Literature on Recent Advances in Applied Micro Methods*

Christine Cai[†]

November 22, 2020

CLICK HERE FOR THE MOST RECENT VERSION

Contents

*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), as well as from some papers that have been circulated on #EconTwitter. Suggestions are welcome!

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

1 OLS

- **Ibragimov and Müller (2016), "Inference with Few Heterogeneous Clusters," REStat**
  "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."

- **Abadie, Athey, Imbens, and Wooldridge (2017), "When Should You Adjust Standard Errors for Clustering?," NBER WP**
  "In empirical work in economics it is common to report standard errors that account for clustering of units. Typically, the motivation given for the clustering adjustments is that unobserved components in outcomes for units within clusters are correlated. However, because correlation may occur across more than one dimension, this motivation makes it difficult to justify why researchers use clustering in some dimensions, such as geographic, but not others, such as age cohorts or gender. It also makes it difficult to explain why one should not cluster with data from a randomized experiment. In this paper, we argue that clustering is in essence a design problem, either a sampling design or an experimental design issue. It is a sampling design issue if sampling follows a two stage process where in the first stage, a subset of clusters were sampled randomly from a population of clusters, while in the second stage, units were sampled randomly from

the sampled clusters. In this case the clustering adjustment is justified by the fact that there are clusters in the population that we do not see in the sample. Clustering is an experimental design issue if the assignment is correlated within the clusters. We take the view that this second perspective best fits the typical setting in economics where clustering adjustments are used. This perspective allows us to shed new light on three questions: (i) when should one adjust the standard errors for clustering, (ii) when is the conventional adjustment for clustering appropriate, and (iii) when does the conventional adjustment of the standard errors matter."

- 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."

- 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."

2 RCT

- 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."

3 Diff-in-Diff & Event Studies

- Athey and Imbens (2018), **"Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption," 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."

- Borusyak and Jaravel (2018), **"Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume," WP**
  "A broad empirical literature uses "event study" research designs for treatment effect estimation, a setting in which all units in the panel receive treatment but at random times. We make four novel points about identification and estimation of causal effects in this setting and show their practical relevance. First, we show that in the presence of unit and time fixed effects, it is impossible to identify the linear component of the path of pre-trends and dynamic treatment effects. Second, we propose graphical and statistical tests for pre-trends. Third, we consider commonly-used "static" regressions, with a treatment dummy instead of a full set of leads and lags around the treatment event, and we show that OLS does not recover a reasonable weighted average of the treatment effects: long-run effects are weighted negatively, and we introduce different estimators robust to this issue. Fourth, we show that equivalent problems of under-identification and negative weighting arise in difference-in-differences settings when the control group is allowed to be on a different time trend or in the presence of unit-specific time trends. We show the practical relevance of these issues in a series of examples from the existing literature. We focus on the estimation of the marginal propensity to consume out of tax rebates: according to our preferred specification, the marginal propensity to consume is much lower than (about half of) the main estimates in the literature. The main message for practitioners is that because of identification issues and negative weighting in event study designs, results from common specifications are likely to seem non-robust. These problems can be alleviated in a principled way by using parametric and semi-parametric estimators and tests."

- 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 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."

- Abraham and Sun (2020), **"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."

- Callaway and Sant'Anna (2020), **"Difference-in-Differences with Multiple Time Periods," WP**
  "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."

- **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."

- **Goodman-Bacon (2020), "Difference-in-Differences with Variation in Treatment Timing," WP**
  "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 general estimator equals a weighted average of all possible two-group/two-period DD estimators in the data. This defines the DD estimand and identifying assumption, a generalization of common trends. I discuss how to interpret DD estimates and propose a new balance test. I show how to decompose the difference between two specifications, and provide a new analysis of models that include time-varying controls."

- **Rambachan and Roth (2020), "An Honest Approach to Parallel Trends," WP**
  "This paper proposes robust inference methods for difference-in-differences and event-study designs that do not require that the parallel trends assumption holds exactly. Instead, the researcher must only impose restrictions on the possible differences in trends between the treated and control groups. Several common intuitions expressed in applied work can be captured by such restrictions, including the notion that pre-treatment differences in trends are informative about counterfactual post-treatment differences in trends. Our methodology then guarantees uniformly valid ("honest") inference when the imposed restrictions are satisfied. We first show that fixed length confidence intervals have near-optimal expected length for a practically-relevant class of restrictions.

We next introduce a novel inference procedure that accommodates a wider range of restrictions, which is based on the observation that inference in our setting is equivalent to testing a system of moment inequalities with a large number of linear nuisance parameters. The resulting confidence sets are consistent, and have optimal local asymptotic power for many parameter configurations. We recommend researchers conduct sensitivity analyses to show what conclusions can be drawn under various restrictions on the possible differences in trends"

- **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."

4 Standard IV

- **Olea and Pflueger (2013), "A Robust Test for Weak Instruments," JBES**
  "We develop a test for weak instruments in linear instrumental variables regression that is robust to heteroscedasticity, autocorrelation, and clustering. Our test statistic is a scaled nonrobust first-stage F statistic. Instruments are considered weak when the two-stage least squares or the limited information maximum likelihood Nagar bias is large relative to a benchmark. We apply our procedures to the estimation of the elasticity of intertemporal substitution, where our test cannot reject the null of weak instruments in a larger number of countries than the test proposed by Stock and Yogo in 2005. Supplementary materials for this article are available online."

- Andrews and Armstrong (2017), **"Unbiased Instrumental Variables Estimation under Known First-Stage Sign," QE**

  "We derive mean-unbiased 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 non-randomized unbiased estimator based on the reduced-form and first-stage 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 two-stage least squares estimator. Our finite-sample results apply to normal models with known variance for the reduced-form errors, and imply analogous results under weak-instrument asymptotics with an unknown error distribution."

- 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 pro- pose 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."

- **Lee, McCrary, Moreira, and Porter (2020), "Valid t-ratio Inference for IV," WP**
  "In the single IV model, current practice relies on the first-stage F exceeding some threshold (e.g., 10) as a criterion for trusting t-ratio inferences, even though this yields an anti-conservative test. We show that a true 5 percent test instead requires an F greater than 104.7. Maintaining 10 as a threshold requires replacing the critical value 1.96 with 3.43. We re-examine 57 AER papers and find that corrected inference causes half of the initially presumed statistically significant results to be insignificant. We introduce a more powerful test, the tF procedure, which provides F-dependent adjusted t-ratio critical values."

- **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."

5 Shift-Share IV[1]

- **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."

- **Borusyak, Hull, and Jaravel (2020), "Quasi-Experimental Shift-Share Research Designs," WP**
  "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)."

- **Borusyak and Hull (2020), "Non-Random Exposure to Exogenous Shocks: Theory and Applications," NBER WP**
  "We develop new tools for causal inference in settings where exogenous shocks affect the treatment status of multiple observations jointly, to different extents. In these settings researchers may construct treatments or instruments that combine the shocks with

[1] 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.

predetermined measures of shock exposure. Examples include measures of spillovers in social and transportation networks, simulated eligibility instruments, and shift-share instruments. We show that leveraging the exogeneity of shocks for identification generally requires a simple but non-standard recentering, derived from the specification of counterfactual shocks that might as well have been realized. We further show how specification of counterfactual shocks can be used for finite-sample inference and specification tests, and we characterize the recentered instruments that are asymptotically efficient. We use this framework to estimate the employment effects of Chinese market access growth due to high-speed rail construction and the insurance coverage effects of expanded Medicaid eligibility."

- **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."

6 RDD

- **Calonico, Cattaneo, and Titiunik (2014), "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," ECMA**
  "In the regression-discontinuity (RD) design, units are assigned to treatment based on whether their value of an observed covariate exceeds a known cutoff. In this design, local polynomial estimators are now routinely employed to construct confidence intervals for treatment effects. The performance of these confidence intervals in applications, however, may be seriously hampered by their sensitivity to the specific bandwidth employed. Available bandwidth selectors typically yield a "large" bandwidth, leading to data-driven confidence intervals that may be biased, with empirical coverage well below their nominal target. We propose new theory-based, more robust confidence interval estimators for average treatment effects at the cutoff in sharp RD, sharp kink RD, fuzzy RD, and fuzzy kink RD designs. Our proposed confidence intervals are constructed

using a bias-corrected RD estimator together with a novel standard error estimator. For practical implementation, we discuss mean squared error optimal bandwidths, which are by construction not valid for conventional confidence intervals but are valid with our robust approach, and consistent standard error estimators based on our new variance formulas. In a special case of practical interest, our procedure amounts to running a quadratic instead of a linear local regression. More generally, our results give a formal justification to simple inference procedures based on increasing the order of the local polynomial estimator employed. We find in a simulation study that our confidence intervals exhibit close-to-correct empirical coverage and good empirical interval length on average, remarkably improving upon the alternatives available in the literature. All results are readily available in R and STATA using our companion software packages described in Calonico, Cattaneo, and Titiunik (2014d, 2014b)."

- 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, which covers most smoothness and/or shape assumptions used in econometrics. We derive finite-sample 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 data-dependent 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 unemployment 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."

- 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."

- 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 mean-squared 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."

- 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."

- 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."

7 Synthetic Control

- **Abadie (2019), "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," JEL**

- **Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2020), "Synthetic Difference in Differences," WP**
  "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."

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 proce-

dure 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."

9 Decomposition

- Fortin, Lemieux, and Firpo (2011), **"Decomposition Methods in Economics," HLE**
  "This chapter provides a comprehensive overview of decomposition methods that have been developed since the seminal work of Oaxaca and Blinder in the early 1970s. These methods are used to decompose the difference in a distributional statistic between two groups, or its change over time, into various explanatory factors. While the original work of Oaxaca and Blinder considered the case of the mean, our main focus is on other distributional statistics besides the mean, such as quantiles, the Gini coefficient or the variance. We discuss the assumptions required for identifying the different elements of the decomposition, as well as various estimation methods proposed in the literature. We also illustrate how these methods work in practice by discussing existing applications and working through a set of empirical examples throughout the paper."

- Chernozhukov, Fernández-Val, and Melly (2013), **"Inference on Counterfactual Distributions," ECMA**
  "Counterfactual distributions are important ingredients for policy analysis and decomposition analysis in empirical economics. In this article, we develop modeling and inference tools for counterfactual distributions based on regression methods. The counterfactual scenarios that we consider consist of ceteris paribus changes in either the

distribution of covariates related to the outcome of interest or the conditional distribution of the outcome given covariates. For either of these scenarios, we derive joint functional central limit theorems and bootstrap validity results for regression-based estimators of the status quo and counterfactual outcome distributions. These results allow us to construct simultaneous confidence sets for function-valued effects of the counterfactual changes, including the effects on the entire distribution and quantile functions of the outcome as well as on related functionals. These confidence sets can be used to test functional hypotheses such as no-effect, positive effect, or stochastic dominance. Our theory applies to general counterfactual changes and covers the main regression methods including classical, quantile, duration, and distribution regressions. We illustrate the results with an empirical application to wage decompositions using data for the United States."

10 General

- **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."

References

Abadie, A. (2019). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, forthcoming.

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017). When Should You Adjust Standard Errors for Clustering? *NBER Working Paper*.

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 (2020). Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics*, forthcoming.

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*.

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.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2020). Synthetic Difference in Differences. *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. and G. W. Imbens (2018). Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption.

Borusyak, K. and P. Hull (2020). Non-random exposure to exogenous shocks: Theory and applications. *NBER Working Paper*.

Borusyak, K., P. Hull, and X. Jaravel (2020). Quasi-Experimental Shift-Share Research Designs. *Working Paper*.

Borusyak, K. and X. Jaravel (2018). Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume. *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*.

Callaway, B. and P. H. Sant'Anna (2020). Difference-in-differences with multiple time periods. *Working Paper*.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica 82*(6), 2295–2326.

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., 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.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013). Inference on Counterfactual Distributions. *Econometrica 81*(6), 2205–2268.

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

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

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

Fortin, N., T. Lemieux, and S. Firpo (2011). Decomposition Methods in Economics. In *Handbook of Labor Economics*, Volume 4, pp. 1–102. Elsevier.

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

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.

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. (2020). Difference-in-Differences with Variation in Treatment Timing. *Working Paper*.

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

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.

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 (2020). Valid t-ratio Inference for IV. *Working Paper*.

Olea, J. L. M. and C. Pflueger (2013). A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics 31*(3), 358–369.

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.

Rambachan, A. and J. Roth (2020). An Honest Approach to Parallel Trends. *Working Paper*.

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

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*.