $\hat{\mathbf{u}}_g^* = \hat{\mathbf{u}}_g$ or $\hat{\mathbf{u}}_g^* = -\hat{\mathbf{u}}_g$; each with probability 0.5 – and the resulting pseudo-sample $\left\{ (\hat{\mathbf{y}}_1^*, \mathbf{X}_1), \ldots, (\hat{\mathbf{y}}_G^*, \mathbf{X}_G) \right\}$.
3. Generate OLS estimate $\hat{\beta}_{1,b}^*$, standard error $s_{\hat{\beta}_{1,b}^*}$ and Wald statistic $w_b^* = \left(\hat{\beta}_{1,b}^* - \beta_1^0 \right) / s_{\hat{\beta}_{1,b}^*}$.
4. Repeat for $b = 1, \ldots, B$.
5. Reject H_0 at level α if $w \notin [w_{\alpha/2}, w_{1-\alpha/2}]$.

Summary

Brewer et al address both serial correlation and within-group correlation in the following steps:

- Aggregate data on state-year level.
- Apply a scaling factor to the residuals: $\sqrt{\frac{G}{G-1}}$.
- Plug the scaled residuals into the cluster-robust variance-covariance matrix to get cluster-robust standard errors (CRSE).
- Use critical values from a t distribution with d.o.f. correction: t_{G-1} instead of a standard normal.

Experimental Design

- They use the same data as Betrand et al on the period 1979-2008 and placebo law changes with tests of nominal 5% size.
- Monte Carlo simulations to show that their procedure allows to build tests with the intended test size.
- Resample states with replacement; half of the states are 'treated'.
- They use OLS and FGLS and compare rejection rates, assuming different inference methods and different number of groups:

  1. Errors i.i.d.
  2. CRSE, unscaled residuals and $N(0,1)$
  3. CRSE, unscaled residuals and t_{G-1}
  4. CRSE, scaled residuals and $N(0,1)$
  5. CRSE, scaled residuals and t_{G-1}
  6. Wild cluster bootstrap-t

  - 6, 10, 20, 50 states resampled.

Experimental Design II

- The purpose is to compare the performance of the different methods in terms of both Type I **and** Type II errors.
- Robustness checks:
  - Robustness to mis-specification of the error process: State-time shocks simulated according to an AR(1) process with varying parameters.
  - Vary the fraction of treated groups to check performance in **unbalanced designs**.

Compare Methods

Table 1. Rejection rates when the null is true. Tests of nominal 5% size with placebo treatments in log-earnings data, 5,000 replications. Equation 3 is estimated by OLS.

Inference method	$G = 50$	$G = 20$	$G = 10$	$G = 6$
i.i.d. errors	0.429	0.424	0.422	0.413
	(0.007)	(0.007)	(0.007)	(0.007)
CRSE, $N(0,1)$ critical values	0.059	0.073	0.110	0.175
	(0.003)	(0.004)	(0.004)	(0.005)
CRSE, t_{G-1} critical values	0.053	0.056	0.066	0.095
	(0.003)	(0.003)	(0.004)	(0.004)
CRSE, $\sqrt{\frac{G}{(G-1)}}$ residuals, $N(0,1)$	0.049	0.056	0.071	0.113
	(0.003)	(0.003)	(0.004)	(0.004)
CRSE, $\sqrt{\frac{G}{(G-1)}}$ residuals, t_{G-1}	0.045	0.041	0.042	0.052
	(0.003)	(0.003)	(0.003)	(0.003)
Wild cluster bootstrap-t	0.044	0.041	0.048	0.059
	(0.003)	(0.003)	(0.003)	(0.003)

Simulation standard errors in parentheses. The treatment parameter has a true coefficient of zero. G number of sampled states. Data from 1976 to 2008 inclusive $(T = 30)$.

Imbalance between Groups

Table 2. Rejection rates when the null is true. Tests of nominal 5% size with placebo treatments in log-earnings data, 5,000 replications. Equation 3 is estimated by OLS.

Inference method	$G1 = 5$	$G1 = 4$	$G1 = 3$	$G1 = 2$
i.i.d. errors	0.422	0.408	0.409	0.405
	(0.007)	(0.007)	(0.007)	(0.007)
CRSE, $N(0,1)$ critical values	0.110	0.125	0.150	0.241
	(0.004)	(0.005)	(0.005)	(0.006)
CRSE, t_{G-1} critical values	0.066	0.079	0.105	0.191
	(0.004)	(0.004)	(0.004)	(0.006)
CRSE, $\sqrt{\frac{G}{(G-1)}}$ residuals, $N(0,1)$	0.071	0.084	0.113	0.199
	(0.004)	(0.004)	(0.004)	(0.006)
CRSE, $\sqrt{\frac{G}{(G-1)}}$ residuals, t_{G-1}	0.042	0.051	0.074	0.150
	(0.003)	(0.003)	(0.004)	(0.005)
Wild cluster bootstrap-t	0.048	0.054	0.052	0.018
	(0.003)	(0.003)	(0.003)	(0.002)

Simulation standard errors in parentheses. The treatment parameter has a true coefficient of zero. $G1$ the number of treated out of a total of 10 states. Data from 1976 to 2008 inclusive $(T = 30)$.

Size *vs* Power

- The proposed combined modifications (scaled CRSE and t_{S-1} critical values) yield good results in most cases: **true test size is within about 1% of nominal test size**.
- Large **imbalance** between the numbers of treatment and control groups $\Rightarrow$ **wild cluster bootstrap-t** procedure performs better.
- However, Brewer et al stress that it is relatively easy to obtain the correct test size.
- The main issue is that the **power** to detect real treatment effects with tests of the correct size is **low**.
- It is **extremely low** when S is small.

(Low) Power with OLS

Table 3. Rejection rates of H_0 : *no treatment effect* when β is the true value of the treatment parameter. Tests of nominal 5% size with placebo treatments in log-earnings data, 5,000 replications. Comparison of different inference methods.

True effect	Inference	$G=50$	$G=20$	$G=10$	$G=6$
$\beta=0.02$	$\sqrt{\frac{G}{(G-1)}}$CRSE, t_{G-1} critical values	0.238	0.134	0.088	0.074
		(0.006)	(0.005)	(0.004)	(0.004)
	wild cluster bootstrap-t	0.225	0.125	0.093	0.074
		(0.006)	(0.005)	(0.004)	(0.004)
$\beta=0.05$	$\sqrt{\frac{G}{(G-1)}}$CRSE, t_{G-1} critical values	0.822	0.513	0.273	0.168
		(0.005)	(0.007)	(0.006)	(0.005)
	wild cluster bootstrap-t	0.799	0.490	0.283	0.167
		(0.006)	(0.007)	(0.006)	(0.005)
$\beta=0.10$	$\sqrt{\frac{G}{(G-1)}}$CRSE, t_{G-1} critical values	1.000	0.919	0.718	0.448
		(0.000)	(0.004)	(0.006)	(0.007)
	wild cluster bootstrap-t	0.999	0.898	0.712	0.429
		(0.000)	(0.004)	(0.006)	(0.007)
$\beta=0.15$	$\sqrt{\frac{G}{(G-1)}}$CRSE, t_{G-1} critical values	1.000	0.995	0.904	0.755
		(.)	(0.001)	(0.004)	(0.006)
	wild cluster bootstrap-t	1.000	0.992	0.896	0.700
		(.)	(0.001)	(0.004)	(0.006)

Simulation standard errors in parentheses. G number of sampled states. Data from 1976 to 2008 inclusive $(T=30)$.

Minimum Detectable Effect

- The power of the two inference methods is similar
- Power is documented more comprehensively when looking at the minimum effect that would be detected (**Minimum Detectable Effect** - MDE).
- Recall: the MDE is defined as

$$MDE(\kappa) = \widehat{SE(\hat{\beta})}\left[c_u + p^t_{1-\kappa}\right]$$

where

κ	Level of power.
$\widehat{SE(\hat{\beta})}$	Scaled CRSE estimate.
c_u	Upper critical value of the t_{S-1} distribution.
$p^t_{1-\kappa}$	$(1-\kappa)th$ percentile of the t-statistic under H_0 : no treatment effect.

Minimum Detectable Effect: Illustration

Figure 2 shows the proportion of times for which H_0 : no treatment effect is rejected when the treatment parameter β has a coefficient between 0 and 0.3.

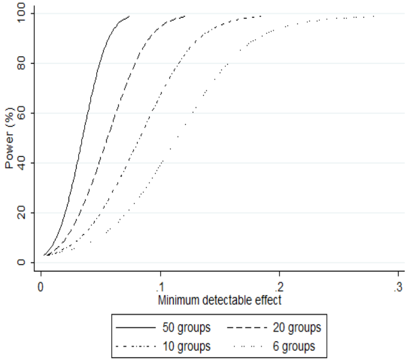

Figure 2. MDE on log-earnings with scaled residuals, t_{G-1} critical values and tests of 5% size, 100,000 replications.

Solution: Increasing power with FGLS

- Hansen (2007b) proposes to use a FGLS estimation under the assumption that the state-year shock follows a **stationary AR(p)** process.
- Coefficients of the AR(p) process can be biased in panel data if the time dimension is short and fixed effects are included (**incidental parameters** problem).
- He also introduces a **bias correction** to account for this.
- He finds that the FGLS estimation clearly dominates OLS also when inference is based on CRSE.
- The FGLS procedure with bias correction (BC-FGLS) is consistent as $S \to \infty$.

Solution: Increasing power with FGLS II

- Brewer et al use simulations to show that it is possible to retain the correct test size and achieve gains in power by using **FGLS** instead of OLS.
- Combine BC-FGLS with robust inference technique (scaled CRSE and critical values from t with d.o.f. adjustment).
- In fact, the size of the test can be controlled using robust inference, even for small S.
- In this way tests have the correct size and FGLS improves power considerably.
- Procedure also robust to mis-specifications of the error process.

A Performance Comparison

Table 4. Rejection rates for tests of nominal 5% size with placebo treatments in log-earnings data, 5,000 replications, comparison of different estimation and inference methods.

Estimation	Inference	$G = 50$	$G = 20$	$G = 10$	$G = 6$
OLS	$\sqrt{\frac{G}{(G-1)}}$CRSE, t_{G-1} critical values	0.045	0.041	0.042	0.052
		(0.003)	(0.003)	(0.003)	(0.003)
FGLS	(no correction)	0.106	0.101	0.120	0.124
		(0.004)	(0.004)	(0.005)	(0.005)
FGLS	$\sqrt{\frac{G}{(G-1)}}$CRSE, t_{G-1} critical values	0.049	0.045	0.054	0.061
		(0.003)	(0.003)	(0.003)	(0.003)
BC-FGLS		0.073	0.070	0.087	0.096
		(0.004)	(0.004)	(0.004)	(0.004)
BC-FGLS	$\sqrt{\frac{\mathbf{G}}{(\mathbf{G-1})}}$**CRSE**, $\mathbf{t_{G-1}}$ **critical values**	0.049	0.045	0.058	0.065
		(0.003)	(0.003)	(0.003)	(0.003)

Simulation standard errors in parentheses. The treatment parameter has a true coefficient of zero. G number of sampled states. Data from 1976 to 2008 inclusive $(T = 30)$.

Assumptions about the Serial Correlation Process

Table 5. Rejection rates for tests of nominal 5% size with placebo treatments in log-earnings data, 5,000 replications, 10 groups. Empirical regression residuals (CPS) replaced by a simulated error term generated according to an AR(2) and a MA(1) process.

Estimation	Inference	CPS residuals	Heterogeneous AR(2)	MA(1)
OLS	$\sqrt{\frac{G}{(G-1)}}$ CRSE, t_{G-1}	0.049	0.040	0.052
		(0.003)	(0.002)	(0.002)
FGLS	(no correction)	0.114	0.101	0.088
		(0.004)	(0.003)	(0.003)
FGLS	$\sqrt{\frac{G}{(G-1)}}$ CRSE, t_{G-1}	0.054	0.055	0.051
		(0.003)	(0.002)	(0.002)
BC-FGLS		0.081	0.072	0.072
		(0.004)	(0.003)	(0.003)
BC-FGLS	$\sqrt{\frac{G}{(G-1)}}$ CRSE, t_{G-1}	0.056	0.059	0.052
		(0.003)	(0.002)	(0.002)

Simulation standard errors in parentheses. The treatment parameter has a true coefficient of zero. G number of sampled states. Data from 1976 to 2008 inclusive $(T = 30)$.

Gains in Power

Table 6. Rejection rates for tests of nominal 5% size with a treatment effect of +0.05 in log earnings data

Estimation	Inference	$G = 50$	$G = 20$	$G = 10$	$G = 6$
OLS	$\sqrt{\frac{G}{(G-1)}}$ CRSE, t_{G-1}	0.810	0.467	0.252	0.168
		(0.006)	(0.007)	(0.006)	(0.005)
FGLS	(no correction)	0.985	0.799	0.573	0.434
		(0.002)	(0.006)	(0.007)	(0.007)
FGLS	$\sqrt{\frac{G}{(G-1)}}$ CRSE, t_{G-1}	0.957	0.670	0.401	0.255
		(0.003)	(0.007)	(0.007)	(0.006)
BC-FGLS		0.978	0.763	0.513	0.384
		(0.002)	(0.006)	(0.007)	(0.007)
BC-FGLS	$\sqrt{\frac{G}{(G-1)}}$ CRSE, t_{G-1}	0.955	0.696	0.423	0.286
		(0.003)	(0.007)	(0.007)	(0.006)

Simulation standard errors in parentheses. Data from 1976 to 2008 inclusive $(T = 30)$.

Summary and Conclusions

- In DID designs: great risk of **underestimating standard errors**:
  - Dependent variables tend to be positively serially correlated.
  - Treatment tends to be serially correlated as well.
- Various methods allow to correct for this problem, e.g. bootstrapping or robust covariance matrix estimators (clustered covariance matrix estimator).
- However, even when test size is correct, one issue is **low power**.
- Brewer et al (2013) – get accurate size (easy), then maximise power:
  - Robust inference (**CRSE** with **scaled residuals** & $\mathbf{t_{S-1}}$ critical values)
  - ...coupled with a **FGLS** estimation procedure.
- Performs quite well also when S is small and is robust to mis-specifications of the error process.

1. Parametric Methods

- Parametric methods specify an **autocorrelation structure**, which is then estimated:
    - It may be either **individual-specific** or **uniform**.
    - This was traditionally the common approach to deal with the problem.
    - OLS residuals used to estimate **autocorrelation parameters** (ρ).
    - Finally, employ ρ's in an FGLS regression.
- **Problem:** In short time series, autocorrelation parameters estimated by OLS are biased **downwards**!
- **Consequence:** Over-rejection remains a problem.
- Hansen's bias correction is consistent for $S \to \infty$ (T fixed).

2. Boostrapping

Simple Bootstrapping.

- **Boostrapping**: a technique used when we are **unable** or **unwilling** to derive the **distribution** of our estimator.
- A simple bootstrapping scheme draws R samples of size N from our original sample.
- On each of these samples, we run our main regression.
- Our R estimated parameters $\hat{\tau}_r$ will mimic the distribution of $\hat{\tau}$.

Block Boostrap

- Block bootstrap preserves the autocorrelation structure by using **series of observations** instead of individual observations.
  1. Bootstrap sample is generated by drawing N_s matrices $(\mathbf{Y_s}, \mathbf{V_s})$:
     - $\mathbf{Y_s}$ is the **entire series** of observations for state s.
     - $\mathbf{V_s}$ is the matrix of D, state & time dummies for state s.
  2. Run OLS on each sample, obtain estimate $\hat{\beta}$ and absolute t statistic

$$t_r = \frac{|\hat{\beta}_r - \hat{\beta}|}{SE\left(\hat{\beta}_r\right)}$$

  3. t_r approaches the sampling distribution of t as R increases.
- **Assessment**: Significant improvement over parametric techniques, but **many groups** required.
- Implemented in Stata by `xtreg yvar treatvar xvars, i(id) fe vce(bootstrap, seed(1234))`.

3. Ignoring Time Series Information

- Simpler alternative: **ignore** time series information.
- For laws implemented **at the same time** in all treated groups, we can simply compute pre- and post-reform **averages** for each group.
- If not, proceed as follows:
  1. Start with a regression leaving treatment indicator out

$$Y_{st} = A_s + B_t + X_{st}\beta + v_{st}$$

  2. Calculate before and after averages for treated groups only:

$$\hat{v}_s^0 = \frac{\sum_{t=1}^{T} \mathbb{1}\left(D_{st} = 0\right) \hat{v}_{st}}{\sum_{t=1}^{T} \mathbb{1}\left(D_{st} = 0\right)} = \frac{\sum_{t=1}^{T} \mathbb{1}\left(D_{st} = 0\right) \left(Y_{st} - \hat{A}_s - \hat{B}_t - X_{st}\hat{\beta}\right)}{\sum_{t=1}^{T} \mathbb{1}\left(D_{st} = 0\right)}$$

$$\hat{v}_s^1 = \frac{\sum_{t=1}^{T} \mathbb{1}\left(D_{st} = 1\right) \hat{v}_{st}}{\sum_{t=1}^{T} \mathbb{1}\left(D_{st} = 1\right)} = \frac{\sum_{t=1}^{T} \mathbb{1}\left(D_{st} = 1\right) \left(Y_{st} - \hat{A}_s - \hat{B}_t - X_{st}\hat{\beta}\right)}{\sum_{t=1}^{T} \mathbb{1}\left(D_{st} = 1\right)}$$

Ignoring Time Series Information II

3. Run the regression

$$\hat{\hat{v}}_{st} = \tau D_{st} + u_{st}$$

- When S is **small**, need to make a correction to the t statistic.
- Simple aggregation performs well, and residual aggregation has reasonable rejection rates as well.
- But **power** tends to be very low!

4. Empirical VCV Matrix

- Parametric corrections unnecessarily inflexible:
  - $S > 1$, so we can estimate the covariance matrix more flexibly...
  - ...if we are willing to assume autocorrelation structure is **the same**...
  - ...and **homoskedastic** (Kiefer, 1980; Hausman and Kuersteiner, 2008).
- Thus, we express the dataset in vector form, where $\mathbf{Y_s}$ is the $T \times 1$ vector of outcome observations.
- We want to estimate the $T \times T$ matrix Σ.

Empirical VCV Matrix II

- Consider the empirical covariance matrix

$$\widehat{\sum}^{*} = \frac{1}{N} \sum_{s=1}^{N} \left[Q \left(\mathbf{Y}_s - \tau D_s - X_s \beta \right) \right] \left[Q \left(\mathbf{Y}_s - \tau D_s - X_s \beta \right) \right]'$$

  where Q performs a **within transformation**.

- And then use the estimated matrix to compute standard errors:

$$\text{Var}\left(\widehat{\beta}^{*} \right) = \left[\sum_{s=1}^{N} Z_s' Q \left(\widehat{\sum}^{*} \right)^{-1} Q Z_s \right]^{-1}$$

- The matrix $\widehat{\sum}^{*}$ has rank $T-1$: we need a **generalised inverse** such as suggested by Hsiao (2003).
- This method performs well when S is **large**.