# Literature on Recent Advances in Applied Micro Methods*

Christine Cai†

June 9, 2024

CLICK HERE FOR THE MOST RECENT VERSION [DARK MODE VERSION]

## Contents



*Many of the references listed here are from the applied econometrics courses taught by Michal Kolesár (ECO 539B, Fall 2019, Princeton University) and Arindrajit Dube (ECON 797B, Fall 2020, UMass Amherst). I also added papers that have circulated on #EconTwitter and those that have been suggested to me after I shared the first version of this draft. In particular, I thank Kirill Borusyak, Otavio Canozzi Conceição, Peter Cribbett, Dylan (@dc_pov), Brian Finley, Kevin Grier, Sean Higgins, Jakob Miethe, Tod Mijanovich, Pedro Picchetti, Esteban Quiñones, Jorge Rodríguez, Davud Rostam-Afschar, and Ben Solow for their suggestions. Additional suggestions are welcome ("more is better" ☺).

†Princeton University, Department of Economics. Email: christine.cai@princeton.edu

# 1 OLS

- **Ibragimov and Müller (2016), "Inference with Few Heterogeneous Clusters," RE-Stat**
  "Suppose estimating a model on each of a small number of potentially heterogeneous clusters yields approximately independent, unbiased, and Gaussian parameter estimators. We make two contributions in this setup. First, we show how to compare a scalar parameter of interest between treatment and control units using a two-sample $t$-statistic, extending previous results for the one-sample $t$-statistic. Second, we develop a test for the appropriate level of clustering; it tests the null hypothesis that clustered standard errors from a much finer partition are correct. We illustrate the approach by revisiting empirical studies involving clustered, time series, and spatially correlated data."

- **Imbens and Kolesár (2016), "Robust Standard Errors in Small Samples: Some Practical Advice," REStat**
  "We study the properties of heteroskedasticity-robust confidence intervals for regression parameters. We show that confidence intervals based on a degrees-of-freedom correction suggested by Bell and McCaffrey (2002) are a natural extension of a principled approach to the Behrens-Fisher problem. We suggest a further improvement for the case with clustering. We show that these standard errors can lead to substantial improvements in coverage rates even for samples with fifty or more clusters.We recommend that researchers routinely calculate the Bell-McCaffrey degrees-of-freedom adjustment to assess potential problems with conventional robust standard errors."

- **Canay, Santos, and Shaikh (2018), "The Wild Bootstrap with a "Small" Number of "Large" Clusters," REStat**
  "This paper studies the wild bootstrap-based test proposed in Cameron et al. (2008). Existing analyses of its properties require that number of clusters is "large." In an asymptotic framework in which the number of clusters is "small," we provide conditions under which an unstudentized version of the test is valid. These conditions include homogeneity-like restrictions on the distribution of covariates. We further establish that a studentized version of the test may only over-reject the null hypothesis by a "small" amount that decreases exponentially with the number of clusters. We obtain qualitatively similar result for "score" bootstrap-based tests, which permit testing in nonlinear models."

- **Cattaneo, Jansson, and Newey (2018), "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity," JASA**

"The linear regression model is widely used in empirical work in economics, statistics, and many other disciplines. Researchers often include many covariates in their linear model specification in an attempt to control for confounders. We give inference methods that allow for many covariates and heteroscedasticity. Our results are obtained using high-dimensional approximations, where the number of included covariates is allowed to grow as fast as the sample size. We find that all of the usual versions of Eicker-White heteroscedasticity consistent standard error estimators for linear models are inconsistent under this asymptotics. We then propose a new heteroscedasticity consistent standard error formula that is fully automatic and robust to both (conditional) heteroscedasticity of unknown form and the inclusion of possibly many covariates. We apply our findings to three settings: parametric linear models with many covariates, linear panel models with many fixed effects, and semiparametric semi-linear models with many technical regressors. Simulation evidence consistent with our theoretical results is provided, and the proposed methods are also illustrated with an empirical application. Supplementary materials for this article are available online."

- **Gibbons, Serrato, and Urbancic (2018), "Broken or Fixed Effects?," JE**
  "We replicate eight influential papers to provide empirical evidence that, in the presence of heterogeneous treatment effects, OLS with fixed effects (FE) is generally not a consistent estimator of the average treatment effect (ATE). We propose two alternative estimators that recover the ATE in the presence of group-specific heterogeneity. We document that heterogeneous treatment effects are common and the ATE is often statistically and economically different from the FE estimate. In all but one of our replications, there is statistically significant treatment effect heterogeneity and, in six, the ATEs are either economically or statistically different from the FE estimates."

- **Pustejovsky and Tipton (2018), "Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models," JBES**
  "In panel data models and other regressions with unobserved effects, fixed effects estimation is often paired with cluster-robust variance estimation (CRVE) to account for heteroscedasticity and un-modeled dependence among the errors. Although asymptotically consistent, CRVE can be biased downward when the number of clusters is small, leading to hypothesis tests with rejection rates that are too high. More accurate tests can be constructed using bias-reduced linearization (BRL), which corrects the CRVE based on a working model, in conjunction with a Satterthwaite approximation for $t$-tests. We propose a generalization of BRL that can be applied in models with arbitrary sets of fixed effects, where the original BRL method is undefined, and describe how to apply

the method when the regression is estimated after absorbing the fixed effects. We also propose a small-sample test for multiple-parameter hypotheses, which generalizes the Satterthwaite approximation for $t$-tests. In simulations covering a wide range of scenarios, we find that the conventional cluster-robust Wald test can severely over-reject while the proposed small-sample test maintains Type I error close to nominal levels. The proposed methods are implemented in an R package called `clubSandwich`. This article has online supplementary materials."

- **Abadie, Athey, Imbens, and Wooldridge (2020), "Sampling-Based Versus Design-Based Uncertainty in Regression Analysis," ECMA**
  "Consider a researcher estimating the parameters of a regression function based on data for all 50 states in the United States or on data for all visits to a website. What is the interpretation of the estimated parameters and the standard errors? In practice, researchers typically assume that the sample is randomly drawn from a large population of interest and report standard errors that are designed to capture sampling variation. This is common even in applications where it is difficult to articulate what that population of interest is, and how it differs from the sample. In this article, we explore an alternative approach to inference, which is partly design-based. In a design-based setting, the values of some of the regressors can be manipulated, perhaps through a policy intervention. Design-based uncertainty emanates from lack of knowledge about the values that the regression outcome would have taken under alternative interventions. We derive standard errors that account for design-based uncertainty instead of, or in addition to, sampling-based uncertainty. We show that our standard errors in general are smaller than the usual infinite-population sampling-based standard errors and provide conditions under which they coincide."

- **Colella, Lalive, Sakalli, and Thoenig (2020), "Inference with Arbitrary Clustering," WP**
  "Analyses of spatial or network data are now very common. Nevertheless, statistical inference is challenging since unobserved heterogeneity can be correlated across neighboring observational units. We develop an estimator for the variance-covariance matrix (VCV) of OLS and 2SLS that allows for arbitrary dependence of the errors across observations in space or network structure and across time periods. As a proof of concept, we conduct Monte Carlo simulations in a geospatial setting based on U.S. metropolitan areas. Tests based on our estimator of the VCV asymptotically correctly reject the null hypothesis, whereas conventional inference methods, e.g., those without clusters or with clusters based on administrative units, reject the null hypothesis too often. We

also provide simulations in a network setting based on the IDEAS structure of coauthorship and real-life data on scientific performance. The Monte Carlo results again show that our estimator yields inference at the correct significance level even in moderately sized samples and that it dominates other commonly used approaches to inference in networks. We provide guidance to the applied researcher with respect to (i) whether or not to include potentially correlated regressors and (ii) the choice of cluster bandwidth. Finally, we provide a companion statistical package (acreg) enabling users to adjust the OLS and 2SLS coefficients standard errors to account for arbitrary dependence."

- **Słoczyński (2020), "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," REStat**
  "Applied work often studies the effect of a binary variable ("treatment") using linear models with additive effects. I study the interpretation of the OLS estimands in such models when treatment effects are heterogeneous. I show that the treatment coefficient is a convex combination of two parameters, which under certain conditions can be interpreted as the average treatment effects on the treated and untreated. The weights on these parameters are inversely related to the proportion of observations in each group. Reliance on these implicit weights can have serious consequences for applied work, as I illustrate with two well-known applications. I develop simple diagnostic tools that empirical researchers can use to avoid potential biases. Software for implementing these methods is available in R and Stata. In an important special case, my diagnostics only require the knowledge of the proportion of treated units."

- **Arkhangelsky, Imbens, Lei, and Luo (2021), "Double-Robust Two-Way-Fixed-Effects Regression For Panel Data," WP**
  "We propose a new estimator for the average causal effects of a binary treatment with panel data in settings with general treatment patterns. Our approach augments the two-way-fixed-effects specification with the unit-specific weights that arise from a model for the assignment mechanism. We show how to construct these weights in various settings, including situations where units opt into the treatment sequentially. The resulting estimator converges to an average (over units and time) treatment effect under the correct specification of the assignment model. We show that our estimator is more robust than the conventional two-way estimator: it remains consistent if either the assignment mechanism or the two-way regression model is correctly specified and performs better than the two-way-fixed-effect estimator if both are locally misspecified. This strong double robustness property quantifies the benefits from modeling the assignment process and motivates using our estimator in practice."

- **Goldsmith-Pinkham, Hull, and Kolesár (2021), "On Estimating Multiple Treatment Effects with Regression," WP**
  "We study the causal interpretation of regressions on multiple dependent treatments and flexible controls. Such regressions are often used to analyze randomized control trials with multiple intervention arms, and to estimate institutional quality (e.g. teacher value-added) with observational data. We show that, unlike with a single binary treatment, these regressions do not generally estimate convex averages of causal effects – even when the treatments are conditionally randomly assigned and the controls fully address omitted variables bias. We discuss different solutions to this issue, and propose as a solution a new class of efficient estimators of weighted average treatment effects."

- **Abadie, Athey, Imbens, and Wooldridge (2022), "When Should You Adjust Standard Errors for Clustering?," QJE**
  "Clustered standard errors, with clusters defined by factors such as geography, are widespread in empirical research in economics and many other disciplines. Formally, clustered standard errors adjust for the correlations induced by sampling the outcome variable from a data-generating process with unobserved cluster-level components. However, the standard econometric framework for clustering leaves important questions unanswered: (i) Why do we adjust standard errors for clustering in some ways but not others, e.g., by state but not by gender, and in observational studies, but not in completely randomized experiments? (ii) Why is conventional clustering an "all-or-nothing" adjustment, while within-cluster correlations can be strong or extremely weak? (iii) In what settings does the choice of whether and how to cluster make a difference? We address these and other questions using a novel framework for clustered inference on average treatment effects. In addition to the common sampling component, the new framework incorporates a design component that accounts for the variability induced on the estimator by the treatment assignment mechanism. We show that, when the number of clusters in the sample is a nonnegligible fraction of the number of clusters in the population, conventional cluster standard errors can be severely inflated, and propose new variance estimators that correct for this bias."

- **MacKinnon, Nielsen, and Webb (2022a), "Cluster-Robust Inference: A Guide to Empirical Practice," JE**
  "Methods for cluster-robust inference are routinely used in economics and many other disciplines. However, it is only recently that theoretical foundations for the use of these methods in many empirically relevant situations have been developed. In this paper, we use these theoretical results to provide a guide to empirical practice. We do not

attempt to present a comprehensive survey of the (very large) literature. Instead, we bridge theory and practice by providing a thorough guide on what to do and why, based on recently available econometric theory and simulation evidence. To practice what we preach, we include an empirical analysis of the effects of the minimum wage on labor supply of teenagers using individual data."

- **MacKinnon, Nielsen, and Webb (2022b), "Testing for the Appropriate Level of Clustering in Linear Regression Models," WP**
  "The overwhelming majority of empirical research that uses cluster-robust inference assumes that the clustering structure is known, even though there are often several possible ways in which a dataset could be clustered. We propose two tests for the correct level of clustering in regression models. One test focuses on inference about a single coefficient, and the other on inference about two or more coefficients. We provide both asymptotic and wild bootstrap implementations. The proposed tests work for a null hypothesis of either no clustering or "fine" clustering against alternatives of "coarser" clustering. We also propose a sequential testing procedure to determine the appropriate level of clustering. Simulations suggest that the bootstrap tests perform very well under the null hypothesis and can have excellent power. An empirical example suggests that using the tests leads to sensible inferences."

- **Müller and Watson (2022), "Spatial Correlation Robust Inference," ECMA**
  "We propose a method for constructing confidence intervals that account for many forms of spatial correlation. The interval has the familiar "estimator plus and minus a standard error times a critical value" form, but we propose new methods for construct- ing the standard error and the critical value. The standard error is constructed using population principal components from a given "worst-case" spatial correlation model. The critical value is chosen to ensure coverage in a benchmark parametric model for the spatial correlations. The method is shown to control coverage in finite sample Gaussian settings in a restricted but nonparametric class of models and in large samples whenever the spatial correlation is weak, i.e., with average pairwise correlations that vanish as the sample size gets large. We also provide results on the efficiency of the method."

- **MacKinnon (2023), "Using Large Samples in Econometrics," JE**
  "As I demonstrate using evidence from a journal data repository that I manage, the datasets used in empirical work are getting larger. When we use very large datasets, it can be dangerous to rely on standard methods for statistical inference. In addition, we need to worry about computational issues. We must be careful in our choice of statistical methods and the algorithms used to implement them."

- **Goldsmith-Pinkham, Hull, and Kolesár (2024), "Contamination Bias in Linear Regressions," NBER WP**
  "We study regressions with multiple treatments and a set of controls that is flexible enough to purge omitted variable bias. We show that these regressions generally fail to estimate convex averages of heterogeneous treatment effects – instead, estimates of each treatments effect are contaminated by non-convex averages of the effects of other treatments. We discuss three estimation approaches that avoid such contamination bias, including the targeting of easiest-to-estimate weighted average effects. A re-analysis of nine empirical applications finds economically and statistically meaningful contamination bias in observational studies; contamination bias in experimental studies is more limited due to idiosyncratic effect heterogeneity."

# 2 RCT

- **Muralidharan, Romero, and Wüthrich (2019), "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments," NBER WP**
  "Cross-cutting or factorial designs are widely used in field experiments. Standard t-tests using the fully-saturated long model provide valid inference on the main treatment effects and all interactions. However, t-tests using a "short"model (without interactions) yield greater power for inference on the main treatment effects if the interactions are zero. We show that the assumption of zero interactions is problematic and leads to a significant increase in incorrect inference regarding the main treatment effects relative to a "business as usual" counterfactual. Further, we show that pre-testing the interactions and ignoring them if they are not significant also leads to incorrect inference (due to the implied model selection). We examine econometric approaches to improve power relative to the long model while controlling size for all values of the interaction. Modest local power improvements are possible, but come at the cost of lower power for most values of the interaction. For the design of new experiments, an alternative is to leave the interaction cells empty. This design-based approach yields global power improvements while controlling size and we recommend it for policy experiments where a "business as usual" counterfactual is especially important."

- **Young (2019), "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," QJE**
  "I follow R. A. Fisher's, *The Design of Experiments* (1935), using randomization statistical inference to test the null hypothesis of no treatment effects in a comprehensive

sample of 53 experimental papers drawn from the journals of the American Economic Association. In the average paper, randomization tests of the significance of individual treatment effects find 13% to 22% fewer significant results than are found using authors methods. In joint tests of multiple treatment effects appearing together in tables, randomization tests yield 33% to 49% fewer statistically significant results than conventional tests. Bootstrap and jackknife methods support and confirm the randomization results."

- **Burlig, Preonas, and Woerman (2020), "Panel Data and Experimental Design," JDE**
  "How should researchers design panel data experiments? We analytically derive the variance of panel estimators, informing power calculations in panel data settings. We generalize Frison and Pocock (1992) to fully arbitrary error structures, thereby extending McKenzie (2012) to allow for non-constant serial correlation. Using Monte Carlo simulations and real-world panel data, we demonstrate that failing to account for arbitrary serial correlation *ex ante* yields experiments that are incorrectly powered under proper inference. By contrast, our serial-correlation-robust power calculations achieve correctly powered experiments in both simulated and real data. We discuss the implications of these results, and introduce a new software package to facilitate proper power calculations in practice."

- **Deeb and de Chaisemartin (2020), "Clustering and External Validity in Randomized Controlled Trials," WP**
  "The randomization inference literature studying randomized controlled trials (RCTs) assumes that units' potential outcomes are deterministic. This assumption is unlikely to hold, as stochastic shocks may take place during the experiment. In this paper, we consider the case of an RCT with individual-level treatment assignment, and we allow for individual-level and cluster-level (e.g. village-level) shocks to affect the potential outcomes. We show that one can draw inference on two estimands: the ATE conditional on the realizations of the cluster-level shocks, using heteroskedasticity-robust standard errors; the ATE netted out of those shocks, using cluster-robust standard errors. By clustering, researchers can test if the treatment would still have had an effect, had the stochastic shocks that occurred during the experiment been different. Then, the decision to cluster or not depends on the level of external validity one would like to achieve."

- **Athey, Bickel, Chen, Imbens, and Pollmann (2021), "Semiparametric Estimation of Treatment Effects in Randomized Experiments," NBER WP**
  "We develop new semiparametric methods for estimating treatment effects. We focus on

a setting where the outcome distributions may be thick tailed, where treatment effects are small, where sample sizes are large and where assignment is completely random. This setting is of particular interest in recent experimentation in tech companies. We propose using parametric models for the treatment effects, as opposed to parametric models for the full outcome distributions. This leads to semiparametric models for the outcome distributions. We derive the semiparametric efficiency bound for this setting, and propose efficient estimators. In the case with a constant treatment effect one of the proposed estimators has an interesting interpretation as a weighted average of quantile treatment effects, with the weights proportional to (minus) the second derivative of the log of the density of the potential outcomes. Our analysis also results in an extension of Huber's model and trimmed mean to include asymmetry and a simplified condition on linear combinations of order statistics, which may be of independent interest."

- **Gabriel, Sjölander, and Sachs (2021), "Nonparametric Bounds for Causal Effects in Imperfect Randomized Experiments," JASA**
  "Nonignorable missingness and noncompliance can occur even in well-designed randomized experiments, making the intervention effect that the experiment was designed to estimate nonidentifiable. Nonparametric causal bounds provide a way to narrow the range of possible values for a nonidentifiable causal effect with minimal assumptions. We derive novel bounds for the causal risk difference for a binary outcome and intervention in randomized experiments with nonignorable missingness that is caused by a variety of mechanisms, with both perfect and imperfect compliance. We show that the so-called worst case imputation, whereby all missing subjects on the intervention arm are assumed to have events and all missing subjects on the control or placebo arm are assumed to be event-free, can be too pessimistic in blinded studies with perfect compliance, and is not bounding the correct estimand with imperfect compliance. We illustrate the use of the proposed bounds in our motivating data example of peanut consumption on the development of peanut allergies in infants. We find that, even accounting for potentially nonignorable missingness and noncompliance, our derived bounds confirm that regular exposure to peanuts reduces the risk of development of peanut allergies, making the results of this study much more compelling."

- **Zhao and Ding (2021), "Covariate-Adjusted Fisher Randomization Tests for the Average Treatment Effect," JE**
  "Fishers randomization test (FRT) delivers exact -values under the strong null hypothesis of no treatment effect on any units whatsoever and allows for flexible covariate adjustment to improve the power. Of interest is whether the resulting covariate-adjusted

procedure could also be valid for testing the weak null hypothesis of zero average treatment effect. To this end, we evaluate two general strategies for conducting covariate adjustment in FRTs: the pseudo-outcome strategy that uses the residuals from an outcome model with only the covariates as the pseudo, covariate-adjusted outcomes to form the test statistic, and the model-output strategy that directly uses the output from an outcome model with both the treatment and covariates as the covariate-adjusted test statistic. Based on theory and simulation, we recommend using the ordinary least squares (OLS) fit of the observed outcome on the treatment, centered covariates, and their interactions for covariate adjustment, and conducting FRT with the robust t-value of the treatment as the test statistic. The resulting FRT is finite-sample exact for testing the strong null hypothesis, asymptotically valid for testing the weak null hypothesis, and more powerful than the unadjusted counterpart under alternatives, all irrespective of whether the linear model is correctly specified or not. We start with complete randomization, and then extend the theory to cluster randomization, stratified randomization, and rerandomization, respectively, giving a recommendation for the test procedure and test statistic under each design. Our theory is design-based, also known as randomization-based, in which we condition on the potential outcomes but average over the random treatment assignment."

- **Vazquez-Bare (2022), "Identification and estimation of spillover effects in randomized experiments," JE**

  "I study identification, estimation and inference for spillover effects in experiments where units outcomes may depend on the treatment assignments of other units within a group. I show that the commonly-used reduced-form linear-in-means regression identifies a weighted sum of spillover effects with some negative weights, and that the difference in means between treated and controls identifies a combination of direct and spillover effects entering with different signs. I propose nonparametric estimators for average direct and spillover effects that overcome these issues and are consistent and asymptotically normal under a precise relationship between the number of parameters of interest, the total sample size and the treatment assignment mechanism. These findings are illustrated using data from a conditional cash transfer program and with simulations. The empirical results reveal the potential pitfalls of failing to flexibly account for spillover effects in policy evaluation: the estimated difference in means and the reduced-form linear-in-means coefficients are all close to zero and statistically insignificant, whereas the nonparametric estimators I propose reveal large, nonlinear and significant spillover effects."

# 3 Diff-in-Diff & Event Studies[1]

- **Brewer, Crossley, and Joyce (2017), "Inference with Difference-in-Differences Revisited," JEM**
  "A growing literature on inference in difference-in-differences (DiD) designs has been pessimistic about obtaining hypothesis tests of the correct size, particularly with few groups. We provide Monte Carlo evidence for four points: (i) it is possible to obtain tests of the correct size even with few groups, and in many settings very straightforward methods will achieve this; (ii) the main problem in DiD designs with grouped errors is instead low power to detect real effects; (iii) feasible GLS estimation combined with robust inference can increase power considerably whilst maintaining correct test size again, even with few groups, and (iv) using OLS with robust inference can lead to a perverse relationship between power and panel length."

- **Athey and Imbens (2018), "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," NBER WP**
  "In this paper we study estimation of and inference for average treatment effects in a setting with panel data. We focus on the setting where units, e.g., individuals, firms, or states, adopt the policy or treatment of interest at a particular point in time, and then remain exposed to this treatment at all times afterwards. We take a design perspective where we investigate the properties of estimators and procedures given assumptions on the assignment process. We show that under random assignment of the adoption date the standard Difference-In-Differences estimator is is an unbiased estimator of a particular weighted average causal effect. We characterize the properties of this estimand, and show that the standard variance estimator is conservative."

- **de Chaisemartin and d'Haultfoeuille (2018), "Fuzzy Differences-in-Differences," REStud**
  "Difference-in-differences (DID) is a method to evaluate the effect of a treatment. In its basic version, a control group is untreated at two dates, whereas a treatment group becomes fully treated at the second date. However, in many applications of the DID method, the treatment rate only increases more in the treatment group. In such fuzzy designs, a popular estimator of the treatment effect is the DID of the outcome divided by the DID of the treatment. We show that this ratio identifies a local average treatment effect only if the effect of the treatment is stable over time, and if the effect of the

[1] See Roth, Sant'Anna, Bilinski, and Poe (2022) for a review of the recent literature on that topic and this webpage by Asjad Naqvi for a summary of all the commands.

treatment is the same in the treatment and in the control group. We then propose two alternative estimands that do not rely on any assumption on treatment effects, and that can be used when the treatment rate does not change over time in the control group. We prove that the corresponding estimators are asymptotically normal. Finally, we use our results to reassess the returns to schooling in Indonesia."

- **Cengiz, Dube, Lindner, and Zipperer (2019), "The Effect of Minimum Wages on Low-Wage Jobs,"[2] QJE**
  "We estimate the effect of minimum wages on low-wage jobs using 138 prominent state-level minimum wage changes between 1979 and 2016 in the United States using a difference-in-differences approach. We first estimate the effect of the minimum wage increase on employment changes by wage bins throughout the hourly wage distribution. We then focus on the bottom part of the wage distribution and compare the number of excess jobs paying at or slightly above the new minimum wage to the missing jobs paying below it to infer the employment effect. We find that the overall number of low-wage jobs remained essentially unchanged over the five years following the increase. At the same time, the direct effect of the minimum wage on average earnings was amplified by modest wage spillovers at the bottom of the wage distribution. Our estimates by detailed demographic groups show that the lack of job loss is not explained by labor-labor substitution at the bottom of the wage distribution. We also find no evidence of disemployment when we consider higher levels of minimum wages. However, we do find some evidence of reduced employment in tradeable sectors. We also show how decomposing the overall employment effect by wage bins allows a transparent way of assessing the plausibility of estimates."

- **Ferman and Pinto (2019), "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," REStat**
  "We derive an inference method that works in differences-in-differences settings with few treated and many control groups in the presence of heteroskedasticity. As a leading example, we provide theoretical justification and empirical evidence that heteroskedasticity generated by variation in group sizes can invalidate existing inference methods, even in data sets with a large number of observations per group. In contrast, our inference method remains valid in this case. Our test can also be combined with feasible generalized least squares, providing a safeguard against misspecification of the serial correlation."

[2] Even though its title makes it sound like it is irrelevant, this paper has been added because it describes another method to deal with heterogeneous treatment effects in event-study designs, by using stacked diff-in-diff by event – see Online Appendix G of that paper for more detail.

- **Freyaldenhoven, Hansen, and Shapiro (2019), "Pre-event Trends in the Panel Event-Study Design," AER**
  "We consider a linear panel event-study design in which unobserved confounds may be related both to the outcome and to the policy variable of interest. We provide sufficient conditions to identify the causal effect of the policy by exploiting covariates related to the policy only through the confounds. Our model implies a set of moment equations that are linear in parameters. The effect of the policy can be estimated by 2SLS, and causal inference is valid even when endogeneity leads to pre-event trends ("pre-trends") in the outcome. Alternative approaches perform poorly in our simulations."

- **de Chaisemartin and d'Haultfoeuille (2020), "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," AER**
  "Linear regressions with period and group fixed effects are widely used to estimate treatment effects. We show that they estimate weighted sums of the average treatment effects (ATE) in each group and period, with weights that may be negative. Due to the negative weights, the linear regression coefficient may for instance be negative while all the ATEs are positive. We propose another estimator that solves this issue. In the two applications we revisit, it is significantly different from the linear regression estimator."

- **Marcus and Sant'Anna (2020), "The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics," JAERE**
  "Difference-in-Differences (DID) research designs usually rely on variation of treatment timing such that, after making an appropriate parallel trends assumption, one can identify, estimate, and make inference about causal effects. In practice, however, different DID procedures rely on different parallel trends assumptions (PTA), and recover different causal parameters. In this paper, we focus on staggered DID (also referred as event-studies) and discuss the role played by the PTA in terms of identification and estimation of causal parameters. We document a "robustness vs. "efficiency trade-off in terms of the strength of the underlying PTA, and argue that practitioners should be explicit about these trade-offs whenever using DID procedures. We propose new DID estimators that reflect these trade-offs and derived their large sample properties. We illustrate the practical relevance of these results by assessing whether the transition from federal to state management of the Clean Water Act affects compliance rates."

- **Sant'Anna and Zhao (2020), "Doubly Robust Difference-in-Differences Estimators," JE**
  "This article proposes doubly robust estimators for the average treatment effect on the

treated (ATT) in difference-in-differences (DID) research designs. In contrast to alternative DID estimators, the proposed estimators are consistent if either (but not necessarily both) a propensity score or outcome regression working models are correctly specified. We also derive the semiparametric efficiency bound for the ATT in DID designs when either panel or repeated cross-section data are available, and show that our proposed estimators attain the semiparametric efficiency bound when the working models are correctly specified. Furthermore, we quantify the potential efficiency gains of having access to panel data instead of repeated cross-section data. Finally, by paying particular attention to the estimation method used to estimate the nuisance parameters, we show that one can sometimes construct doubly robust DID estimators for the ATT that are also doubly robust for inference. Simulation studies and an empirical application illustrate the desirable finite-sample performance of the proposed estimators. Open-source software for implementing the proposed policy evaluation tools is available."

- **Schmidheiny and Siegloch (2020), "On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization," WP**
  "We discuss properties and pitfalls of panel-data event study designs. We derive three main results. First, assuming constant treatment effects before and/or after some event time, also known as binning, is a natural restriction imposed on theoretically infinite effect windows. Binning identifies dynamic treatment effects in the absence of never-treated units and is particularly suitable in case of multiple events. Second, event study designs with binned endpoints and distributed-lag models are numerically identical leading to the same parameter estimates after correct reparametrization. Third, classic dummy variable event study designs can be generalized to models that account for multiple events of different sign and intensity of the treatment, which are common in public and labor economics. We demonstrate the practical relevance of our methodological points in an application studying the effects of unemployment benefit duration on job search effort."

- **Abraham and Sun (2021), "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," JE**
  "To estimate the dynamic effects of an absorbing treatment, researchers often use two-way fixed effects regressions that include leads and lags of the treatment. We show that in settings with variation in treatment timing across units, the coefficient on a given lead or lag can be contaminated by effects from other periods, and apparent pretrends can arise solely from treatment effects heterogeneity. We propose an alternative estimator that is free of contamination, and illustrate the relative shortcomings of two-way fixed

effects regressions with leads and lags through an empirical application."

- **Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021), "Synthetic Difference in Differences," AER**
  "We present a new estimator for causal effects with panel data that builds on insights behind the widely used difference in differences and synthetic control methods. Relative to these methods, we find, both theoretically and empirically, that the proposed synthetic difference in differences estimator has desirable robustness properties, and that it performs well in settings where the conventional estimators are commonly used in practice. We study the asymptotic behavior of the estimator when the systematic part of the outcome model includes latent unit factors interacted with latent time factors, and we present conditions for consistency and asymptotic normality."

- **Baker, Larcker, and Wang (2021), "How Much Should We Trust Staggered Differences-In-Differences Estimates?," WP**
  "Difference-in-differences analysis with staggered treatment timing is frequently used to assess the impact of policy changes on corporate outcomes in academic research. However, recent advances in econometric theory show that such designs are likely to be biased in the presence of treatment effect heterogeneity. Given the pronounced use of staggered treatment designs in accounting and applied corporate finance research, this finding potentially impacts a large swath of prior findings in these fields. We survey the nascent literature and document how and when such bias arises from treatment effect heterogeneity. We then apply recently proposed methods to a set of prior published results. We find that correcting for the bias induced by the staggered nature of policy adoption frequently impacts the estimated effect from standard difference-in-difference studies. In many cases, the reported effects in prior research become indistinguishable from zero."

- **Butts (2021), "Difference-in-Differences Estimation with Spatial Spillovers," WP**
  "Empirical work often uses treatment assigned following geographic boundaries. When the effects of treatment cross over borders, classical difference-in-differences estimation produces biased estimates for the average treatment effect. In this paper, I introduce a potential outcomes framework to model spillover effects and decompose the estimate's bias in two parts: (1) the control group no longer identifies the counterfactual trend because their outcomes are affected by treatment and (2) changes in treated units' outcomes reflect the effect of their own treatment status and the effect from the treatment status of "close" units. I propose estimation strategies that can remove both

sources of bias and semi-parametrically estimate the spillover effects themselves. I extend Callaway and Sant'Anna (2020) to allow for event-study estimates that control for spillovers. To highlight the importance of spillover effects, I revisit analyses of three place-based interventions."

- **Callaway and Sant'Anna (2021), "Difference-in-Differences with Multiple Time Periods," JE**
  "In this article, we consider identification, estimation, and inference procedures for treatment effect parameters using Difference-in-Differences (DiD) with (i) multiple time periods, (ii) variation in treatment timing, and (iii) when the parallel trends assumption holds potentially only after conditioning on observed covariates. We show that a family of causal effect parameters are identified in staggered DiD setups, even if differences in observed characteristics create non-parallel outcome dynamics between groups. Our identification results allow one to use outcome regression, inverse probability weighting, or doubly-robust estimands. We also propose different aggregation schemes that can be used to highlight treatment effect heterogeneity across different dimensions as well as to summarize the overall effect of participating in the treatment. We establish the asymptotic properties of the proposed estimators and prove the validity of a computationally convenient bootstrap procedure to conduct asymptotically valid simultaneous (instead of pointwise) inference. Finally, we illustrate the relevance of our proposed tools by analyzing the effect of the minimum wage on teen employment from 2001–2007. Open-source software is available for implementing the proposed methods."

- **Freyaldenhoven, Hansen, Pérez, and Shapiro (2021), "Visualization, Identification, and Estimation in the Linear Panel Event-Study Design," NBER WP**
  "Linear panel models, and the "event-study plots" that often accompany them, are popular tools for learning about policy effects. We discuss the construction of event-study plots and suggest ways to make them more informative. We examine the economic content of different possible identifying assumptions. We explore the performance of the corresponding estimators in simulations, highlighting that a given estimator can perform well or poorly depending on the economic environment. An accompanying Stata package, -xtevent-, facilitates adoption of our suggestions."

- **Gardner (2021), "Two-Stage Differences-in-Differences," WP**
  "A recent literature has shown that when adoption of a treatment is staggered and average treatment effects vary across groups and over time, difference-in-differences regression does not identify an easily interpretable measure of the typical effect of the treatment. In this paper, I extend this literature in two ways. First, I provide some

simple underlying intuition for why difference-in-differences regression does not identify a group $\times$ period average treatment effect. Second, I propose an alternative two-stage estimation framework, motivated by this intuition. In this framework, group and period effects are identified in a first stage from the sample of untreated observations, and average treatment effects are identified in a second stage by comparing treated and untreated outcomes, after removing these group and period effects. The two-stage approach is robust to treatment-effect heterogeneity under staggered adoption, and can be used to identify a host of different average treatment effect measures. It is also simple, intuitive, and easy to implement. I establish the theoretical properties of the two-stage approach and demonstrate its effectiveness and applicability using Monte-Carlo evidence and an example from the literature."

- Goodman-Bacon (2021), **"Difference-in-Differences with Variation in Treatment Timing," JE**
  "The canonical difference-in-differences (DD) estimator contains two time periods, "pre" and "post", and two groups, "treatment" and "control". Most DD applications, however, exploit variation across groups of units that receive treatment at different times. This paper shows that the two-way fixed effects estimator equals a weighted average of all possible two-group/two-period DD estimators in the data. A causal interpretation of two-way fixed effects DD estimates requires both a parallel trends assumption and treatment effects that are constant over time. I show how to decompose the difference between two specifications, and provide a new analysis of models that include time-varying controls."

- Wooldridge (2021), **"Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," WP**
  "I establish the equivalence between the two-way fixed effects (TWFE) estimator and an estimator obtained from a pooled ordinary least squares regression that includes unit-specific time averages and time-period specific cross-sectional averages, which I call the two-way Mundlak (TWM) regression. This equivalence furthers our understanding of the anatomy of TWFE, and has several applications. The equivalence between TWFE and TWM implies that various estimators used for intervention analysis – with a common entry time into treatment or staggered entry, with or without covariates – can be computed using TWFE or pooled OLS regressions that control for time-constant treatment intensities, covariates, and interactions between them. The approach allows considerable heterogeneity in treatment effects across treatment intensity, calendar time, and covariates. The equivalence implies that standard strategies for heterogeneous

trends are available to relax the common trends assumption. Further, the two-way Mundlak regression is easily adapted to nonlinear models such as exponential models and logit and probit models."

- **Caetano, Callaway, Stroud, and Sant'Anna Rodrigues (2022), "Difference in Differences with Time-Varying Covariates," WP**
  "This paper considers identification and estimation of causal effect parameters from participating in a binary treatment in a difference in differences (DID) setup when the parallel trends assumption holds after conditioning on observed covariates. Relative to existing work in the econometrics literature, we consider the case where the value of covariates can change over time and, potentially, where participating in the treatment can affect the covariates themselves. We propose new empirical strategies in both cases. We also consider two-way fixed effects (TWFE) regressions that include time-varying regressors, which is the most common way that DID identification strategies are implemented under conditional parallel trends. We show that, even in the case with only two time periods, these TWFE regressions are not generally robust to (i) time-varying covariates being affected by the treatment, (ii) treatment effects and/or paths of untreated potential outcomes depending on the level of time-varying covariates in addition to only the change in the covariates over time, (iii) treatment effects and/or paths of untreated potential outcomes depending on time-invariant covariates, (iv) treatment effect heterogeneity with respect to observed covariates, and (v) violations of strong functional form assumptions, both for outcomes over time and the propensity score, that are unlikely to be plausible in most DID applications. Thus, TWFE regressions can deliver misleading estimates of causal effect parameters in a number of empirically relevant cases. We propose both doubly robust estimands and regression adjustment/imputation strategies that are robust to these issues while not being substantially more challenging to implement."

- **Roth (2022), "Pre-Test with Caution: Event-Study Estimates after Testing for Parallel Trends," AER I**
  "This paper discusses two important limitations of the common practice of testing for pre-existing differences in trends ("pre-trends") when using difference-in-differences and related methods. First, conventional pre-trends tests may have low power. Second, conditioning the analysis on the result of a pre-test can distort estimation and inference, potentially exacerbating the bias of point estimates and undercoverage of confidence intervals. I analyze these issues both in theory and in simulations calibrated to a survey of recent papers in leading economics journals, which suggest that these limitations are

important in practice. I conclude with practical recommendations for mitigating these issues."

- **Roth and Sant'Anna (2022), "When Is Parallel Trends Sensitive to Functional Form?," ECMA**
  "This paper assesses when the validity of difference-in-differences depends on functional form. We provide a novel characterization: the parallel trends assumption holds under all strictly monotonic transformations of the outcome if and only if a stronger "parallel trends"-type condition holds for the cumulative distribution function of untreated potential outcomes. This condition for parallel trends to be insensitive to functional form is satisfied if and essentially only if the population can be partitioned into a subgroup for which treatment is effectively randomly assigned and a remaining subgroup for which the distribution of untreated potential outcomes is stable over time. These conditions have testable implications, and we introduce falsification tests for the null that parallel trends is insensitive to functional form."

- **Roth and Sant'Anna (2022), "Efficient Estimation for Staggered Rollout Designs," WP**
  "This paper studies efficient estimation of causal effects when treatment is (quasi-) randomly rolled out to units at different points in time. We solve for the most efficient estimator in a class of estimators that nests two-way fixed effects models and other popular generalized difference-in-differences methods. A feasible plug-in version of the efficient estimator is asymptotically unbiased with efficiency (weakly) dominating that of existing approaches. We provide both t-based and permutation-test based methods for inference. We illustrate the performance of the plug-in efficient estimator in simulations and in an application to Wood et al (2020)'s study of the staggered rollout of a procedural justice training program for police officers. We find that confidence intervals based on the plug-in efficient estimator have good coverage and can be as much as eight times shorter than confidence intervals based on existing state-of-the-art methods. As an empirical contribution of independent interest, our application provides the most precise estimates to date on the effectiveness of procedural justice training programs for police officers."

- **Roth, Sant'Anna, Bilinski, and Poe (2022), "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," JE**
  "This paper synthesizes recent advances in the econometrics of difference-in-differences (DiD) and provides concrete recommendations for practitioners. We begin by articulating a simple set of "canonical" assumptions under which the econometrics of DiD are

well-understood. We then argue that recent advances in DiD methods can be broadly classified as relaxing some components of the canonical DiD setup, with a focus on (i) multiple periods and variation in treatment timing, (ii) potential violations of parallel trends, or (iii) alternative frameworks for inference. Our discussion highlights the different ways that the DiD literature has advanced beyond the canonical model, and helps to clarify when each of the papers will be relevant for empirical work. We conclude by discussing some promising areas for future research. ”

- **Sun and Shapiro (2022), “Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure,” NBER WP**
  “Linear panel models featuring unit and time fixed effects appear in many areas of empirical economics. An active literature studies the interpretation of the ordinary least squares estimator of the model, commonly called the two-way fixed effects (TWFE) estimator, in the presence of unmodeled coefficient heterogeneity. We illustrate some implications for the case where the research design takes advantage of variation across units (say, US states) in exposure to some treatment (say, a policy change). In this case, the TWFE can fail to estimate the average (or even a weighted average) of the units’ coefficients. Under some conditions, there exists no estimator that is guaranteed to estimate even a weighted average. Building on the literature, we note that when there is a unit totally unaffected by treatment, it is possible to estimate an average effect by replacing the TWFE with an average of difference-in-differences estimators.”

- **Arkhangelsky and Imbens (2023), “Causal Models for Longitudinal and Panel Data: A Survey,” WP**
  “This survey discusses the recent causal panel data literature. This recent literature has focused on credibly estimating causal effects of binary interventions in settings with longitudinal data, with an emphasis on practical advice for empirical researchers. It pays particular attention to heterogeneity in the causal effects, often in situations where few units are treated. The literature has extended earlier work on difference-in-differences or two-way-fixed-effect estimators and more generally incorporated factor models or interactive fixed effects. It has also developed novel methods using synthetic control approaches.”

- **Dube, Girardi, Jordà, and Taylor (2023), “A Local Projections Approach to Difference-in-Differences Event Studies,” NBER WP**
  “We propose a local projection (LP) based difference-in-differences approach that subsumes many of the recent solutions proposed in the literature to address possible biases arising from negative weighting. We combine LPs with a flexible clean control condition

to define appropriate sets of treated and control units. Our proposed LP-DiD estimator can be implemented with various weighting and normalization schemes for different target estimands, accommodates controls for pre-treatment values of the outcome and of other time-varying covariates, and is simple and fast to implement. Simulations and two empirical applications demonstrate that the LP-DiD estimator performs well in common applied settings."

- **Faletto (2023), "Fused Extended Two-Way Fixed Effects for Difference-in-Differences with Staggered Adoptions," WP**
  "We study the role of selection into treatment in difference-in-differences (DiD) designs. We derive necessary and sufficient conditions for parallel trends assumptions under general classes of selection mechanisms. These conditions characterize the empirical content of parallel trends. For settings where the necessary conditions are questionable, we propose tools for selection-based sensitivity analysis. We also provide templates for justifying DiD in applications with and without covariates. A reanalysis of the causal effect of NSW training programs demonstrates the usefulness of our selection-based approach to sensitivity analysis."

- **Ghanem, Sant'Anna, and Wüthrich (2023), "Selection and Parallel Trends," WP**
  "We study the role of selection into treatment in difference-in-differences (DiD) designs. We derive necessary and sufficient conditions for parallel trends assumptions under general classes of selection mechanisms. These conditions characterize the empirical content of parallel trends. For settings where the necessary conditions are questionable, we propose tools for selection-based sensitivity analysis. We also provide templates for justifying DiD in applications with and without covariates. A reanalysis of the causal effect of NSW training programs demonstrates the usefulness of our selection-based approach to sensitivity analysis."

- **Pichetti and Pinto (2023), "Marginal Treatment Effects in Difference-in-Differences," WP**
  "Difference-in-Differences (DiD) is a popular method used to evaluate the effect of a treatment that exploits variation in treatment status that comes from the exposure to a shock, usually in the form of a policy change. When there is imperfect compliance towards the shock, the usual DiD estimand fails to recover relevant causal parameters. This article presents an identification strategy in DiD settings with imperfect compliance that identifies Marginal treatment effects (MTE). We show how to combine and modify standard instrumental variables (IV) and DiD assumptions to identify treatment effects in DiD settings where individuals enter into treatment with at least partial knowledge

of their unobservable gains. We propose two estimators for the MTE that are consistent under different assumptions regarding the functional form of potential outcomes and prove their asymptotically normality. Furthermore, we derive an estimator for the local average treatment effect (LATE) that is robust to misspecification of the MTE model. We assert the desirable finite-sample properties through simulation studies of a linear MTE model. Finally, we use our results to investigate heterogeneity on the returns to primary education attendance in Indonesia."

- **Rambachan and Roth (2023), "A More Credible Approach to Parallel Trends," REStud**
  "This paper proposes tools for robust inference in difference-in-differences and event-study designs where the parallel trends assumption may be violated. Instead of requiring that parallel trends holds exactly, we impose restrictions on how different the post-treatment violations of parallel trends can be from the pre-treatment differences in trends ("pre-trends"). The causal parameter of interest is partially identified under these restrictions. We introduce two approaches that guarantee uniformly valid inference under the imposed restrictions, and we derive novel results showing that they have desirable power properties in our context. We illustrate how economic knowledge can inform the restrictions on the possible violations of parallel trends in two economic applications. We also highlight how our approach can be used to conduct sensitivity analyses showing what causal conclusions can be drawn under various restrictions on the possible violations of the parallel trends assumption."

- **Borusyak, Jaravel, and Spiess (2024), "Revisiting Event Study Designs: Robust and Efficient Estimation," REStud**
  "We develop a framework for difference-in-differences designs with staggered treatment adoption and heterogeneous causal effects. We show that conventional regression-based estimators fail to provide unbiased estimates of relevant estimands absent strong restrictions on treatment-effect homogeneity. We then derive the efficient estimator addressing this challenge, which takes an intuitive imputation form when treatment-effect heterogeneity is unrestricted. We characterize the asymptotic behaviour of the estimator, propose tools for inference, and develop tests for identifying assumptions. Our method applies with time-varying controls, in triple-difference designs, and with certain non-binary treatments. We show the practical relevance of our results in a simulation study and an application. Studying the consumption response to tax rebates in the U.S., we find that the notional marginal propensity to consume is between 8 and 11% in the first quarter–about half as large as benchmark estimates used to calibrate macroeconomic

models–and predominantly occurs in the first month after the rebate."

- **de Chaisemartin and D'Haultfœuille (2024), "Two-way Fixed Effects and Differences-in-Differences Estimators in Heterogeneous Adoption Designs," WP**
  "We consider treatment-effect estimation under a parallel trends assumption, in heterogeneous adoption designs where no unit is treated at period one, and units receive a weakly positive dose at period two. First, we develop a test of the assumption that the treatment effect is mean independent of the treatment, under which the commonly-used two-way-fixed-effects estimator is consistent. When this test is rejected, we propose alter- native, robust estimators. If there are stayers with a period-two treatment equal to 0, the robust estimator is a difference-in-differences (DID) estimator using stayers as the control group. If there are quasi-stayers with a period-two treatment arbitrarily close to zero, the robust estimator is a DID using units with a period-two treatment below a bandwidth as controls. Finally, without stayers or quasi-stayers, we propose non-parametric bounds, and an estimator relying on a parametric specification of treatment-effect heterogeneity. We use our results to revisit Pierce and Schott (2016) and Enikolopov et al. (2011)."

# 4 Standard IV

- **Andrews and Armstrong (2017), "Unbiased Instrumental Variables Estimation under Known First-Stage Sign," QE**
  "We derive meanunbiased estimators for the structural parameter in instrumental variables models with a single endogenous regressor where the sign of one or more first-stage coefficients is known. In the case with a single instrument, there is a unique nonrandomized unbiased estimator based on the reducedform and firststage regression estimates. For cases with multiple instruments we propose a class of unbiased estimators and show that an estimator within this class is efficient when the instruments are strong. We show numerically that unbiasedness does not come at a cost of increased dispersion in models with a single instrument: in this case the unbiased estimator is less dispersed than the twostage least squares estimator. Our finitesample results apply to normal models with known variance for the reducedform errors, and imply analogous results under weakinstrument asymptotics with an unknown error distribution."

- **Choi, Gu, and Shen (2018), "Weak-Instrument Robust Inference for Two-Sample Instrumental Variables Regression," JAE**
  "Instrumental variable (IV) methods for regression are well established. More recently,

methods have been developed for statistical inference when the instruments are weakly correlated with the endogenous regressor, so that estimators are biased and no longer asymptotically normally distributed. This paper extends such inference to the case where two separate samples are used to implement instrumental variables estimation. We also relax the restrictive assumptions of homoskedastic error structure and equal moments of exogenous covariates across two samples commonly employed in the two-sample IV literature for strong IV inference. Monte Carlo experiments show good size properties of the proposed tests regardless of the strength of the instruments. We apply the proposed methods to two seminal empirical studies that adopt the two-sample IV framework."

- **Mogstad and Torgovitsky (2018), "Identification and Extrapolation of Causal Effects with Instrumental Variables," ARE**
  "Instrumental variables (IV) are widely used in economics to address selection on unobservables. Standard IV methods produce estimates of causal effects that are specific to individuals whose behavior can be manipulated by the instrument at hand. In many cases, these individuals are not the same as those who would be induced to treatment by an intervention or policy of interest to the researcher. The average causal effect for the two groups can differ significantly if the effect of the treatment varies systematically with unobserved factors that are correlated with treatment choice. We review the implications of this type of unobserved heterogeneity for the interpretation of standard IV methods and for their relevance to policy evaluation. We argue that making inferences about policy-relevant parameters typically requires extrapolating from the individuals affected by the instrument to the individuals who would be induced to treatment by the policy under consideration. We discuss a variety of alternatives to standard IV methods that can be used to rigorously perform this extrapolation. We show that many of these approaches can be nested as special cases of a general framework that embraces the possibility of partial identification."

- **Andrews, Stock, and Sun (2019), "Weak Instruments in Instrumental Variables Regression: Theory and Practice," ARE**
  "When instruments are weakly correlated with endogenous regressors, conventional methods for instrumental variables (IV) estimation and inference become unreliable. A large literature in econometrics has developed procedures for detecting weak instruments and constructing robust confidence sets, but many of the results in this literature are limited to settings with independent and homoskedastic data, while data encountered in practice frequently violate these assumptions. We review the literature on weak

instruments in linear IV regression with an emphasis on results for nonhomoskedastic (heteroskedastic, serially correlated, or clustered) data. To assess the practical importance of weak instruments, we also report tabulations and simulations based on a survey of papers published in the *American Economic Review* from 2014 to 2018 that use IV. These results suggest that weak instruments remain an important issue for empirical practice, and that there are simple steps that researchers can take to better handle weak instruments in applications."

- **Evdokimov and Kolesár (2019), "Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects," WP**
  "We study inference in an instrumental variables model with heterogeneous treatment effects and possibly many instruments and/or covariates. In this case two-step estimators such as the two-stage least squares (TSLS) or versions of the jackknife instrumental variables (JIV) estimator estimate a particular weighted average of the local average treatment effects. The weights in these estimands depend on the first-stage coefficients, and either the sample or population variability of the covariates and instruments, depending on whether they are treated as fixed (conditioned upon) or random. We give new asymptotic variance formulas for the TSLS and JIV estimators, and propose consistent estimators of these variances. The heterogeneity of the treatment effects generally increases the asymptotic variance. Moreover, when the treatment effects are heterogeneous, the conditional asymptotic variance is smaller than the unconditional one. Our results are also useful when the treatment effects are constant, because they provide the asymptotic distribution and valid standard errors for the estimators that are robust to the presence of many covariates."

- **Choi and Shen (2019), "Two-Sample Instrumental Variables Regression with Potentially Weak Instruments," SJ**
  "We develop a command, weaktsiv, for two-sample instrumental variables regression models with one endogenous regressor and potentially weak instruments. weaktsiv includes the classic two-sample two-stage least-squares estimator whose inference is valid only under the assumption of strong instruments. It also includes statistical tests and confidence sets with correct size and coverage probabilities even when the instruments are weak."

- **Finley (2020), "Testing for Weak-Instrument Bias in Just-Identified 2SLS," WP**
  "We propose a test and confidence procedure to gauge the possible impact of weak instruments in the linear model with one excluded instrument and one endogenous regressor, the model typically used with instrumental variables in applied work. Where

$\hat{\beta}$ is the two-stage least squares estimator of the endogenous regressor's coefficient, $\beta$, we perform inference on worst-case asymptotic values of $P[\beta < \hat{\beta}]$. The deviation of $P[\beta < \hat{\beta}]$ from .5 can be intuitively read as a deviation from median unbiasedness, providing an interpretable bias test for the just-identified model, where the mean bias $E[\hat{\beta} - \beta]$ is undefined. These inference procedures can easily be made robust to error heteroskedasticity and dependence such as clustering and serial correlation."

- **Huntington-Kleina (2020), "Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness," JCI**
  "In Instrumental Variables (IV) estimation, the effect of an instrument on an endogenous variable may vary across the sample. In this case, IV produces a local average treatment effect (LATE), and if monotonicity does not hold, then no effect of interest is identified. In this paper, I calculate the weighted average of treatment effects that is identified under general first-stage effect heterogeneity, which is generally not the average treatment effect among those affected by the instrument. I then describe a simple set of data-driven approaches to modeling variation in the effect of the instrument. These approaches identify a Super-Local Average Treatment Effect (SLATE) that weights treatment effects by the corresponding instrument effect more heavily than LATE. Even when first-stage heterogeneity is poorly modeled, these approaches considerably reduce the impact of small-sample bias compared to standard IV and unbiased weak-instrument IV methods, and can also make results more robust to violations of monotonicity. In application to a published study with a strong instrument, the preferred approach reduces error by about 19% in small ($N \approx 1,000$) subsamples, and by about 13% in larger ($N \approx 33,000$) subsamples."

- **Mogstad, Torgovitsky, and Walters (2020a), "Policy Evaluation with Multiple Instrumental Variables," WP**
  "Marginal treatment effect methods are widely used for causal inference and policy evaluation with instrumental variables. However, they fundamentally rely on the well-known monotonicity (threshold-crossing) condition on treatment choice behavior. Recent research has shown that this condition cannot hold with multiple instruments unless treatment choice is effectively homogeneous. Based on these findings, we develop a new marginal treatment effect framework under a weaker, partial monotonicity condition. The partial monotonicity condition is implied by standard choice theory and allows for rich heterogeneity even in the presence of multiple instruments. The new framework can be viewed as having multiple different choice models for the same observed treatment variable, all of which must be consistent with the data and with each other.

Using this framework, we develop a methodology for partial identification of clearly stated, policy-relevant target parameters while allowing for a wide variety of nonparametric shape restrictions and parametric functional form assumptions. We show how the methodology can be used to combine multiple instruments together to yield more informative empirical conclusions than one would obtain by using each instrument separately. The methodology provides a blueprint for extracting and aggregating information about treatment effects from multiple controlled or natural experiments while still allowing for rich heterogeneity in both treatment effects and choice behavior."

- **Mogstad, Torgovitsky, and Walters (2020b), "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," NBER WP**
  "Empirical researchers often combine multiple instrumental variables (IVs) for a single treatment using two-stage least squares (2SLS). When treatment effects are heterogeneous, a common justification for including multiple IVs is that the 2SLS estimand can be given a causal interpretation as a positively-weighted average of local average treatment effects (LATEs). This justification requires the well-known monotonicity condition. However, we show that with more than one instrument, this condition can only be satisfied if choice behavior is effectively homogenous. Based on this finding, we consider the use of multiple IVs under a weaker, partial monotonicity condition. We characterize empirically verifiable sufficient and necessary conditions for the 2SLS estimand to be a positively-weighted average of LATEs under partial monotonicity. We apply these results to an empirical analysis of the returns to college with multiple instruments. We show that the standard monotonicity condition is at odds with the data. Nevertheless, our empirical checks show that the 2SLS estimate retains a causal interpretation as a positively-weighted average of the effects of college attendance among complier groups."

- **Young (2020), "Consistency Without Inference: Instrumental Variables in Practical Application," WP**
  "I use Monte Carlo simulations, the jackknife and multiple forms of the bootstrap to study a comprehensive sample of 1359 instrumental variables regressions in 31 papers published in the journals of the American Economic Association. Monte Carlo simulations based upon published regressions show that non-iid error processes in highly leveraged regressions, both prominent features of published work, adversely affect the size and power of IV estimates, while increasing the bias of IV relative to OLS. Weak instrument pre-tests based upon F-statistics are found to be largely uninformative of both size and bias. In published papers, statistically significant IV results generally depend

upon only one or two observations or clusters, IV has little power as, despite producing substantively different estimates, it rarely rejects the OLS point estimate or the null that OLS is unbiased, while the statistical significance of excluded instruments is substantially exaggerated."

- **Andresen and Huber (2021), "Instrument-Based Estimation with Binarized Treatments: Issues and Tests for the Exclusion Restriction," EJ**
  "When estimating local average and marginal treatment effects using instrumental variables (IV), multivalued endogenous treatments are frequently converted to binary measures, supposedly to improve interpretability or policy relevance. Such binarization introduces a violation of the IV exclusion if (i) the IV affects the multivalued treatment within support areas below and/or above the threshold and (ii) such IV-induced changes in the multivalued treatment affect the outcome. We discuss assumptions that satisfy the IV exclusion restriction with a binarized treatment and permit identifying the average effect of (i) the binarized treatment and (ii) unit-level increases in the original multivalued treatment among specific compliers. We derive testable implications of these assumptions and propose tests, which we apply to the estimation of the returns to college graduation instrumented by college proximity."

- **Słoczyński (2021), "When Should We (Not) Interpret Linear IV Estimands as LATE?," WP**
  "In this paper I revisit the interpretation of the linear instrumental variables (IV) estimand as a weighted average of conditional local average treatment effects (LATEs). I focus on a practically relevant situation in which additional covariates are required for identification while the reduced-form and first-stage regressions implicitly restrict the effects of the instrument to be homogeneous, and are thus possibly misspecified. I show that the weights on some conditional LATEs are negative and the IV estimand is no longer interpretable as a causal effect under a weaker version of monotonicity, i.e. when there are compliers but no defiers at some covariate values and defiers but no compliers elsewhere. The problem of negative weights disappears in the overidentified specification of Angrist and Imbens (1995) and in an alternative method, termed "reordered IV," that I also develop. Even if all weights are positive, the IV estimand in the just identified specification is not interpretable as the unconditional LATE parameter unless the groups with different values of the instrument are roughly equal sized. I illustrate my findings in an application to causal effects of college education using the college proximity instrument. The benchmark estimates suggest that college attendance yields earnings gains of about 60 log points, which is well outside the range of estimates in the

recent literature. I demonstrate that this result is driven by the existence of defiers and the presence of negative weights. Corrected estimates indicate that attending college causes earnings to be roughly 20% higher."

- **Angrist and Kolesár (2022), "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV," WP**
  "We revisit the finite-sample behavior of just-identified instrumental variables (IV) estimators, arguing that in most microeconometric applications, just-identified IV bias is negligible and the usual inference strategies likely reliable. Three widely-cited applications are used to explain why this is so. We then consider pretesting strategies of the form t1 $>$ c, where t1 is the first-stage t-statistic, and the first-stage sign is given. Although pervasive in empirical practice, pretesting on the first-stage F-statistic exacerbates bias and distorts inference. We show, however, that median bias is both minimized and roughly halved by setting c $=$ 0, that is by screening on the sign of the estimated first stage. This bias reduction is a free lunch: conventional confidence interval coverage is unchanged by screening on the estimated first-stage sign. To the extent that IV analysts sign-screen already, these results strengthen the case for a sanguine view of the finite-sample behavior of just-ID IV."

- **Lee, McCrary, Moreira, and Porter (2022), "Valid t-ratio Inference for IV," AER**
  "In the single-IV model, researchers commonly rely on t-ratio-based inference, even though the literature has quantified its potentially severe large-sample distortions. Building on Stock and Yogo (2005), we introduce the tF critical value function, leading to a standard error adjustment that is a smooth function of the first-stage F-statistic. For one-quarter of specifications in 61 AER papers, corrected standard errors are at least 49 and 136 percent larger than conventional 2SLS standard errors at the 5-percent and 1-percent significance levels, respectively. tF confidence intervals have shorter expected length than those of Anderson and Rubin (1949), whenever both are bounded."

- **Mikusheva and Sun (2022), "Inference with Many Weak Instruments," REStud**
  "We develop a concept of weak identification in linear instrumental variable models in which the number of instruments can grow at the same rate or slower than the sample size. We propose a jackknifed version of the classical weak identification-robust Anderson–Rubin (AR) test statistic. Large-sample inference based on the jackknifed AR is valid under heteroscedasticity and weak identification. The feasible version of this statistic uses a novel variance estimator. The test has uniformly correct size and good power properties. We also develop a pre-test for weak identification that is related to

the size property of a Wald test based on the Jackknife Instrumental Variable Estimator. This new pre-test is valid under heteroscedasticity and with many instruments."

- **Keane and Neal (2023), "Instrument Strength in IV Estimation and Inference: A Guide to Theory and Practice," JE**
  "Two stage least squares (2SLS) has poor properties if instruments are exogenous but weak. But how strong do instruments need to be for 2SLS estimates and test statistics to exhibit acceptable properties? A common standard is that first-stage $\geq 10$. This is adequate to ensure two-tailed $t$-tests have modest size distortions. But other problems persist: In particular, we show 2SLS standard errors are artificially small in samples where the estimate is most contaminated by the OLS bias. Hence, if the bias is positive, the $t$-test has little power to detect true negative effects, and inflated power to find positive effects. This phenomenon, which we call a "power asymmetry," persists even if first-stage $F$ is in the thousands. Robust tests like Anderson-Rubin perform better, and should be used in lieu of the $t$-test even with strong instruments. We also show how 2SLS test statistics typically suffer from very low power if first-stage $F$ is only 10, leading us to suggest a higher standard of instrument strength in empirical practice."

# 5 Shift-Share IV[3]

- **Adão, Kolesár, and Morales (2019), "Shift-Share Designs: Theory and Inference," QJE**
  "We study inference in shift-share regression designs, such as when a regional outcome is regressed on a weighted average of sectoral shocks, using regional sector shares as weights. We conduct a placebo exercise in which we estimate the effect of a shift-share regressor constructed with randomly generated sectoral shocks on actual labor market outcomes across U.S. commuting zones. Tests based on commonly used standard errors with 5% nominal significance level reject the null of no effect in up to 55% of the placebo samples. We use a stylized economic model to show that this overrejection problem arises because regression residuals are correlated across regions with similar sectoral shares, independent of their geographic location. We derive novel inference methods that are valid under arbitrary cross-regional correlation in the regression residuals. We show using popular applications of shift-share designs that our methods may lead to substantially wider confidence intervals in practice."

[3]The references in this section are taken from a guest lecture that Peter Hull gave in Arindrajit Dube's ECON 797B Fall 2020 course at UMass Amherst – see lecture slides here.

- **Borusyak and Hull (2021), "Non-Random Exposure to Exogenous Shocks," WP**
  "We develop a new approach to estimating the causal effects of treatments or instruments that combine multiple sources of variation according to a known formula. Examples include treatments capturing spillovers in social or transportation networks and simulated instruments for policy eligibility. We show how exogenous shocks to some, but not all, determinants of such variables can be leveraged while avoiding omitted variables bias. Our solution involves specifying counterfactual shocks that may as well have been realized and adjusting for a summary measure of non-randomness in shock exposure: the average treatment (or instrument) across shock counterfactuals. We use this approach to address bias when estimating employment effects of market access growth from Chinese high-speed rail construction."

- **Goldsmith-Pinkham, Sorkin, and Swift (2020), "Bartik Instruments: What, When, Why, and How," AER**
  "The Bartik instrument is formed by interacting local industry shares and national industry growth rates. We show that the typical use of a Bartik instrument assumes a pooled exposure research design, where the shares measure differential exposure to common shocks, and identification is based on exogeneity of the shares. Next, we show how the Bartik instrument weights each of the exposure designs. Finally, we discuss how to assess the plausibility of the research design. We illustrate our results through two applications: estimating the elasticity of labor supply, and estimating the elasticity of substitution between immigrants and natives."

- **Borusyak, Hull, and Jaravel (2022), "Quasi-Experimental Shift-Share Research Designs," REStud**
  "Many studies use shift-share (or Bartik) instruments, which average a set of shocks with exposure share weights. We provide a new econometric framework for shift-share instrumental variable (SSIV) regressions in which identification follows from the quasi-random assignment of shocks, while exposure shares are allowed to be endogenous. The framework is motivated by an equivalence result: the orthogonality between a shift-share instrument and an unobserved residual can be represented as the orthogonality between the underlying shocks and a shock-level unobservable. SSIV regression coefficients can similarly be obtained from an equivalent shock-level regression, motivating shock-level conditions for their consistency. We discuss and illustrate several practical insights delivered by this framework in the setting of Autor et al. (2013)."

# 6 RD Designs[4]

- **Shen and Zhang (2016), "Distributional Tests for Regression Discontinuity: Theory and Empirical Examples," REStat**
  "This paper proposes consistent testing methods for examining the effect of a policy treatment on the whole distribution of a response outcome within the setting of a regression discontinuity design. These methods are particularly useful when a policy is expected to produce treatment effects that are heterogeneous along some unobserved characteristics. The test statistics are Kolmogorov-Smirnov-type and are asymptotically distribution free when the data are i.i.d. The proposed tests are applied to three seminal RD studies (Pop-Eleches & Urquiola, 2013; Abdulkadiroglu, Angrist, & Pathak, 2014; and Battistin et al., 2009)."

- **Arai and Ichimura (2018), "Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator," QE**
  "A new bandwidth selection method that uses different bandwidths for the local linear regression estimators on the left and the right of the cutoff point is proposed for the sharp regression discontinuity design estimator of the average treatment effect at the cutoff point. The asymptotic mean squared error of the estimator using the proposed bandwidth selection method is shown to be smaller than other bandwidth selection methods proposed in the literature. The approach that the bandwidth selection method is based on is also applied to an estimator that exploits the sharp regression kink design. Reliable confidence intervals compatible with both of the proposed bandwidth selection methods are also proposed as in the work of Calonico, Cattaneo, and Titiunik (2014a). An extensive simulation study shows that the proposed method's performances for the samples sizes 500 and 2000 closely match the theoretical predictions. Our simulation study also shows that the common practice of halving and doubling an optimal bandwidth for sensitivity check can be unreliable."

- **Armstrong and Kolesár (2018), "Optimal Inference in a Class of Regression Models," ECMA**
  "We consider the problem of constructing confidence intervals (CIs) for a linear functional of a regression function, such as its value at a point, the regression discontinuity parameter, or a regression coefficient in a linear or partly linear regression. Our main assumption is that the regression function is known to lie in a convex function class,

[4] See also this RD tutorial by Mattias Cattaneo, made for the 2020 Chamberlain Online Seminar Series. Here is a webpage with several built-in commands to run RDDs.

which covers most smoothness and/or shape assumptions used in econometrics. We derive finitesample optimal CIs and sharp efficiency bounds under normal errors with known variance. We show that these results translate to uniform (over the function class) asymptotic results when the error distribution is not known. When the function class is centrosymmetric, these efficiency bounds imply that minimax CIs are close to efficient at smooth regression functions. This implies, in particular, that it is impossible to form CIs that are substantively tighter using datadependent tuning parameters, and maintain coverage over the whole function class. We specialize our results to inference on the regression discontinuity parameter, and illustrate them in simulations and an empirical application."

- **Canay and Kamat (2018), "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design," REStud**
  "In the regression discontinuity design (RDD), it is common practice to assess the credibility of the design by testing whether the means of baseline covariates do not change at the cut-off (or threshold) of the running variable. This practice is partly motivated by the stronger implication derived by Lee (2008), who showed that under certain conditions the distribution of baseline covariates in the RDD must be continuous at the cut-off. We propose a permutation test based on the so-called induced ordered statistics for the null hypothesis of continuity of the distribution of baseline covariates at the cut-off; and introduce a novel asymptotic framework to analyse its properties. The asymptotic framework is intended to approximate a small sample phenomenon: even though the total number $n$ of observations may be large, the number of effective observations local to the cut-off is often small. Thus, while traditional asymptotics in RDD require a growing number of observations local to the cut-off as $n \to \infty$, our framework keeps the number $q$ of observations local to the cut-off fixed as $n \to \infty$. The new test is easy to implement, asymptotically valid under weak conditions, exhibits finite sample validity under stronger conditions than those needed for its asymptotic validity, and has favourable power properties relative to tests based on means. In a simulation study, we find that the new test controls size remarkably well across designs. We then use our test to evaluate the plausibility of the design in Lee (2008), a well-known application of the RDD to study incumbency advantage."

- **Ganong and Jäger (2018), "A Permutation Test for the Regression Kink Design," JASA**
  "The regression kink (RK) design is an increasingly popular empirical method for estimating causal effects of policies, such as the effect of unemployment benefits on unem-

ployment duration. Using simulation studies based on data from existing RK designs, we empirically document that the statistical significance of RK estimators based on conventional standard errors can be spurious. In the simulations, false positives arise as a consequence of nonlinearities in the underlying relationship between the outcome and the assignment variable, confirming concerns about the misspecification bias of discontinuity estimators pointed out by Calonico, Cattaneo, and Titiunik. As a complement to standard RK inference, we propose that researchers construct a distribution of placebo estimates in regions with and without a policy kink and use this distribution to gauge statistical significance. Under the assumption that the location of the kink point is random, this permutation test has exact size in finite samples for testing a sharp null hypothesis of no effect of the policy on the outcome. We implement simulation studies based on existing RK applications that estimate the effect of unemployment benefits on unemployment duration and show that our permutation test as well as inference procedures proposed by Calonico, Cattaneo, and Titiunik improve upon the size of standard approaches, while having sufficient power to detect an effect of unemployment benefits on unemployment duration. Supplementary materials for this article are available online."

- **Kolesár and Rothe (2018), "Inference in Regression Discontinuity Designs with a Discrete Running Variable," AER**
  "We consider inference in regression discontinuity designs when the running variable only takes a moderate number of distinct values. In particular, we study the common practice of using confidence intervals (CIs) based on standard errors that are clustered by the running variable as a means to make inference robust to model misspecification (Lee and Card 2008). We derive theoretical results and present simulation and empirical evidence showing that these CIs do not guard against model misspecification, and that they have poor coverage properties. We therefore recommend against using these CIs in practice. We instead propose two alternative CIs with guaranteed coverage properties under easily interpretable restrictions on the conditional expectation function."

- **Calonico, Cattaneo, Farrell, and Titiunik (2019), "Regression Discontinuity Designs Using Covariates," REStat**
  "We study regression discontinuity designs when covariates are included in the estimation. We examine local polynomial estimators that include discrete or continuous covariates in an additive separable way, but without imposing any parametric restrictions on the underlying population regression functions. We recommend a covariate-adjustment approach that retains consistency under intuitive conditions and characterize the poten-

tial for estimation and inference improvements. We also present new covariate-adjusted mean-squared error expansions and robust bias-corrected inference procedures, with heteroskedasticity-consistent and cluster-robust standard errors. We provide an empirical illustration and an extensive simulation study. All methods are implemented in R and Stata software packages."

- **Gelman and Imbens (2019), "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs," JBES**
  "It is common in regression discontinuity analysis to control for third, fourth, or higher-degree polynomials of the forcing variable. There appears to be a perception that such methods are theoretically justified, even though they can lead to evidently nonsensical results. We argue that controlling for global high-order polynomials in regression discontinuity analysis is a flawed approach with three major problems: it leads to noisy estimates, sensitivity to the degree of the polynomial, and poor coverage of confidence intervals. We recommend researchers instead use estimators based on local linear or quadratic polynomials or other smooth functions."

- **Hsu and Shen (2019), "Testing for Treatment Effect Heterogeneity in Regression Discontinuity Design," JE**
  "Treatment effect heterogeneity is frequently studied in regression discontinuity (RD) applications. This paper proposes, under the RD setup, the first set of formal tests for treatment effect heterogeneity among subpopulations with different characteristics. The proposed tests study whether a policy treatment is 1) beneficial for at least some subpopulations defined by covariate values, 2) has any impact on at least some subpopulations, and 3) has a heterogeneous impact across subpopulations. Monte Carlo simulations show good small sample performance of the proposed tests. The empirical section applies the tests to study the impact of attending a better high school and discover interesting patterns of treatment effect heterogeneity neglected by previous studies."

- **Imbens and Wager (2019), "Optimized Regression Discontinuity Designs," REStat**
  "The increasing popularity of regression discontinuity methods for causal inference in observational studies has led to a proliferation of different estimating strategies, most of which involve first fitting nonparametric regression models on both sides of a treatment assignment boundary and then reporting plug-in estimates for the effect of interest. In applications, however, it is often difficult to tune the nonparametric regressions in a way that is well calibrated for the specific target of inference; for example, the model with the best global in-sample fit may provide poor estimates of the discontinuity parameter,

which depends on the regression function at boundary points. We propose an alternative method for estimation and statistical inference in regression discontinuity designs that uses numerical convex optimization to directly obtain the finite-sample-minimax linear estimator for the regression discontinuity parameter, subject to bounds on the second derivative of the conditional response function. Given a bound on the second derivative, our proposed method is fully data driven and provides uniform confidence intervals for the regression discontinuity parameter with both discrete and continuous running variables. The method also naturally extends to the case of multiple running variables."

- **Armstrong and Kolesár (2020), "Simple and Honest Confidence Intervals in Nonparametric Regression," QE**
  "We consider the problem of constructing honest confidence intervals (CIs) for a scalar parameter of interest, such as the regression discontinuity parameter, in nonparametric regression based on kernel or local polynomial estimators. To ensure that our CIs are honest, we use critical values that take into account the possible bias of the estimator upon which the CIs are based. We show that this approach leads to CIs that are more efficient than conventional CIs that achieve coverage by undersmoothing or subtracting an estimate of the bias. We give sharp efficiency bounds of using different kernels, and derive the optimal bandwidth for constructing honest CIs. We show that using the bandwidth that minimizes the maximum meansquared error results in CIs that are nearly efficient and that in this case, the critical value depends only on the rate of convergence. For the common case in which the rate of convergence is $n^{-2/5}$, the appropriate critical value for 95% CIs is 2.18, rather than the usual 1.96 critical value. We illustrate our results in a Monte Carlo analysis and an empirical application."

- **Bertanha and Imbens (2020), "External Validity in Fuzzy Regression Discontinuity Designs," JBES**
  "Fuzzy regression discontinuity designs identify the local average treatment effect (LATE) for the subpopulation of compliers, and with forcing variable equal to the threshold. We develop methods that assess the external validity of LATE to other compliance groups at the threshold, and allow for identification away from the threshold. Specifically, we focus on the equality of outcome distributions between treated compliers and always-takers, and between untreated compliers and never-takers. These equalities imply continuity of expected outcomes conditional on both the forcing variable and the treatment status. We recommend that researchers plot these conditional expectations and test for discontinuities at the threshold to assess external validity. We provide new commands

in STATA and MATLAB to implement our proposed procedures."

- **Bugni and Canay (2020), "Testing Continuity of a Density via $g$-order statistics in the Regression Discontinuity Design," JE**
  "In the regression discontinuity design (RDD), it is common practice to assess the credibility of the design by testing the continuity of the density of the running variable at the cut-off, e.g., McCrary (2008). In this paper we propose an approximate sign test for continuity of a density at a point based on the so-called $g$-order statistics, and study its properties under two complementary asymptotic frameworks. In the first asymptotic framework, the number $q$ of observations local to the cut-off is fixed as the sample size $n$ diverges to infinity, while in the second framework $q$ diverges to infinity slowly as $n$ diverges to infinity. Under both of these frameworks, we show that the test we propose is asymptotically valid in the sense that it has limiting rejection probability under the null hypothesis not exceeding the nominal level. More importantly, the test is easy to implement, asymptotically valid under weaker conditions than those used by competing methods, and exhibits finite sample validity under stronger conditions than those needed for its asymptotic validity. In a simulation study, we find that the approximate sign test provides good control of the rejection probability under the null hypothesis while remaining competitive under the alternative hypothesis. We finally apply our test to the design in Lee (2008), a well-known application of the RDD to study incumbency advantage."

- **Calonico, Cattaneo, and Farrell (2020), "Optimal Bandwidth Choice for Robust Bias Corrected Inference in Regression Discontinuity Designs," EJ**
  "Modern empirical work in regression discontinuity (RD) designs often employs local polynomial estimation and inference with a mean square error (MSE) optimal bandwidth choice. This bandwidth yields an MSE-optimal RD treatment effect estimator, but is by construction invalid for inference. Robust bias-corrected (RBC) inference methods are valid when using the MSE-optimal bandwidth, but we show that they yield suboptimal confidence intervals in terms of coverage error. We establish valid coverage error expansions for RBC confidence interval estimators and use these results to propose new inference-optimal bandwidth choices for forming these intervals. We find that the standard MSE-optimal bandwidth for the RD point estimator is too large when the goal is to construct RBC confidence intervals with the smaller coverage error rate. We further optimize the constant terms behind the coverage error to derive new optimal choices for the auxiliary bandwidth required for RBC inference. Our expansions also establish that RBC inference yields higher-order refinements (relative to traditional undersmoothing)

in the context of RD designs. Our main results cover sharp and sharp kink RD designs under conditional heteroskedasticity, and we discuss extensions to fuzzy and other RD designs, clustered sampling, and pre-intervention covariates adjustments. The theoretical findings are illustrated with a Monte Carlo experiment and an empirical application, and the main methodological results are available in R and Stata packages."

- **Cattaneo, Idrobo, and Titiunik (2020), "A Practical Introduction to Regression Discontinuity Designs: Foundations," WP**
  "In this Element and its accompanying second Element, A Practical Introduction to Regression Discontinuity Designs: Extensions, Matias Cattaneo, Nicolás Idrobo, and Rocío Titiunik provide an accessible and practical guide for the analysis and interpretation of regression discontinuity (RD) designs that encourages the use of a common set of practices and facilitates the accumulation of RD-based empirical evidence. In this Element, the authors discuss the foundations of the canonical Sharp RD design, which has the following features: (i) the score is continuously distributed and has only one dimension, (ii) there is only one cutoff, and (iii) compliance with the treatment assignment is perfect. In the second Element, the authors discuss practical and conceptual extensions to this basic RD setup."

- **Cattaneo, Jansson, and Ma (2020), "Simple Local Polynomial Density Estimators," JASA**
  "This article introduces an intuitive and easy-to-implement nonparametric density estimator based on local polynomial techniques. The estimator is fully boundary adaptive and automatic, but does not require prebinning or any other transformation of the data. We study the main asymptotic properties of the estimator, and use these results to provide principled estimation, inference, and bandwidth selection methods. As a substantive application of our results, we develop a novel discontinuity in density testing procedure, an important problem in regression discontinuity designs and other program evaluation settings. An illustrative empirical application is given. Two companion Stata and R software packages are provided."

- **Cattaneo, Keele, Titiunik, and Vazquez-Bare (2020), "Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs," JASA**
  "In nonexperimental settings, the regression discontinuity (RD) design is one of the most credible identification strategies for program evaluation and causal inference. However, RD treatment effect estimands are necessarily local, making statistical methods for the extrapolation of these effects a key area for development. We introduce a new method for extrapolation of RD effects that relies on the presence of multiple

cutoffs, and is therefore design-based. Our approach employs an easy-to-interpret identifying assumption that mimics the idea of common trends in difference-in-differences designs. We illustrate our methods with data on a subsidized loan program on post-education attendance in Colombia, and offer new evidence on program effects for students with test scores away from the cutoff that determined program eligibility. Supplementary materials for this article are available online."

- **Cattaneo and Titiunik (2022), "Regression Discontinuity Designs," ARE**
  "The regression discontinuity (RD) design is one of the most widely used nonexperimental methods for causal inference and program evaluation. Over the last two decades, statistical and econometric methods for RD analysis have expanded and matured, and there is now a large number of methodological results for RD identification, estimation, inference, and validation. We offer a curated review of this methodological literature organized around the two most popular frameworks for the analysis and interpretation of RD designs: the continuity framework and the local randomization framework. For each framework, we discuss three main topics: (a) designs and parameters, focusing on different types of RD settings and treatment effects of interest; (b) estimation and inference, presenting the most popular methods based on local polynomial regression and methods for the analysis of experiments, as well as refinements, extensions, and alternatives; and (c) validation and falsification, summarizing an array of mostly empirical approaches to support the validity of RD designs in practice."

- **Cattaneo, Idrobo, and Titiunik (2023), "A Practical Introduction to Regression Discontinuity Designs: Extensions," WP**
  "This monograph, together with its accompanying first part Cattaneo, Idrobo and Titiunik (2020), collects and expands the instructional materials we prepared for more than 40 short courses and workshops on Regression Discontinuity (RD) methodology that we taught between 2014 and 2022. In this second monograph, we discuss several topics in RD methodology that build on and extend the analysis of RD designs introduced in Cattaneo, Idrobo and Titiunik (2020). Our first goal is to present an alternative RD conceptual framework based on local randomization ideas. This methodological approach can be useful in RD designs with discretely-valued scores, and can also be used more broadly as a complement to the continuity-based approach in other settings. Then, employing both continuity-based and local randomization approaches, we extend the canonical Sharp RD design in multiple directions: fuzzy RD designs, RD designs with discrete scores, and multi-dimensional RD designs. The goal of our two-part monograph is purposely practical and hence we focus on the empirical analysis of RD designs."

- **Noack and Rothe (2023), "Donut Regression Discontinuity Designs," WP**
  "We study the econometric properties of so-called donut regression discontinuity (RD) designs, a robustness exercise which involves repeating estimation and inference without the data points in some area around the treatment threshold. This approach is often motivated by concerns that possible systematic sorting of units, or similar data issues, in some neighborhood of the treatment threshold might distort estimation and inference of RD treatment effects. We show that donut RD estimators can have substantially larger bias and variance than contentional RD estimators, and that the corresponding confidence intervals can be substantially longer. We also provide a formal testing framework for comparing donut and conventional RD estimation results."

- **Callaway, Goodman-Bacon, and Sant'Anna (2024b), "Event-Studies with a Continuous Treatment," NBER WP**
  "This paper considers methods for defining aggregate parameters of interest in a difference-in-differences design with a continuous and staggered treatment. It also discusses how aggregation choices often simplify estimation."

- **Callaway, Goodman-Bacon, and Sant'Anna (2024a), "Difference-in-Differences with a Continuous Treatment," NBER WP**
  "This paper analyzes difference-in-differences designs with a continuous treatment. We show that treatment effect on the treated-type parameters can be identified under a generalized parallel trends assumption that is similar to the binary treatment setup. However, interpreting differences in these parameters across different values of the treatment can be particularly challenging due to selection bias that is not ruled out by the parallel trends assumption. We discuss alternative, typically stronger, assumptions that alleviate these challenges. We also provide a variety of treatment effect decomposition results, highlighting that parameters associated with popular linear two-way fixed-effect (TWFE) specifications can be hard to interpret, even when there are only two time periods. We introduce alternative estimation procedures that do not suffer from these TWFE drawbacks, and show in an application that they can lead to different conclusions."

- **Lehner (2024), "A Note on Spatial Regression Discontinuity Designs," Working Paper**
  "Implementations of spatial regression discontinuity estimation and inference vary considerably in the literature. I show that many commonly used estimators for spatial RDDs, albeit working well in common scenarios, generally do not identify the local

average treatment effect at the RD boundary. Second, I propose a way to report heterogeneous treatment effects alongside the RD cutoff. Third, I introduce randomization inference to the spatial RD framework by creating a set of functions that allow to randomly shift borders. These tools might be interesting for other identification strategies that rely on the shift of boundaries. A companion R-package, SpatialRDD, includes all the tools necessary to carry out spatial RD estimation, including the proposed improvements. The package makes such applications more transparent, and, more importantly ensures easy and straightforward replicability of all necessary steps."

- Roth (2024), **"Interpreting Event-Studies from Recent Difference-in-Differences Methods," WP**
  "This note discusses the interpretation of event-study plots produced by recent difference-in-differences methods. I show that even when specialized to the case of non-staggered treatment timing, the default plots produced by software for three of the most popular recent methods (de Chaisemartin and DHaultfoeuille, 2020; Callaway and SantAnna, 2021; Borusyak, Jaravel and Spiess, 2024) do not match those of tra- ditional two-way fixed effects (TWFE) event-studies: the new methods may show a kink or jump at the time of treatment even when the TWFE event-study shows a straight line. This difference stems from the fact that the new methods construct the pre-treatment coefficients asymmetrically from the post-treatment coefficients. As a result, visual heuristics for analyzing TWFE event-study plots should not be immedi- ately applied to those from these methods. I conclude with practical recommendations for constructing and interpreting event-study plots when using these methods."

- Xie (2024), **"Nonlinear and Nonseparable Structural Functions in Regression Discontinuity Designs with a Continuous Treatment," JE**
  "Many empirical examples of regression discontinuity (RD) designs concern a continuous treatment variable, but the theoretical aspects of such models are less studied. This study examines the identification and estimation of the structural function in fuzzy RD designs with a continuous treatment variable. The structural function fully describes the causal impact of the treatment on the outcome. We show that the nonlinear and nonseparable structural function can be nonparametrically identified at the RD cutoff under shape restrictions, including monotonicity and smoothness conditions. Based on the nonparametric identification equation, we propose a three-step semiparametric estimation procedure and establish the asymptotic normality of the estimator. The semiparametric estimator achieves the same convergence rate as in the case of a binary treatment variable. As an application of the method, we estimate the causal effect of

sleep time on health status by using the discontinuity in natural light timing at time zone boundaries."

# 7 Synthetic Control

- **Botosaru and Ferman (2019), "On the Role of Covariates in the Synthetic Control Method," EJ**
  "Abadie et al. (2010) derive bounds on the bias of the synthetic control estimator under a perfect balance assumption on both observed covariates and pre-treatment outcomes. In the absence of a perfect balance on covariates, we show that it is still possible to derive such bounds, albeit at the expense of relying on stronger assumptions about the effects of observed and unobserved covariates and of generating looser bounds. We also show that a perfect balance on pre-treatment outcomes does not generally imply an approximate balance for all covariates, even when they are all relevant. Our results have important implications for the implementation of the method."

- **Abadie (2020), "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," JEL**

- **Athey, Bayati, Doudchenko, Imbens, and Khosravi (2020), "Matrix Completion Methods for Causal Panel Data Models," WP**
  "In this paper we study methods for estimating causal effects in settings with panel data, where some units are exposed to a treatment during some periods and the goal is estimating counterfactual (untreated) outcomes for the treated unit/period combinations. We develop a class of matrix completion estimators that uses the observed elements of the matrix of control outcomes corresponding to untreated unit/periods to impute the "missing" elements of the control outcome matrix, corresponding to treated units/periods. The approach estimates a matrix that well-approximates the original (incomplete) matrix, but has lower complexity according to the nuclear norm for matrices. We generalize results from the matrix completion literature by allowing the patterns of missing data to have a time series dependency structure. We present novel insights concerning the connections between the matrix completion literature, the literature on interactive fixed effects models and the literatures on program evaluation under unconfoundedness and synthetic control methods. We show that all these estimators can be viewed as focusing on the same objective function. They differ in the way they deal with lack of identification, in some cases solely through regularization (our proposed nuclear norm matrix completion estimator) and in other cases primarily through imposing

hard restrictions (the unconfoundedness and synthetic control approaches). proposed method outperforms unconfoundedness-based or synthetic control estimators."

- **Ferman, Pinto, and Possebom (2020), "Cherry Picking with Synthetic Controls," JPAM**
  "We evaluate whether a lack of guidance on how to choose the matching variables used in the Synthetic Control (SC) estimator creates specificationsearching opportunities. We provide theoretical results showing that specificationsearching opportunities are asymptotically irrelevant if we restrict to a subset of SC specifications. However, based on Monte Carlo simulations and simulations with real datasets, we show significant room for specification searching when the number of pretreatment periods is in line with common SC applications, and when alternative specifications commonly used in SC applications are also considered. This suggests that such lack of guidance generates a substantial level of discretion in the choice of the comparison units in SC applications, undermining one of the advantages of the method. We provide recommendations to limit the possibilities for specification searching in the SC method. Finally, we analyze the possibilities for specification searching and provide our recommendations in a series of empirical applications."

- **Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021), "Synthetic Difference in Differences," AER**
  "We present a new estimator for causal effects with panel data that builds on insights behind the widely used difference in differences and synthetic control methods. Relative to these methods, we find, both theoretically and empirically, that the proposed synthetic difference in differences estimator has desirable robustness properties, and that it performs well in settings where the conventional estimators are commonly used in practice. We study the asymptotic behavior of the estimator when the systematic part of the outcome model includes latent unit factors interacted with latent time factors, and we present conditions for consistency and asymptotic normality."

- **Ben-Michael, Feller, and Rothstein (2021a), "Synthetic Controls with Staggered Adoption," WP**
  "Staggered adoption of policies by different units at different times creates promising opportu- nities for observational causal inference. Estimation remains challenging, however, and common regression methods can give misleading results. A promising alternative is the synthetic control method (SCM), which finds a weighted average of control units that closely balances the treated units pre-treatment outcomes. In this paper, we generalize SCM, originally designed to study a single treated unit, to the staggered

adoption setting. We first bound the error for the average effect and show that it depends on both the imbalance for each treated unit separately and the imbalance for the average of the treated units. We then propose "partially pooled" SCM weights to minimize a weighted combination of these measures; approaches that focus only on balancing one of the two components can lead to bias. We extend this approach to incorporate unit-level intercept shifts and auxiliary covariates. We assess the performance of the proposed method via extensive simulations and apply our results to the question of whether teacher collective bargaining leads to higher school spending, finding minimal impacts. We implement the proposed method in the `augsynth R` package."

- **Ben-Michael, Feller, and Rothstein (2021b), "The Augmented Synthetic Control Method," WP**
  "The synthetic control method (SCM) is a popular approach for estimating the impact of a treatment on a single unit in panel data settings. The "synthetic control" is a weighted average of control units that balances the treated unit's pre-treatment outcomes as closely as possible. A critical feature of the original proposal is to use SCM only when the fit on pre-treatment outcomes is excellent. We propose Augmented SCM as an extension of SCM to settings where such pre-treatment fit is infeasible. Analogous to bias correction for inexact matching, Augmented SCM uses an outcome model to estimate the bias due to imperfect pre-treatment fit and then de-biases the original SCM estimate. Our main proposal, which uses ridge regression as the outcome model, directly controls pre-treatment fit while minimizing extrapolation from the convex hull. This estimator can also be expressed as a solution to a modified synthetic controls problem that allows negative weights on some donor units. We bound the estimation error of this approach under different data generating processes, including a linear factor model, and show how regularization helps to avoid over-fitting to noise. We demonstrate gains from Augmented SCM with extensive simulation studies and apply this framework to estimate the impact of the 2012 Kansas tax cuts on economic growth. We implement the proposed method in the new `augsynth R` package."

- **Ferman and Pinto (2021), "Synthetic Controls with Imperfect Pre-Treatment Fit," QE**
  "We analyze the properties of the Synthetic Control (SC) and related estimators when the pre- treatment fit is imperfect. In this framework, we show that these estimators are generally biased if treatment assignment is correlated with unobserved confounders, even when the number of pre-treatment periods goes to infinity. Still, we show that a demeaned version of the SC method can improve in terms of bias and variance rela-

tive to the difference-in-difference estimator. We also derive a specification test for the demeaned SC estimator in this setting with imperfect pre-treatment fit. Given our theoretical results, we provide practical guidance for applied researchers on how to justify the use of such estimators in empirical applications."

- Gunsilius (2022), **"Distributional Synthetic Controls," WP**
  "The method of synthetic controls is a widely-adopted tool for evaluating causal effects of policy changes in settings with observational data. In many settings where it is applicable, researchers want to identify causal effects of policy changes on a treated unit at an aggregate level while having access to data at a finer granularity. We develop a simple extension of the synthetic controls estimator that takes advantage of this additional structure and provides nonparametric estimates of the heterogeneity within the aggregate unit. The idea is to replicate the quantile function associated with the treated unit by a weighted average of quantile functions of the control units. This estimator has desirable properties and relies on the same mathematical theory as the changes-in-changes estimator. It operates in both repeated cross-sections and panel data with as little as a single pre-treatment period and provides a unique counterfactual quantile function for continuous, discrete or mixed distributions."

# 8 Matching

- Otsu and Rai (2017), **"Bootstrap Inference of Matching Estimators for Average Treatment Effects," JASA**
  "It is known that the naive bootstrap is not asymptotically valid for a matching estimator of the average treatment effect with a fixed number of matches. In this article, we propose asymptotically valid inference methods for matching estimators based on the weighted bootstrap. The key is to construct bootstrap counterparts by resampling based on certain linear forms of the estimators. Our weighted bootstrap is applicable for the matching estimators of both the average treatment effect and its counterpart for the treated population. Also, by incorporating a bias correction method in Abadie and Imbens (2011), our method can be asymptotically valid even for matching based on a vector of covariates. A simulation study indicates that the weighted bootstrap method is favorably comparable with the asymptotic normal approximation. As an empirical illustration, we apply the proposed method to the National Supported Work data. Supplementary materials for this article are available online."

- Adusumilli (2018), **"Bootstrap Inference for Propensity Score Matching," WP**

"Propensity score matching, where the propensity scores are estimated in a first step, is widely used for estimating treatment effects. In this context, the naive bootstrap is invalid (Abadie and Imbens, 2008). This paper proposes a novel bootstrap procedure for the propensity score matching estimator, and demonstrates its consistency. The proposed bootstrap is built around the notion of potential errors, introduced in this paper. Precisely, each observation is associated with two potential error terms, corresponding to each of the potential states - treated or control - only one of which is realized. Thus, the variability of the estimator stems not only from the randomness of the potential errors themselves, but also from the probabilistic nature of treatment assignment, which randomly realizes one of the potential error terms. The proposed bootstrap takes both sources of randomness into account by resampling the potential errors as a pair as well as re-assigning new values for the treatments. Simulations and real data examples demonstrate the superior performance of the proposed method relative to using the asymptotic distribution for inference, especially when the degree of overlap in propensity scores is poor. General versions of the procedure can also be applied to other causal effect estimators such as inverse probability weighting and propensity score sub-classification, potentially leading to higher order refinements for inference in such contexts."

- **Imai, Kim, and Wang (2020), "Matching Methods for Causal Inference with Time-Series Cross-Sectional Data," WP**
  "Matching methods improve the validity of causal inference by reducing model dependence and offering intuitive diagnostics. While they have become a part of the standard tool kit across disciplines, matching methods are rarely used when analyzing time-series cross-sectional data. We fill this methodological gap. In the proposed approach, we first match each treated observation with control observations from other units in the same time period that have an identical treatment history up to the pre-specified number of lags. We use standard matching and weighting methods to further refine this matched set so that the treated and matched control observations have similar covariate values. Assessing the quality of matches is done by examining covariate balance. Finally, we estimate both short-term and long-term average treatment effects using the difference-in-differences estimator, accounting for a time trend. We illustrate the proposed methodology through simulation and empirical studies. An open-source software package is available for implementing the proposed methods."

- **Ferman (2021b), "Matching Estimators with Few Treated and Many Control Observations," JE**

"We analyze the properties of matching estimators when there are few treated, but many control observations. We show that, under standard assumptions, the nearest neighbor matching estimator for the average treatment effect on the treated is asymptotically unbiased in this framework. However, when the number of treated observations is fixed, the estimator is not consistent, and it is generally not asymptotically normal. Since standard inference methods are inadequate, we propose alternative inference methods, based on the theory of randomization tests under approximate symmetry, that are asymptotically valid in this framework. We show that these tests are valid under relatively strong assumptions when the number of treated observations is fixed, and under weaker assumptions when the number of treated observations increases, but at a lower rate relative to the number of control observations."

# 9 Bunching[5]

- **Kleven (2016), "Bunching," ARE**
  "Recent years have seen a surge of applied work using bunching approaches, a development that is closely linked to the increased availability of administrative data. These approaches exploit the incentives for bunching created by discontinuities in the slope of choice sets (kinks) or in the level of choice sets (notches) to study the behavior of individuals and firms. Although the bunching approach was originally developed in the context of taxation, it is beginning to find applications in many other areas, such as social security, social insurance, welfare programs, education, regulation, private sector prices, and reference-dependent preferences. This review provides a guide to bunching estimation, discusses its strengths and weaknesses, surveys a range of applications across fields, and considers reasons for the ubiquity of kinks and notches."

- **Blomquist, Newey, Kumar, and Liang (2019), "On Bunching and Identification of the Taxable Income Elasticity," NBER WP**
  "The taxable income elasticity is a key parameter for predicting the effect of tax reform or designing an income tax. Bunching at kinks and notches in a single budget set have been used to estimate the taxable income elasticity. We show that when the heterogeneity distribution is unrestricted the amount of bunching at a kink or a notch is not informative about the size of the taxable income elasticity, and neither is the entire distribution of taxable income for a convex budget set. Kinks do provide information

[5] See this 2018 Bunching Estimator Workshop webpage for more references on bunching, including recent applications (thanks to Ben Solow for sharing this link).

about the size of the elasticity when a priori restrictions are placed on the heterogeneity distribution. They can identify the elasticity when the heterogeneity distribution is specified across the kink and provide bounds under restrictions on the heterogeneity distribution. We also show that variation in budget sets can identify the taxable income elasticity when the distribution of preferences is unrestricted and stable across budget sets. For nonparametric utility with general heterogeneity we show that kinks only provide elasticity information about individuals at the kink and we give bounds analogous to those for isoelastic utility. Identification becomes more difficult with optimization errors We show in examples how results are affected by optimization errors."

- **Caetano, Caetano, and Nelson (2020), "Correcting for Endogeneity in Models with Bunching," WP**
  "We show that in models with endogeneity, bunching at the lower or upper boundary of the distribution of the treatment variable may be used to build a correction for endogeneity. We derive the asymptotic distribution of the parameters of the corrected model, provide an estimator of the standard errors, and prove the consistency of the bootstrap. An empirical application reveals that time spent watching television, corrected for endogeneity, has roughly no net effect on cognitive skills and a significant negative net effect on non-cognitive skills in children."

- **Marx (2020), "Dynamic Bunching Estimation with Panel Data," WP**
  "Bunching estimation of distortions in a distribution around a policy threshold provides a means of studying behavioral parameters. Standard cross-sectional bunching estimators rely on identification assumptions about heterogeneity that I show can be violated by serial dependence of the choice variable or attrition related to the threshold. I propose a dynamic bunching estimation design that exploits panel data to obtain identification from relative within-agent changes in income and to estimate new parameters. Simulations using household income data demonstrate the benefits of the panel design. An application to charitable organizations demonstrates opportunities for estimating elasticity correlates, causal effects, and extensive-margin responses."

- **Caetano, Caetano, Fe, and Nielsen (2021), "A Dummy Test of Identification in Models with Bunching," WP**
  "We propose a simple test of the main identification assumption in models where the treatment variable takes multiple values and has bunching. The test consists of adding an indicator of the bunching point to the estimation model and testing whether the coefficient of this indicator is zero. Although similar in spirit to the test in Caetano (2015),

the dummy test has important practical advantages: it is more powerful at detecting endogeneity, and it also detects violations of the functional form assumption. The test does not require exclusion restrictions and can be implemented in many approaches popular in empirical research, including linear, two-way fixed effects, and discrete choice models. We apply the test to the estimation of the effect of a mothers working hours on her child's skills in a panel data context (James-Burdumy 2005)."

# 10 Sufficient Statistics

- **Lee, Leung, O'Leary, Pei, and Quach (2020), "Are Sufficient Statistics Necessary? Nonparametric Measurement of Deadweight Loss from Unemployment Insurance," JOLE**
  "Central to the welfare analysis of income transfer programs is the deadweight loss associated with possible reforms. To aid analytical tractability, its measurement typically requires specifying a simplified model of behavior. We employ a complementary "decomposition" approach that compares the behavioral and mechanical components of a policy's total impact on the government budget to study the deadweight loss of two unemployment insurance policies. Experimental and quasi-experimental estimates using state administrative data show that increasing the weekly benefit is more efficient (with a fiscal externality of 53 cents per dollar of mechanical transferred income) than reducing the programs implicit earnings tax."

- **Kleven (2021), "Sufficient Statistics Revisited," ARE**
  "This paper reviews and generalizes the sufficient statistics approach to policy evaluation. The idea of the approach is that the welfare effect of policy changes can be expressed in terms estimable reduced-form elasticities, allowing for policy evaluation without estimating the structural primitives of fully specified models. The approach relies on three assumptions: that policy changes are small, that government policy is the only source of market imperfection, and that a set of high-level restrictions on the environment and on preferences can be used to reduce the number of elasticities to be estimated. We generalize the approach in all three dimensions. It is possible to develop transparent sufficient statistics formulas under very general conditions, but the estimation requirements increase greatly. Starting from such general formulas elucidates that feasible empirical implementations are in fact structural approaches."

# 11 Decomposition

- **Callaway, Li, and Oka (2018), "Quantile Treatment Effects in Difference in Differences Models under Dependence Restrictions and with Only Two Time Periods," JE**
  "This paper shows that the Conditional Quantile Treatment Effect on the Treated is identified under (i) a Conditional Distributional Difference in Differences assumption and (ii) a new assumption that the dependence (the copula) between the change in untreated potential outcomes and the initial level of untreated potential outcomes is the same for the treated group and untreated group. We consider estimation and inference with discrete covariates and propose a uniform inference procedure based on the exchangeable bootstrap. Finally, we estimate the effect of increasing the minimum wage on the distribution of earnings for subgroups defined by race, gender, and education."

- **Firpo, Fortin, and Lemieux (2018), "Decomposing Wage Distributions using Recentered Influence Function Regressions," Econometrics**
  "This paper provides a detailed exposition of an extension of the Oaxaca-Blinder decomposition method that can be applied to various distributional measures. The two-stage procedure first divides distributional changes into a wage structure effect and a composition effect using a reweighting method. Second, the two components are further divided into the contribution of each explanatory variable using recentered influence function (RIF) regressions. We illustrate the practical aspects of the procedure by analyzing how the polarization of U.S. male wages between the late 1980s and the mid 2010s was affected by factors such as de-unionization, education, occupations, and industry changes."

- **d'Haultfœuille, Maurel, and Zhang (2018), "Extremal Quantile Regressions for Selection Models and the Black-White Wage Gap," JE**
  "We consider models with endogenous selection and no instrument nor large support regressors. Identification relies on the independence between the covariates and selection, when the outcome tends to infinity. We propose a simple estimator based on extremal quantile regressions and apply it to the evolution of the black-white wage gap in the US."

- **Wüthrich (2019), "A Closed-Form Estimator for Quantile Treatment Effects with Endogeneity," JE**
  "This paper studies the estimation of quantile treatment effects based on the instrumental variable quantile regression (IVQR) model (Chernozhukov and Hansen, 2005).

I develop a class of flexible plug-in estimators based on closed-form solutions derived from the IVQR moment conditions. The proposed estimators remain tractable and root-n-consistent, while allowing for rich patterns of effect heterogeneity. Functional central limit theorems and bootstrap validity results for the estimators of the quantile treatment effects and other functionals are provided. Monte Carlo simulations demonstrate favorable finite sample properties of the proposed approach. I apply my method to reanalyze the causal effect of 401(k) plans."

- **Hsu, Lai, and Lieli (2020), "Counterfactual Treatment Effects: Estimation and Inference," JBES**
  "This article proposes statistical methods to evaluate the quantile counterfactual treatment effect (QCTE) if one were to change the composition of the population targeted by a status quo program. QCTE enables a researcher to carry out an ex-ante assessment of the distributional impact of certain policy interventions or to investigate the possible explanations for treatment effect heterogeneity. Assuming unconfoundedness and invariance of the conditional distributions of the potential outcomes, QCTE is identified and can be nonparametrically estimated by a kernel-based method. Viewed as a random function over the continuum of quantile indices, the estimator converges weakly to a zero mean Gaussian process at the parametric rate. We propose a multiplier bootstrap procedure to construct uniform confidence bands, and provide similar results for average effects and for the counterfactually treated subpopulation. We also present Monte Carlo simulations and two counterfactual exercises that provide insight into the heterogeneous earnings effects of the Job Corps training program in the United States."

- **Chernozhukov, Fernández-Val, and Weidner (2020), "Network and Panel Quantile Effects via Distribution Regression," JE**
  "This paper provides a method to construct simultaneous confidence bands for quantile functions and quantile effects in nonlinear network and panel models with unobserved two-way effects, strictly exogenous covariates, and possibly discrete outcome variables. The method is based upon projection of simultaneous confidence bands for distribution functions constructed from fixed effects distribution regression estimators. These fixed effects estimators are debiased to deal with the incidental parameter problem. Under asymptotic sequences where both dimensions of the data set grow at the same rate, the confidence bands for the quantile functions and effects have correct joint coverage in large samples. An empirical application to gravity models of trade illustrates the applicability of the methods to network data."

- **Hausman, Liu, Luo, and Palmer (2020), "Errors in the Dependent Variable of Quan-**

**tile Regression Models," ECMA**

"We study the consequences of measurement error in the dependent variable of random-coefficients models, focusing on the particular case of quantile regression. The popular quantile regression estimator of Koenker and Bassett (1978) is biased if there is an additive error term. Approaching this problem as an errors-in- variables problem where the dependent variable suffers from classical measurement error, we present a sieve maximum-likelihood approach that is robust to left-hand side measurement error. After providing sufficient conditions for identification, we demonstrate that when the number of knots in the quantile grid is chosen to grow at an adequate speed, the sieve maximum-likelihood estimator is consistent and asymptotically normal, permitting inference via bootstrapping. Monte Carlo evidence verifies our method outperforms quantile regression in mean bias and MSE. Finally, we illustrate our estimator with an application to the returns to education highlighting changes over time in the returns to education that have previously been masked by measurement-error bias."

- **Słoczyński (2020), "Average Gaps and Oaxaca–Blinder Decompositions: A Cautionary Tale about Regression Estimates of Racial Differences in Labor Market Outcomes," ILRR**

  "Using a recent result from the program evaluation literature, the author demonstrates that the interpretation of regression estimates of between-group differences in wages and other economic outcomes depends on the relative sizes of subpopulations under study. When the disadvantaged group is small, regression estimates are similar to the average loss for disadvantaged individuals. When this group is a numerical majority, regression estimates are similar to the average gain for advantaged individuals. The author analyzes racial test score gaps using ECLS-K data and racial wage gaps using CPS, NLSY79, and NSW data, and shows that the interpretation of regression estimates varies substantially across data sets. Methodologically, he develops a new version of the OaxacaBlinder decomposition, in which the unexplained component recovers a parameter referred to as the average outcome gap. Under additional assumptions, this estimand is equivalent to the average treatment effect. Finally, the author reinterprets the Reimers, Cotton, and Fortin decompositions in the context of the program evaluation literature, with attention to the limitations of these approaches."

- **Borden, Haupt, and Wiborg (2021), "A New Framework for Estimation of Unconditional Quantile Treatment Effects: The Residualized Quantile Regression (RQR) Model," WP**

  "The identification of unconditional quantile treatment effects (QTE) has become in-

creasingly popular within social sciences. However, current methods to identify unconditional QTEs of continuous treatment variables are incomplete. Contrary to popular belief, the unconditional quantile regression model introduced by Firpo, Fortin, and Lemieux (2009) does not identify QTE, while the propensity score framework of Firpo (2007) allows for only a binary treatment variable, and the generalized quantile regression model of Powell (2020) is unfeasible with high-dimensional fixed effects. This paper introduces a two-step approach to estimate unconditional QTEs. In the first step, the treatment variable is regressed on the control variables using an OLS. Then, in the second step, the residuals from the OLS are used as the treatment variable in a quantile regression model. The intuition is that the first step decomposes the variance of the treatment variable into a piece explained by the observed control variables and a residual piece independent of the controls. Since the control variables purge the treatment of confounding, they are redundant in the second step. Therefore, our RQR approach circumvents the problem that the inclusion of controls together with the treatment variable in CQR changes the interpretation of the treatment coefficients. Unlike much of the literature on quantile regression, this two-step residualized quantile regression framework is easy to understand, computationally fast, and can include high-dimensional fixed effects."

- **Franguridi, Gafarov, and Wüthrich (2021), "Conditional Quantile Estimators: A Small Sample Theory," WP**
  "We study the small sample properties of conditional quantile estimators such as classical and IV quantile regression. First, we propose a higher-order analytical framework for comparing competing estimators in small samples and assessing the accuracy of common inference procedures. Our framework is based on a novel approximation of the discontinuous sample moments by a Hölder-continuous process with a negligible error. For any consistent estimator, this approximation leads to asymptotic linear expansions with nearly optimal rates. Second, we study the higher-order bias of exact quantile estimators up to O $(1/n)$. Using a novel non-smooth calculus technique, we uncover previously unknown non-negligible bias components that cannot be consistently estimated and depend on the employed estimation algorithm. To circumvent this problem, we propose a "symmetric" bias correction, which admits a feasible implementation. Our simulations confirm the empirical importance of bias correction."

- **Guo and Basse (2021), "The Generalized Oaxaca-Blinder Estimator," JASA**
  "After performing a randomized experiment, researchers often use ordinary-least squares (OLS) regression to adjust for baseline covariates when estimating the average treat-

ment effect. It is widely known that the resulting confidence interval is valid even if the linear model is misspecified. In this paper, we generalize that conclusion to covariate adjustment with nonlinear models. We introduce an intuitive way to use any "simple" nonlinear model to construct a covariate-adjusted confidence interval for the average treatment effect. The confidence interval derives its validity from randomization alone, and when nonlinear models fit the data better than linear models, it is narrower than the usual interval from OLS adjustment."

- **Chetverikov, Liu, and Tsyvinski (2022), "Weighted-Average Quantile Regression," NBER WP**
  "In this paper, we introduce the weighted-average quantile regression model. We argue that this model is of interest in many applied settings and develop an estimator for parameters of this model. We show that our estimator is $\sqrt{T}$-consistent and asymptotically normal with mean zero under weak conditions, where T is the sample size. We demonstrate the usefulness of our estimator in two empirical settings. First, we study the factor structures of the expected shortfalls of the industry portfolios. Second, we study inequality and social welfare dependence on individual characteristics."

# 12 General

- **Wan, Xie, and Zhou (2017), "A Varying Coefficient Approach to Estimating Hedonic Housing Price Functions and their Quantiles," JAS**
  "The varying coefficient (VC) model introduced by Hastie and Tibshirani [26] is arguably one of the most remarkable recent developments in nonparametric regression theory. The VC model is an extension of the ordinary regression model where the coefficients are allowed to vary as smooth functions of an effect modifier possibly different from the regressors. The VC model reduces the modelling bias with its unique structure while also avoiding the curse of dimensionality problem. While the VC model has been applied widely in a variety of disciplines, its application in economics has been minimal. The central goal of this paper is to apply VC modelling to the estimation of a hedonic house price function using data from Hong Kong, one of the world's most buoyant real estate markets. We demonstrate the advantages of the VC approach over traditional parametric and semi-parametric regressions in the face of a large number of regressors. We further combine VC modelling with quantile regression to examine the heterogeneity of the marginal effects of attributes across the distribution of housing prices."

- **Abadie and Cattaneo (2018), "Econometric Methods for Program Evaluation," ARE**
  "Program evaluation methods are widely applied in economics to assess the effects of policy interventions and other treatments of interest. In this article, we describe the main methodological frameworks of the econometrics of program evaluation. In the process, we delineate some of the directions along which this literature is expanding, discuss recent developments, and highlight specific areas where new research may be particularly fruitful."

- **Chen, Wan, Tso, and Zhang (2018), "A Model Averaging Approach for the Ordered Probit and Nested Logit Models with Applications," JAS**
  "This paper considers model averaging for the ordered probit and nested logit models, which are widely used in empirical research. Within the frameworks of these models, we examine a range of model averaging methods, including the jackknife method, which is proved to have an optimal asymptotic property in this paper. We conduct a large-scale simulation study to examine the behaviour of these model averaging estimators in finite samples, and draw comparisons with model selection estimators. Our results show that while neither averaging nor selection is a consistently better strategy, model selection results in the poorest estimates far more frequently than averaging, and more often than not, averaging yields superior estimates. Among the averaging methods considered, the one based on a smoothed version of the Bayesian Information criterion frequently produces the most accurate estimates. In three real data applications, we demonstrate the usefulness of model averaging in mitigating problems associated with the replication crisis that commonly arises with model selection."

- **Słoczyński and Wooldridge (2020), "A General Double Robustness Result for Estimating Average Treatment Effects," ET**
  "In this paper we study doubly robust estimators of various average and quantile treatment effects under unconfoundedness; we also consider an application to a setting with an instrumental variable. We unify and extend much of the recent literature by providing a very general identification result which covers binary and multi-valued treatments; unnormalized and normalized weighting; and both inverse-probability weighted (IPW) and doubly robust estimators. We also allow for subpopulation-specific average treatment effects where subpopulations can be based on covariate values in an arbitrary way. Similar to Wooldridge (2007), we then discuss estimation of the conditional mean using quasi-log likelihoods (QLL) from the linear exponential family."

- **Broderick, Giordano, and Meager (2020), "An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?," WP**

"We propose a method to assess the sensitivity of econometric analyses to the removal of a small fraction of the sample. Analyzing all possible data subsets of a certain size is computationally prohibitive, so we provide a finite-sample metric to approximately compute the number (or fraction) of observations that has the greatest influence on a given result when dropped. We call our resulting metric the Approximate Maximum Influence Perturbation. Our approximation is automatically computable and works for common estimators (including OLS, IV, GMM, MLE, and variational Bayes). We provide explicit finite-sample error bounds on our approximation for linear and instrumental variables regressions. At minimal computational cost, our metric provides an exact finite-sample lower bound on sensitivity for any estimator, so any non-robustness our metric finds is conclusive. We demonstrate that the Approximate Maximum Influence Perturbation is driven by a low signal-to-noise ratio in the inference problem, is not reflected in standard errors, does not disappear asymptotically, and is not a product of misspecification. Several empirical applications show that even 2-parameter linear regression analyses of randomized trials can be highly sensitive. While we find some applications are robust, in others the sign of a treatment effect can be changed by dropping less than 1% of the sample even when standard errors are small."

- **Chernozhukov, Wüthrich, and Zhu (2020), "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," WP**
  "We introduce new inference procedures for counterfactual and synthetic control methods for policy evaluation. We recast the causal inference problem as a counter- factual prediction and a structural breaks testing problem. This allows us to exploit insights from conformal prediction and structural breaks testing to develop permutation inference procedures that accommodate modern high-dimensional estimators, are valid under weak and easy-to-verify conditions, and are provably robust against misspecification. Our methods work in conjunction with many different approaches for predicting counterfactual mean outcomes in the absence of the policy intervention. Examples include synthetic controls, difference-in-differences, factor and matrix completion models, and (fused) time series panel data models. Our approach demonstrates an excellent small-sample performance in simulations and is taken to a data application where we re-evaluate the consequences of decriminalizing indoor prostitution."

- **Cunningham (2020), *Causal Inference: The Mixtape* [HTML version]**

- **Ferman (2021a), "A Simple Way to Assess Inference Methods," WP**
  We propose a simple way to assess whether inference methods are reliable. The assessment can detect problems when the asymptotic theory that justifies the inference

method is invalid and/or provides a poor approximation given the design of the empirical application. It can be easily applied to a wide range of applications. We show that, despite being a simple idea and despite its limitations, this assessment has the potential of making scientific evidence more reliable, if it becomes widely used by applied researchers. We analyze in detail the cases of Difference-in-Differences with few treated clusters, shift-share designs, weighted OLS, stratied experiments, and matching estimators."

- **Han (2021), "Identification in Nonparametric Models for Dynamic Treatment Effects," JE**
  "This paper develops a nonparametric model that represents how sequences of outcomes and treatment choices influence one another in a dynamic manner. In this setting, we are interested in identifying the average outcome for individuals in each period, had a particular treatment sequence been assigned. The identification of this quantity allows us to identify the average treatment effects (ATE's) and the ATEs on transitions, as well as the optimal treatment regimes, namely, the regimes that maximize the (weighted) sum of the average potential outcomes, possibly less the cost of the treatments. The main contribution of this paper is to relax the sequential randomization assumption widely used in the biostatistics literature by introducing a flexible choice-theoretic framework for a sequence of endogenous treatments. This framework allows non-compliance of subjects in experimental studies or endogenous treatment decisions in observational settings. We show that the parameters of interest are identified under each periods exclusion restrictions, which are motivated by, e.g., a sequence of randomized treatment assignments or encouragements and a behavioral/information assumption on agents who receive treatments."

# References

Abadie, A. (2020). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, forthcoming.

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2022). When Should You Adjust Standard Errors for Clustering? *Quarterly Journal of Economics*.

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2020). Sampling-Based Versus Design-Based Uncertainty in Regression Analysis. *Econometrica 88*(1), 265–296.

Abadie, A. and M. D. Cattaneo (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics 10*, 465–503.

Abraham, S. and L. Sun (2021). Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics 225*(2), 175–199.

Adão, R., M. Kolesár, and E. Morales (2019). Shift-Share Designs: Theory and Inference. *Quarterly Journal of Economics 134*(4), 1949–2010.

Adusumilli, K. (2018). Bootstrap Inference for Propensity Score Matching. *Working Paper*.

Andresen, E. M. and M. Huber (2021). Instrument-based estimation with binarized treatments: Issues and tests for the exclusion restriction. *Econometrics Journal*.

Andrews, I. and T. B. Armstrong (2017). Unbiased Instrumental Variables Estimation under Known First-Stage Sign. *Quantitative Economics 8*(2), 479–503.

Andrews, I., J. H. Stock, and L. Sun (2019). Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics 11*, 727–753.

Angrist, J. and M. Kolesár (2022). One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV. *Working Paper*.

Arai, Y. and H. Ichimura (2018). Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator. *Quantitative Economics 9*(1), 441–482.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021). Synthetic Difference in Differences. *American Economic Review*.

Arkhangelsky, D. and G. Imbens (2023). Causal Models for Longitudinal and Panel Data: A Survey. *Working Paper*.

Arkhangelsky, D., G. W. Imbens, L. Lei, and X. Luo (2021). Double-Robust Two-Way-Fixed-Effects Regression For Panel Data. *Working Paper*.

Armstrong, T. B. and M. Kolesár (2018). Optimal Inference in a Class of Regression Models. *Econometrica 86*(2), 655–683.

Armstrong, T. B. and M. Kolesár (2020). Simple and Honest Confidence Intervals in Nonparametric Regression. *Quantitative Economics 11*(1), 1–39.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2020). Matrix Completion Methods for Causal Panel Data Models. *Working Paper*.

Athey, S., P. J. Bickel, A. Chen, G. Imbens, and M. Pollmann (2021). Semiparametric Estimation of Treatment Effects in Randomized Experiments. *NBER Working Paper*.

Athey, S. and G. W. Imbens (2018). Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption. *NBER Working Paper*.

Baker, A. C., D. F. Larcker, and C. C. Y. Wang (2021). How Much Should We Trust Staggered Differences-In-Differences Estimates? *Working Paper*.

Ben-Michael, E., A. Feller, and J. Rothstein (2021a). Synthetic Controls with Staggered Adoption. *NBER Working Paper*.

Ben-Michael, E., A. Feller, and J. Rothstein (2021b). The Augmented Synthetic Control Method. *NBER Working Paper*.

Bertanha, M. and G. W. Imbens (2020). External Validity in Fuzzy Regression Discontinuity Designs. *Journal of Business and Economic Statistics 38*(3), 593–612.

Blomquist, S., W. Newey, A. Kumar, and C.-Y. Liang (2019). On Bunching and Identification of the Taxable Income Elasticity. *NBER Working Paper*.

Borden, N. T., A. Haupt, and . Wiborg (2021). A New Framework for Estimation of Unconditional Quantile Treatment Effects: The Residualized Quantile Regression (RQR) Model. *Working Paper*.

Borusyak, K. and P. Hull (2021). Non-random exposure to exogenous shocks. *Working Paper*.

Borusyak, K., P. Hull, and X. Jaravel (2022). Quasi-Experimental Shift-Share Research Designs. *Review of Economic Studies 89*(1), 181–213.

Borusyak, K., X. Jaravel, and J. Spiess (2024). Revisiting Event Study Designs: Robust and Efficient Estimation. *Review of Economic Studies*, rdae007.

Botosaru, I. and B. Ferman (2019). On the Role of Covariates in the Synthetic Control Method. *Econometrics Journal 22*(2), 117–130.

Brewer, M., T. F. Crossley, and R. Joyce (2017). Inference with Difference-in-Differences Revisited. *Journal of Econometric Methods 7*(1).

Broderick, T., R. Giordano, and R. Meager (2020). An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions? *Working Paper*.

Bugni, F. A. and I. A. Canay (2020). Testing Continuity of a Density via g-order statistics in the Regression Discontinuity Design. *Journal of Econometrics*.

Burlig, F., L. Preonas, and M. Woerman (2020). Panel Data and Experimental Design. *Journal of Development Economics*, 102458.

Butts, K. (2021). Difference-in-Differences Estimation with Spatial Spillovers. *Working Paper*.

Caetano, C., G. Caetano, H. Fe, and E. Nielsen (2021). A Dummy Test of Identification in Models with Bunching. *Working Paper*.

Caetano, C., G. Caetano, and E. Nelson (2020). Correcting for Endogeneity in Models with Bunching.

Caetano, C., B. Callaway, P. Stroud, and H. Sant'Anna Rodrigues (2022). Difference in Differences with Time-Varying Covariates. *Working Paper*.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna (2024a). Difference-in-Differences with a Continuous Treatment. *NBER WP*.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna (2024b). Event-Studies with a Continuous Treatment. *NBER WP*.

Callaway, B., T. Li, and T. Oka (2018). Quantile Treatment Effects in Difference in Differences Models under Dependence Restrictions and with Only Two Time Periods. *Journal of Econometrics 206*(2), 395–413.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-Differences with Multiple Time Periods. *Journal of Econometircs 225*(2), 200–230.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2020). Optimal Bandwidth Choice for Robust Bias-Corrected Inference in Regression Riscontinuity Designs. *Econometrics Journal 23*(2), 192–210.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019). Regression Discontinuity Designs Using Covariates. *Review of Economics and Statistics 101*(3), 442–451.

Canay, I. A. and V. Kamat (2018). Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design. *Review of Economic Studies 85*(3), 1577–1608.

Canay, I. A., A. Santos, and A. M. Shaikh (2018). The Wild Bootstrap with a "Small" Number of "Large" Clusters. *Review of Economics and Statistics*, 1–45.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2020). *A Practical Introduction to Regression Discontinuity Designs: Foundations*.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2023). *A Practical Introduction to Regression Discontinuity Designs: Extensions*.

Cattaneo, M. D., M. Jansson, and X. Ma (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association 115*(531), 1449–1455.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018). Inference in Linear Regression Models with Many Covariates and Heteroscedasticity. *Journal of the American Statistical Association 113*(523), 1350–1361.

Cattaneo, M. D., L. Keele, R. Titiunik, and G. Vazquez-Bare (2020). Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs. *Journal of the American Statistical Association*, 1–12.

Cattaneo, M. D. and R. Titiunik (2022). Regression Discontinuity Designs. *Annual Review of Economics 14*, 821–851.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). The Effect of Minimum Wages on Low-Wage Jobs. *Quarterly Journal of Economics 134*(3), 1405–1454.

Chen, L., A. T. Wan, G. Tso, and X. Zhang (2018). A Model Averaging Approach for the Ordered Probit and Nested Logit Models with Applications. *Journal of Applied Statistics 45*(16), 3012–3052.

Chernozhukov, V., I. Fernández-Val, and M. Weidner (2020). Network and Panel Quantile Effects via Distribution Regression. *Journal of Econometrics*.

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2020). An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. *Working Paper*.

Chetverikov, D., Y. Liu, and A. Tsyvinski (2022). Weighted-Average Quantile Regression. *NBER Working Paper*.

Choi, J., J. Gu, and S. Shen (2018). Weak-Instrument Robust Inference for Two-Sample Instrumental Variables Regression. *Journal of Applied Econometrics 33*(1), 109–125.

Choi, J. and S. Shen (2019). Two-sample instrumental-variables regression with potentially weak instruments. *Stata Journal 19*(3), 581–597.

Colella, F., R. Lalive, S. O. Sakalli, and M. Thoenig (2020). Inference with Arbitrary Clustering. *Working Paper*.

Cunningham, S. (2020). *Causal Inference: The Mixtape*. Yale University Press.

de Chaisemartin, C. and X. d'Haultfoeuille (2018). Fuzzy Differences-in-Differences. *Review of Economic Studies 85*(2), 999–1028.

de Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review 110*(9), 2964–96.

de Chaisemartin, C. and X. D'Haultfœuille (2024). Two-Way Fixed Effects and Differences-in-Differences Estimators in Heterogeneous Adoption Designs. *Working Paper*.

Deeb, A. and C. de Chaisemartin (2020). Clustering and External Validity in Randomized Controlled Trials. *Working Paper*.

d'Haultfœuille, X., A. Maurel, and Y. Zhang (2018). Extremal Quantile Regressions for Selection Models and the Black-White Wage Gap. *Journal of Econometrics 203*(1), 129–142.

Dube, A., D. Girardi, . Jordà, and A. M. Taylor (2023). A Local Projections Approach to Difference-in-Differences Event Studies. *NBER Working Paper*.

Evdokimov, K. S. and M. Kolesár (2019). Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects. *Working Paper*.

Faletto, G. (2023). Fused Extended Two-Way Fixed Effects for Difference-in-Differences with Staggered Adoptions. *Working Paper*.

Ferman, B. (2021a). A Simple Way to Assess Inference Methods. *Working Paper*.

Ferman, B. (2021b). Matching estimators with few treated and many control observations. *Journal of Econometircs 225*(2), 295–307.

Ferman, B. and C. Pinto (2019). Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity. *Review of Economics and Statistics 101*(3), 452–467.

Ferman, B. and C. Pinto (2021). Synthetic Controls with Imperfect Pre-Treatment Fit. *Quantitative Economics*.

Ferman, B., C. Pinto, and V. Possebom (2020). Cherry Picking with Synthetic Controls. *Journal of Policy Analysis and Management 39*(2), 510–532.

Finley, B. (2020). Testing for Weak-Instrument Bias in Just-Identified 2SLS. *Working Paper*.

Firpo, S. P., N. M. Fortin, and T. Lemieux (2018). Decomposing Wage Distributions using Recentered Influence Function Regressions. *Econometrics 6*(2), 28.

Franguridi, G., B. Gafarov, and K. Wüthrich (2021). Conditional Quantile Estimators: A Small Sample Theory. *Working Paper*.

Freyaldenhoven, S., C. Hansen, J. P. Pérez, and J. M. Shapiro (2021). Visualization, Identification, and Estimation in the Linear Panel Event-Study Design. *NBER Working Paper*.

Freyaldenhoven, S., C. Hansen, and J. M. Shapiro (2019). Pre-event Trends in the Panel Event-Study Design. *American Economic Review 109*(9), 3307–38.

Gabriel, E., A. Sjölander, and M. C. Sachs (2021). Nonparametric Bounds for Causal Effects in Imperfect Randomized Experiments. *Journal of American Statistical Association*.

Ganong, P. and S. Jäger (2018). A Permutation Test for the Regression Kink Design. *Journal of the American Statistical Association 113*(522), 494–504.

Gardner, J. (2021). Two-Stage Differences-in-Differences. *Working Paper*.

Gelman, A. and G. Imbens (2019). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics 37*(3), 447–456.

Ghanem, D., P. H. C. Sant'Anna, and K. Wüthrich (2023). Selection and Parallel Trends. *Working Paper*.

Gibbons, C. E., J. C. S. Serrato, and M. B. Urbancic (2018). Broken or Fixed Effects? *Journal of Econometric Methods 8*(1).

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2024). Contamination Bias in Linear Regressions. *NBER WP*.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2021). On Estimating Multiple Treatment Effects with Regression. *Working Paper*.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020). Bartik Instruments: What, When, Why, and How. *American Economic Review 110*(8), 2586–2624.

Goodman-Bacon, A. (2021). Difference-in-Differences with Variation in Treatment Timing. *Journal of Econometircs 225*(2), 254–277.

Gunsilius, F. F. (2022). Distributional Synthetic Controls. *Working Paper*.

Guo, K. and G. Basse (2021). The Generalized Oaxaca-Blinder Estimator. *Journal of American Statistical Association*.

Han, S. (2021). Identification in nonparametric models for dynamic treatment effects. *Journal of Econometircs 225*(2), 132–147.

Hausman, J. A., H. Liu, Y. Luo, and C. Palmer (2020). Errors in the Dependent Variable of Quantile Regression Models. *Econometrica*, forthcoming.

Hsu, Y.-C., T.-C. Lai, and R. P. Lieli (2020). Counterfactual Treatment Effects: Estimation and Inference. *Journal of Business and Economic Statistics*, 1–16.

Hsu, Y.-C. and S. Shen (2019). Testing Treatment Effect Heterogeneity in Regression Discontinuity Designs. *Journal of Econometrics 208*(2), 468–486.

Huntington-Kleina, N. (2020). Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness. *Journal of Causal Inference*.

Ibragimov, R. and U. K. Müller (2016). Inference with Few Heterogeneous Clusters. *Review of Economics and Statistics 98*(1), 83–96.

Imai, K., I. S. Kim, and E. Wang (2020). Matching Methods for Causal Inference with Time-Series Cross-Sectional Data. *Working Paper*.

Imbens, G. and S. Wager (2019). Optimized Regression Discontinuity Designs. *Review of Economics and Statistics 101*(2), 264–278.

Imbens, G. W. and M. Kolesár (2016). Robust Standard Errors in Small Samples: Some Practical Advice. *Review of Economics and Statistics 98*(4), 701–712.

Keane, M. and T. Neal (2023). Instrument Strength in IV Estimation and Inference: A Guide to Theory and Practice. *Journal of Econometrics*.

Kleven, H. (2021). Sufficient Statistics Revisited. *Annual Review of Economics*, forthcoming.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics 8*, 435–464.

Kolesár, M. and C. Rothe (2018). Inference in Regression Discontinuity Designs with a Discrete Running Variable. *American Economic Review 108*(8), 2277–2304.

Lee, D. L., J. McCrary, M. J. Moreira, and J. Porter (2022). Valid $t$-ratio Inference for IV. *American Economic Review*, forthcoming.

Lee, D. S., P. Leung, C. J. O'Leary, Z. Pei, and S. Quach (2020). Are Sufficient Statistics Necessary? Nonparametric Measurement of Deadweight Loss from Unemployment Insurance. *Journal of Labor Economics*, forthcoming.

Lehner, A. (2024). A Note on Spatial Regression Discontinuity Designs. *Working Paper*.

MacKinnon, J. G. (2023). Using Large Samples in Econometrics. *Journal of Econometrics 235*(2), 922–926.

MacKinnon, J. G., M. . Nielsen, and M. D. Webb (2022a). Cluster-Robust Inference: A Guide to Empirical Practice. *Journal of Econometrics*.

MacKinnon, J. G., M. . Nielsen, and M. D. Webb (2022b). Testing for the Appropriate Level of Clustering in Linear Regression Models. *Working Paper*.

Marcus, M. and P. H. Sant'Anna (2020). The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics. *Journal of the Association of Environmental and Resource Economists*, forthcoming.

Marx, B. M. (2020). Dynamic Bunching Estimation with Panel Data. *Working Paper*.

Mikusheva, A. and L. Sun (2022). Inference with Many Weak Instruments. *Review of Economic Studies 89*(5), 2663–2686.

Mogstad, M. and A. Torgovitsky (2018). Identification and Extrapolation of Causal Effects with Instrumental Variables. *Annual Review of Economics 10*, 577–613.

Mogstad, M., A. Torgovitsky, and C. Walters (2020a). Policy Evaluation with Multiple Instrumental Variables. *Working Paper*.

Mogstad, M., A. Torgovitsky, and C. R. Walters (2020b). The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables. *Working Paper*.

Muralidharan, K., M. Romero, and K. Wüthrich (2019). Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments. *NBER WP*.

Müller, U. K. and M. W. Watson (2022). Spatial Correlation Robust Inference. *Econometrica*, forthcoming.

Noack, C. and C. Rothe (2023). Donut Regression Discontinuity Designs. *Working Paper*.

Otsu, T. and Y. Rai (2017). Bootstrap Inference of Matching Estimators for Average Treatment Effects. *Journal of the American Statistical Association 112*(520), 1720–1732.

Pichetti, P. and C. Pinto (2023). Marginal Treatment Effects in Difference-in-Differences. *Working Paper*.

Pustejovsky, J. E. and E. Tipton (2018). Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business & Economic Statistics 36*(4), 672–683.

Rambachan, A. and J. Roth (2023). A More Credible Approach to Parallel Trends. *Review of Economic Studies*, rdad018.

Roth, J. (2022). Pre-test with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights*, forthcoming.

Roth, J. (2024). Interpreting Event-Studies from Recent Difference-in-Differences Methods. *Working Paper*.

Roth, J. and P. H. Sant'Anna (2022). Efficient Estimation for Staggered Rollout Designs. *Working Paper*.

Roth, J., P. H. C. Sant'Anna, A. Bilinski, and J. Poe (2022). What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. *Journal of Econometrics*.

Sant'Anna, P. H. and J. Zhao (2020). Doubly Robust Difference-in-Differences Estimators. *Journal of Econometrics 219*(1), 101–122.

Schmidheiny, K. and S. Siegloch (2020). On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization. *Working Paper*.

Shen, S. and X. Zhang (2016). Distributional Tests for Regression Discontinuity: Theory and Empirical Examples. *Review of Economics and Statistics 98*(4), 685–700.

Słoczyński, T. (2020). Average Gaps and Oaxaca–Blinder Decompositions: A Cautionary Tale about Regression Estimates of Racial Differences in Labor Market Outcomes. *Industrial and Labor Relations Review 73*(3), 705–729.

Słoczyński, T. (2020). Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights. *Review of Economics and Statistics 0*(ja), 1–27.

Słoczyński, T. (2021). When Should We (Not) Interpret Linear IV Estimands as LATE? *Working Paper*.

Słoczyński, T. and J. Wooldridge (2020). A General Double Robustness Result for Estimating Average Treatment Effects. *Econometric Theory*.

Sun, L. and J. M. Shapiro (2022). Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure. *NBER Working Paper*.

Vazquez-Bare, G. (2022). Identification and Estimation of Spillover Effects in Randomized Experiments. *Journal of Econometrics*.

Wan, A. T., S. Xie, and Y. Zhou (2017). A Varying Coefficient Approach to Estimating Hedonic Housing Price Functions and their Quantiles. *Journal of Applied Statistics 44*(11), 1979–1999.

Wooldridge, J. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. *Working Paper*.

Wüthrich, K. (2019). A Closed-Form Estimator for Quantile Treatment Effects with Endogeneity. *Journal of Econometrics 210*(2), 219–235.

Xie, H. (2024). Nonlinear and Nonseparable Structural Functions in Regression Discontinuity Designs with a Continuous Treatment. *Journal of Econometrics 242*(1), 105784.

Young, A. (2019). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *Quarterly Journal of Economics 134*(2), 557–598.

Young, A. (2020). Consistency Without Inference: Instrumental Variables in Practical Application. *Working Paper*.

Zhao, A. and P. Ding (2021). Covariate-Adjusted Fisher Randomization Tests for the Average Treatment Effect. *Journal of Econometircs 225*(2), 278–294.