Journal of Agricultural Economics

JAE Journal of Agricultural Economics

ORIGINAL ARTICLE OPEN ACCESS

Estimating Causal Effects With Observational Data: Guidelines for Agricultural and Applied Economists

Arne Henningsen[1] | Guy Low[2,3] | David Wuepper[4] | Tobias Dalhaus[2,5] | Hugo Storm[6] | Dagim Belay[1] | Stefan Hirsch[7]

[1]Department of Food and Resource Economics, University of Copenhagen, Frederiksberg C, Denmark | [2]Business Economics Group, Wageningen University & Research, Wageningen, the Netherlands | [3]Department of Agricultural Economics and Rural Development, University of Göttingen, Göttingen, Germany | [4]Land Economics Group, Institute for Food and Resource Economics, University of Bonn, Bonn, Germany | [5]Agribusiness, Sustainable Food Systems Research Centre, Rhine-Waal University of Applied Sciences, Kleve, Germany | [6]Data Science in Agricultural Economics, Institute for Food and Resource Economics, University of Bonn, Bonn, Germany | [7]Department of Management in Agribusiness, University of Hohenheim, Stuttgart, Germany

Correspondence: Arne Henningsen (arne@ifro.ku.dk)

Received: 20 December 2024 | **Revised:** 24 November 2025 | **Accepted:** 26 November 2025

Keywords: causal inference | difference in differences | instrumental variables | observational data | regression discontinuity | synthetic controls

ABSTRACT

Most research questions in agricultural and applied economics are causal in nature: they study how changes in one or more variables (such as policies, prices or weather) affect one or more other variables (e.g., income, crop yields or pollution). Only a minority of these research questions can be studied with experimental methods, so most empirical studies in agricultural and applied economics rely on observational data. However, estimating causal effects with observational data requires an appropriate research design and a transparent discussion of all identifying assumptions, together with a critical discussion of how plausible they are. This paper provides an overview of approaches that are frequently used in agricultural and applied economics to estimate causal effects with observational data. It then provides advice and guidelines for agricultural and applied economists seeking to estimate causal effects with observational data, including how to assess and discuss the identification strategies adopted in their analysis.

JEL Classification: C21, C23, C24, C26, C51, C52

1 | Introduction

Today, around 50% of empirical economics articles focus on causal inference (Imbens 2024). However, a commonly observed problem in empirical research is that there is not always an obvious path to causal identification. Sometimes, the researcher might only be able to approximate causality without fully achieving it, for example by adjusting for some but not all confounders, or by addressing reverse causality but still failing to account for a systematic measurement error. In these cases, outstanding challenges to causal identification could continue to bias an estimate away from the true causal effect.

McKenzie et al. (2010) compared experimental and non-experimental methods in an empirical application and found that estimates from Ordinary Least Squares (OLS) regression, matching approaches, and Difference-in-Differences (DID) methods based on observational data overstate the effect of interest by 20%–82% compared to an experimental benchmark. However, econometric estimates based on observational data are often interpreted causally, without paying attention to the validity of the assumptions that allow this (Gibson 2019).

The misinterpretation of statistical associations as causal effects, together with insufficient robustness and replicability of

empirical analyses, has motivated the "credibility revolution" in quantitative economics research and a call for higher standards in statistical identification (Angrist and Pischke 2010; Bellemare 2012; Gibson 2019).[1] More recently, the availability of large data sets has led to the "big data fallacy", whereby decision makers often incorrectly consider a large data set as a substitute for a proper identification strategy and consequently misinterpret association as causation when the sample size is large (Vosgerau et al. 2025). While the "credibility revolution" has its origin in labour economics, it has increasingly reached into agricultural economics (Bellemare 2012). However, many empirical studies in agricultural economics continue to interpret estimated relationships between economic variables causally using terms such as "effect" or "impact", even when the underlying research design and econometric framework are not based on a valid identification strategy, or at least not a sufficiently described and motivated one. For example, some studies use OLS or matching methods, which rely on a selection-on-observables (conditional independence) assumption, although there is a high risk of this assumption being violated. The use of these methods possibly moves the estimates in the direction of the actual causal effect but often not sufficiently far that the estimates can be causally interpreted. Other examples are studies that use Instrumental Variables (IVs), such as 2-Stage Least Squares (2SLS) or endogenous switching regression, but do not sufficiently discuss or justify the validity of the IVs, or studies that use a DID design but do not scrutinise the parallel-trends assumption they are relying on. The mere application of such econometric approaches, yet without sufficient verification of the underlying assumptions, is often falsely regarded as a sufficient condition for the causal interpretation of the results. Incorrect use of causal identification approaches may even make the estimate worse and move it away from the actual causal effect. Examples are an erroneous null-finding because the parallel-trends assumption for the chosen DID estimator does not hold, or an exaggerated statistical significance because the IV does not produce a strong first-stage estimate.

The correct identification of causal effects is highly relevant for agricultural economics research because decisions and recommendations by policymakers, international organisations (e.g., FAO, IFAD, World Bank), NGOs, and the private sector (e.g., agribusinesses, farmers) in the area of agriculture and food often have crucial implications for society, in areas such as environmental sustainability, food safety, and food security (Finger et al. 2023).[2] Hence, empirical agricultural economics papers that aim to identify causal effects should include a clear description and justification of the underlying "identification strategy" (Imbens 2024). This refers to the identification of the exogenous variation in an endogenous covariate or treatment variable of interest, that is, the part of the variation in this variable that is not related to unobserved confounders (e.g., Gibson 2019; Lal et al. 2024). Only for this part of the variation in the endogenous covariate or treatment variable is it possible to say that it *affects* the dependent variable (e.g., Gibson 2019). Moreover, the limitations of the identification strategy should be clearly outlined and possible implications for the reliability of the results should be acknowledged and investigated.[3] If a specific method is used to address the non-experimental nature of the data, the added value compared to simpler approaches such as OLS should be highlighted. If the added value cannot be clearly shown, it may be preferable to stick with a simpler method and interpret the results as associations. Especially problematic are analyses in which an outcome is regressed on a set of explanatory variables and each coefficient is interpreted as reflecting the causal effect of the respective variable. This is usually inappropriate as, in most empirical applications, it is impossible to present a credible identification strategy for multiple explanatory variables and to avoid "bad controls" for estimating the causal effect of each of these explanatory variables.

The "gold standard" for internal validity is the use of Randomised Controlled Trials (RCTs) (Gibson 2019), and numerous examples can be found in the agricultural and applied economics literature (e.g., Bulte et al. 2014; Wilebore et al. 2019; El Benni et al. 2025).[4] However, RCTs also have important limitations (see, e.g., Barrett and Carter 2010). Most of the highly relevant research questions in agricultural and applied economics cannot be answered with experiments because they would be problematic, impractical, or infeasible for various reasons. For example, randomly assigning import tariffs, randomly assigning different levels of education to future farmers at their birth, increasing food prices in randomly selected regions, or restricting food aid to specific regions while excluding others that are also in need (Buchanan-Smith et al. 2016, 36) would be problematic for multiple reasons, including ethical considerations.[5] However, highly relevant research questions should not be neglected just because they cannot be answered by applying experimental methods. Instead, observational data must be used to answer these research questions as thoroughly as possible.

This paper discusses research designs and empirical methods that are frequently used in agricultural and applied economics to estimate causal effects with observational data. These discussions should help researchers, analysts and reviewers assess the suitability of these empirical approaches in their specific analysis, choose the most appropriate approach, justify their choice of approach, and interpret their results appropriately. Therefore, we extend previous literature that provides overviews (Imbens 2024) or guidelines on how to conduct econometric identification methods using IVs (e.g., Jiang 2017; Young 2022; Lal et al. 2024) for different disciplines, and tailor our guidelines to research questions and commonly used econometric approaches in agricultural and applied economics. We focus on the most common empirical research designs used in agricultural economics. For focus and brevity, we omit approaches that are less frequently used in agricultural economics so far, such as the regression kink approach (Cattaneo and Titiunik 2022), bunching (Caetano et al. 2025), and the front-door criterion (Bellemare et al. 2024).

Section 2 discusses the use of various methods that are based on the "selection on observables" identification strategy such as OLS and matching methods (e.g., propensity score matching). Section 3 explores methods based on IVs (or exclusion restrictions) such as 2SLS regression and endogenous switching regression. Section 4 discusses fixed-effects estimations and DID approaches. Section 5 describes the synthetic control method, while Section 6 examines regression discontinuity designs.

Section 7 provides general suggestions that apply to all methods and Section 8 concludes the paper.

2 | Selection on Observables

The selection-on-observables identification strategy is based on the assumption that we can observe and control for all variables that are correlated with both the treatment and the error term. This implies that there are no unobserved factors that are correlated with the treatment and affect the outcome through pathways that are not blocked by control variables. This assumption is also sometimes called the Conditional Independence Assumption (CIA), conditional ignorability, or conditional unconfoundedness.

2.1 | Classical Regression Methods

Classical regression analyses (e.g., OLS, logit, probit, tobit, or Poisson regression) can be affected by three potential sources of statistical endogeneity:[6] (a) omitted variables/unobserved heterogeneity; (b) measurement error (any type of measurement error in the explanatory variables or non-random measurement error in the dependent variable); and (c) reverse causality/simultaneity from which it follows that the dependent variable also influences the explanatory variable of interest. When discussing potential endogeneity in a regression analysis, it is advisable to focus on each of the three potential reasons separately (see, e.g., Bellemare and Novak 2017). Theoretically, all the explanatory variables must be uncorrelated with the error term, while in practice the discussion of endogeneity usually focuses on one or a few explanatory variables that are of particular interest for the research question, for example, treatment variables. If a control variable is correlated with the error term, the bias of the estimated coefficient(s) of interest depends on the relationship between this endogenous control variable and the explanatory variable of interest, that is, whether there is a direct correlation or indirect relationship through other control variables (see Frölich 2008; Bellemare 2015, the latter provides an illustrative example with only one control variable).[7]

2.2 | Directed Acyclic Graphs (DAGs)

Whether a selection-on-observables identification strategy may be feasible can be assessed, for example, by using Directed Acyclic Graphs (DAGs). DAGs are useful for at least two purposes. First, they clearly communicate and discuss assumptions about relationships between variables. Second, by applying certain rules or algorithms to DAGs (either manually or through available software tools[8]), sets of suitable control variables can be determined (Morgan and Winship 2014; Pearl and Mackenzie 2018).[9] This also includes the identification of variables that should *not* be used as control variables, that is, variables on the causal path from the treatment variable to the outcome variable ("bad controls"). DAGs were originally developed in computer science (Pearl and Mackenzie 2018), but are increasingly being used in economics (Imbens 2020; Hünermund and Bareinboim 2025). However, it is important to emphasise that a DAG should not be considered as the only "true" and universally valid presentation of the real world, but rather as a tool to communicate the underlying assumptions of an empirical analysis.

2.3 | Approaches to Address Endogeneity

Some studies aim to address unobserved heterogeneity by using a control variable that indicates the marginal utility of joining or leaving the "treatment" (Verhofstadt and Maertens 2014; Bellemare and Novak 2017; Ruml and Qaim 2021; Aïhounton and Henningsen 2024). Theoretically, this approach seems promising, but in practice it can be problematic because the control variable is usually observed after the decision to participate in the treatment has been made and, thus, it can be influenced by the treatment itself, which can introduce endogeneity (Aïhounton and Henningsen 2024).

Some empirical researchers try to address endogeneity by using lagged values instead of concurrent values of explanatory variables. Bellemare et al. (2017) show theoretically that using lagged values of explanatory variables addresses endogeneity only under the untestable assumption of "no dynamics among unobservables". Their Monte Carlo simulation shows that using lagged values of explanatory variables can result in substantially biased estimates and incorrect inference even if there are only low levels of dynamics among unobservables (Bellemare et al. 2017). Providing convincing arguments that there are no dynamics in any unobservable variables seems to be very difficult or impossible for most empirical studies.

There are methods to assess the sensitivity of the results to unobserved heterogeneity (e.g., Oster 2019; Cinelli and Hazlett 2020; Diegert et al. 2023), which have often been used in recent applied economics research. However, these methods are, in general, based on bold assumptions, and it is difficult or impossible to assess whether these assumptions are fulfilled in a specific empirical application. However, when applying a selection-on-observables identification strategy, these methods can contribute to assessing the suitability of the identification strategy if their assumptions are discussed appropriately and their results are interpreted carefully.

Bernard et al. (2024) recommend presenting confidence intervals that adjust conventional confidence intervals by incorporating the uncertainty about the bias that occurs due to using observational data (e.g., unobserved heterogeneity, violation of the Stable Unit Treatment Value Assumption (SUTVA), insufficient common support).

2.4 | Relaxing Functional Form Assumptions

Classical regression methods usually rely on strict assumptions about the functional form of the relationship between treatment variables, control variables, and the dependent variable. These restrictive assumptions can be partly or fully relaxed by using, for instance, the Augmented Inverse Propensity Weighted (AIPW) estimator,[10] semi- or nonparametric regression methods, matching methods such as propensity score matching (PSM),[11] or machine learning approaches. However, except for assumptions about the functional form, these methods are based

on the same identifying assumptions as regression methods (e.g., Angrist and Pischke 2009; Blattman 2010; Mullally and Chakravarty 2018). Therefore, the same discussion as for the use of regression methods is required.

In recent years, machine learning methods have rapidly advanced and are being increasingly used in agricultural and applied economics. It is important to note that most machine learning methods are unsuitable when they are used directly to estimate causal effects, even if all variables that are correlated with both the outcome and the treatment variable are observed. This is because machine learning methods are generally designed for prediction and not for the direct estimation of causal relationships. For example, machine learning approaches for variable selection (such as Lasso) select the subset of covariates that optimises out-of-sample prediction performance, but this selection likely introduces omitted-variable biases as it drops highly correlated control variables, including covariates that are correlated with both the outcome and the treatment variable.

However, machine learning methods can be used within established econometrics frameworks for causal identification such as under the selection-on-observables assumption or for IV estimation (see Section 3 and Appendix A). These methods are then called "causal machine learning." Despite this name, it should be clear that these methods are not new concepts for causal identification but rather extensions of the established econometrics frameworks of causal identification in which specific parts are replaced by machine learning methods. Hence, they come with the same identification assumptions that apply to "classical" econometric approaches and, thus, the same requirements to carefully consider and motivate an appropriate identification strategy. The basic idea of causal machine learning is to leverage the predictive capabilities of machine learning methods and their flexibility to approximate potentially complex relationships within these frameworks (Storm et al. 2020; Baylis et al. 2021). For example, under the selection-on-observables assumption, causal machine learning methods can be used to relax restrictive functional form assumptions, such as in the case of Double/Debiased Machine Learning (DML) (Chernozhukov et al. 2018), which assumes that the outcome model is a separable additive function, but that treatment effects, the influence of controls on outcomes, and the treatment assignment are unknown nonlinear functions. The approach allows the use of any machine learning algorithm to approximate these nonlinear functions and to derive average treatment effects.

2.5 | Estimating Treatment Heterogeneity

The "Causal Forests" method (Wager and Athey 2018), which is a special case of Generalised Random Forests (RF) (Athey et al. 2019), extends the DML approach allowing the estimation of heterogeneous treatment effects, that is, treatment effects that depend on observed characteristics (Conditional Average Treatment Effects, CATE). From an applied perspective, a crucial advantage is that treatment heterogeneity is estimated in a transparent and data-driven way, and thus avoids the need to predefine and potentially cherry-pick treatment groups. In agricultural economics, Causal Forests have already been applied in various contexts to study treatment heterogeneity (e.g., Deines et al. 2019, 2023; Stetter et al. 2022; Schulz et al. 2024), while Brignoli et al. (2024) conduct simulation studies to compare the performance of classical econometric methods, Causal Forests, and other machine-learning methods in the estimation of (heterogeneous) treatment effects with typical cross-sectional farm-level data.

2.6 | Suggestions

In summary, when relying on a selection-on-observables identification strategy, we suggest doing the following (in addition to following the general suggestions that we provide in Section 7):

- Clearly state the assumptions that the chosen method and model specification require for obtaining unbiased and/or consistent estimates.
- Use a DAG to find a suitable model specification (e.g., which control variables to include and which not to include) and to discuss the credibility of the chosen identification strategy, including potential unobserved confounders.
- Separately discuss the three potential sources of statistical endogeneity: (a) omitted variables/unobserved heterogeneity; (b) measurement error; and (c) reverse causality/simultaneity.
- Discuss the potential statistical endogeneity not only of the explanatory variable of interest but also of the control variables.
- Consider using placebo tests with outcome variables that should not be affected by the explanatory variable of interest, for example, using lagged values of the outcome variable (e.g., see Imbens and Wooldridge 2009, and Chabé-Ferret 2025b, section 8.3, for criticism of the former and also an alternative specification).
- Consider using methods for assessing the sensitivity of the results to unobserved heterogeneity (e.g., Oster 2019; Cinelli and Hazlett 2020; Diegert et al. 2023; Bernard et al. 2024).
- Consider using methods that do not rely on strict parametric assumptions.

3 | Instrumental-Variable Methods

Instrumental-variable (IV) methods are often used in cases in which selection-on-observables cannot be justified (Lal et al. 2024). We define "IV methods" in a broad sense. While this section focusses on the use of IVs in linear IV and 2-Stage Least Squares (2SLS) regression (which is identical to IV-regression if the number of IVs[12] is equal to the number of endogenous regressors), these discussions and the practical advice given in this section also apply to other estimators that rely on IVs, including machine-learning IV methods (see Appendix A). A brief overview of special types of IVs is presented in Appendix B.

The assumptions required by IV approaches are sophisticated and difficult to test empirically (Lal et al. 2024). However, this does not imply that we want to discourage their use. Rather, our aim is to provide some suggestions and tools on how to

implement credible IV-based identification strategies in empirical research. This is important as invalid IVs can exacerbate the problem so that the bias in the 2SLS estimator may even exceed the OLS endogeneity bias (Lal et al. 2024). By construction, IV estimates are less precise than OLS estimates. For example, Lal et al. (2024) analyse 70 IV designs and show that 2SLS estimates have, on average, six times higher standard errors than OLS estimates, although this decreases with the strength of the IV.[13]

Using an IV approach to estimate a causal effect is possible if one has at least as many IVs as endogenous regressors. These IVs must fulfil the following two criteria: (a) they must be "relevant", i.e., strongly related to the endogenous regressors (even after controlling for all exogenous regressors); and (b) they must be statistically "exogenous", i.e., no direct effect on the outcome variable (exclusion restriction) and not related to the error term (statistical independence).

3.1 | Assessing the Strength of IVs

The first criterion can be empirically investigated with tests for weak IVs. Traditionally, an IV was considered relevant (i.e., not weak) if an F-test of its relevance in the first-stage regression had a test statistic of 10 or higher (Staiger and Stock 1997). However, more recent research indicates that a test statistic of 10 is insufficient in most empirical applications. For instance, Keane and Neal (2024) show that OLS estimates are often closer to the "true" causal effects than 2SLS estimates if the IV's F-statistic in the first stage is below 20. They also demonstrate that, in cases where there is only one IV, the evaluation of the strength of the IV should be based on an F-statistic that exceeds 50. Moreover, estimation results (e.g., t-tests) are often unreliable even in cases where there are much higher values for the F-statistic (e.g., Lee et al. 2022; Keane and Neal 2023, 2024). In addition, Lal et al. (2024) show that first-stage F-statistics are frequently overestimated if the test is not robust towards heteroskedasticity, clustering, and autocorrelation, which implies that IVs in such cases may incorrectly be treated as relevant.

3.2 | Assessing the Exogeneity of IVs

The statistical exogeneity of the IV implies that it influences the dependent variable only via its effect on the endogenous explanatory variable (exclusion restriction) and it is not correlated with the error term (statistical independence). If the endogenous explanatory variable is continuous, the exogeneity of the IVs cannot be empirically investigated without further assumptions (Pearl 1995a, 1995b; Gunsilius 2021).[14] For instance, if there are more potential IVs than endogenous regressors available, it is possible to apply the Sargan-Hansen test/Sargan's J test/Hansen's J test for overidentifying restrictions. If, based on theoretical considerations, it is certain that there are at least as many *exogenous* IVs as there are endogenous regressors, the test indicates (under some assumptions, e.g., correct model specification) whether the additional IVs, that is, those that are not certain to be exogenous, are indeed exogenous. However, without clear theoretical justifications that ascertain the exogeneity of at least as many IVs as there are exogenous regressors, the test is basically uninformative.

In contrast, if the endogenous explanatory variable is discrete, the exogeneity of the IVs can be tested. Pearl (1995a, 1995b) derives testable inequalities, which have been extended by Kitagawa (2015, for binary endogenous explanatory variables and discrete IVs) and Kédagni and Mourifié (2020, for discrete endogenous explanatory variables and any kind of IVs). The intuition behind these inequalities is that for observations with the same value of the endogenous explanatory variable, the potential outcomes should not depend on the value of the IV. However, these inequalities have rarely been used in empirical research.[15]

In addition, it is helpful to think of placebo estimates that can be used to test specific violations of the exogeneity assumption. For instance, in some cases, the effect on the pre-treatment outcome can be estimated (see, e.g., Chabé-Ferret 2025b, section 8.2.1). In other empirical applications, the IV might affect the treatment via a specific mechanism that only matters for some observations (e.g., specific locations, farmers, or crops) but not for others. In this case, a useful placebo test would be to obtain reduced-form estimates of the correlation between the outcome and the IV for a (sub)sample of observations, where the outcome and the IV should be uncorrelated. If the main concern is that the IV might affect the outcome through a specific pathway other than the endogenous regressor, and this potential other pathway is measurable, one can directly test this violation of the exogeneity assumption by regressing this pathway on the IV. For example, if an IV is supposed to affect the farmers' access to credit but is assumed not to affect their access to insurance, one can regress farmers' access to insurance on the IV.

One weakness of all the tests mentioned above is that they can never "prove" that an IV is exogenous because they all rely on the null hypothesis that the IVs are exogenous, and not rejecting the null hypothesis does not necessarily mean that it is true, particularly if the test has low statistical power, for example, caused by a small number of observations, multicollinearity, or a large error variance. Hence, it is always necessary to strongly motivate the exogeneity of IVs based on solid theoretical argumentation (e.g., Lal et al. 2024) and critically discuss the assumption of statistical exogeneity for each IV used, for example, by debating whether potential (unobserved) variables may be related to both the treatment variable and the outcome variable. This is important as McKenzie et al. (2010) show that using IVs for which the exogeneity assumption is potentially violated may lead to the overestimation of the effect of up to 82% compared to the effect found from an experimental benchmark study. This is more than the overestimation that occurs when simply applying OLS (35%), matching (20%) or DID (22%), which implies that a badly identified 2SLS estimation only amplifies estimation bias. As a general rule, the less specific the effects of the chosen IV, the less likely the exogeneity assumption is fulfilled (see, e.g., Mellon (2024) for a discussion of rainfall as an IV).

3.3 | Interpretation and Practical Aspects of 2SLS

In the case of a weak IV or a violation of the exogeneity assumption, an IV estimation can lead to greater bias than an OLS regression (Lal et al. 2024). In such cases, it is advisable to apply non-causal estimators, interpret the results as associations, and draw conclusions with due caution. Here we

refer, for example, to Groher et al. (2020) and Aïhounton and Henningsen (2024) for examples of correlational wording. Lal et al. (2024) note that 2SLS estimates are in many cases much larger than standard OLS estimates, although the aim of the IV estimation is usually to tackle a positive omitted variable bias of OLS. Therefore, it is also advisable to discuss the direction of the bias that the IV estimation is intended to address and assess the extent to which the IV approach was able to address this bias (for examples, see e.g., Basu 2018; Hirsch et al. 2023).

For the interpretation of results, it is important to note that 2SLS estimates indicate Average Treatment Effects (ATE) only under restrictive assumptions (e.g., that the treatment effect is homogeneous across all subjects with the same values of the control variables) (e.g., Heckman 1997; Aronow and Carnegie 2013).[16] However, these assumptions are unlikely to be fulfilled in most empirical analyses. Under less restrictive assumptions (e.g., monotonicity of the effect of the IV on the endogenous explanatory variable), 2SLS estimates indicate Local Average Treatment Effects (LATE), which indicate the effect of the part of the variation in the endogenous explanatory variable that is caused by variation in the IV (e.g., Imbens and Angrist 1994). For instance, in the case of a binary IV and a binary endogenous explanatory variable, the LATE indicates the average treatment effect on those subjects that "comply" with the IV. The effects on the "always takers" and the "never takers" remain unidentified, and in most cases it is unknown who the "compliers" actually are. While the LATE may provide relevant information in some empirical analyses, in others it might not identify the effect we are interested in (Angrist and Pischke 2009; Aronow and Carnegie 2013).

For estimating 2SLS, modern statistical software offers various packages. It is advisable to use these rather than manually estimating 2SLS by first estimating the first-stage OLS and then manually inserting the predicted values into a separately estimated second-stage OLS regression. A common mistake when using the "manual" procedure is failing to include the same control variables in both stages, resulting in inconsistent 2SLS estimates (Angrist and Pischke 2009). Furthermore, the manual procedure results in incorrect OLS standard errors in the second stage. However, unless the IVs are very strong, even the standard errors obtained by software packages for 2SLS estimations do not correctly reflect the uncertainty of 2SLS estimates and, thus, they need to be further adjusted (Lee et al. 2022; Lal et al. 2024). For example, Lal et al. (2024) analyse 70 IV designs and report that the estimated standard errors of 2SLS estimates systematically underestimate the uncertainties of these estimates.

3.4 | Beyond Linear 2SLS

Although the above discussions refer to linear IV and 2SLS regression, they are largely transferable to a large number of other methods that rely on IVs or exclusion restrictions such as endogenous switching regression models or methods for non-continuous dependent or endogenous explanatory variables (see Appendix A for details). It is important to note that additional pitfalls exist when using IVs in regression models with non-linear terms (e.g., quadratic, interaction terms) and/or in non-linear regression models (e.g., probit, logit) (see Appendix A for details).

While the availability of a valid IV is a crucial requirement for obtaining unbiased estimates using any IV approach, it is also crucial to consider the functional form assumption that underlies the employed methods. For instance, Okui et al. (2012) show that 2SLS regression may result in substantially biased estimates of the treatment effect if the functional relationship between the control variables and the outcome variable is incorrectly specified. Interestingly, in applied settings, much of the discussion seems to focus on the validity of the IV, while often the strong functional form assumptions seem to be more readily accepted and less critically discussed. However, depending on the degree of heterogeneity or nonlinearity, they may be equally critical (Okui et al. 2012).

Existing nonparametric versions of IV estimators relax these functional form assumptions and require only that the outcome is the sum of an (unknown) nonlinear function of a treatment variable and observed covariates (that are uncorrelated with unobserved confounders) and an additive error term that may be correlated with the treatment variable (Newey and Powell 2003). However, early nonparametric approaches based on basis functions/splines or kernel methods struggle with a larger number of covariates or IVs and large sample sizes. Building on these early nonparametric estimators, an active field of research at the intersection of machine learning and econometrics has developed extensions that leverage the predictive capabilities of modern machine learning methods to improve nonparametric IV estimators.

Although these new machine learning-based IV approaches offer some interesting extensions of existing approaches, it is important to emphasise that they do not change the requirement of having a solid identification strategy and valid IVs.[17]

Generally, the promise of IV estimation is that it can estimate unbiased effects despite unobserved confounders. However, any IV approach comes at the cost of a substantial reduction in the statistical power of the estimation. This is particularly relevant to consider when estimating heterogeneous treatment effects (given that estimating not just one value but infinitely many or a function of values is a substantially more complex task). Hence, applying IV methods with the aim of identifying treatment heterogeneity typically requires large datasets.

3.5 | Suggestions

If various assessments indicate that an IV-based method should be considered, we suggest performing the following checks that comprise a combination of theory-based considerations and suitable statistical tests (e.g., Lal et al. 2024) (in addition to following the general suggestions that we provide in Section 7):

- If an explanatory variable is incorrectly treated as endogenous, estimates based on IV regression (e.g., 2SLS) are less efficient than estimates based on corresponding selection-on-observables regression methods (e.g.,

OLS). Therefore, it is important to consider and discuss, based on theoretical argumentation, whether a potentially endogenous explanatory variable should indeed be instrumented. In all cases, it is advisable to provide and compare the results for both the IV regression and the OLS estimation.

When using an IV regression method, it is important to assess the strength of the IVs based on the following criteria:

- Always report complete first-stage results including all model diagnostics.
- Only use IV-based methods when the IV(s) are sufficiently strong, that is, there is a sufficiently high correlation between the endogenous explanatory variable and the IV(s) after controlling for exogenous control variables.
- Assess the strength of the IV(s) by applying an F-test to the first stage of the IV estimation that tests the statistical significance of the IVs [i.e., tests a specification with the exogenous control variables but not the IV(s) as explanatory variables against the complete first-stage regression with the IV(s)].
- If the F-statistic of the statistical significance of the IVs in the first stage is below 20, consider presenting OLS estimates instead of 2SLS estimates as OLS estimates are often closer to the "true" causal effects than are 2SLS estimates. In the case of a single IV, the F-statistic should exceed 50 (Keane and Neal 2024).
- If the first-stage F-statistic is below 100, standard errors may need to be adjusted as described by Lee et al. (2022) or Keane and Neal (2024).
- In the case of heteroskedasticity, clustering, or autocorrelation of the error term in the first stage, it is important to conduct an F-test that is robust to these conditions, as a standard F-test overestimates the F-statistic (Lal et al. 2024). See, for example, the Cragg-Donald F-statistic (Cragg and Donald 1993) or the Kleibergen-Paap statistic (Kleibergen and Paap 2006) and the guidance on these statistics provided, for example, in Bazzi and Clemens (2013) or Windmeijer (2024).

We refer to previous parts of this section and the literature (e.g., Lal et al. 2024, section 2.2.1) for a more in-depth discussion of the options to investigate the strength of IVs.

If the IVs are sufficiently strong (so that the use of IV regression is not abandoned), it is important to assess the appropriateness of the exogeneity assumption, that is, the exclusion restriction and the statistical independence assumption. We suggest doing the following:

- Use strong theoretical considerations to rule out any direct effect of the IVs on the dependent variable or any relationship with omitted factors (error term), see, for example, Mellon (2024), who discusses the use of weather as an IV.
- If the endogenous explanatory variable(s) are discrete, use statistical tests to test the exogeneity of the IV(s), for example, the tests suggested by Kitagawa (2015) or Kédagni and Mourifié (2020).
- If possible, use placebo tests to assess the exogeneity assumption (see above).
- In the rare situations when the model is overidentified (i.e., the number of IVs is larger than the number of endogenous explanatory variables) and there are clear theoretical justifications for the exogeneity of at least as many IVs as there are exogenous regressors, use a Sargan-Hansen test/Sargan's J test/Hansen's J test to test the exogeneity of the additional IVs, that is, those that are not certain to be exogenous. However, in most cases, one cannot be certain that there are at least as many exogenous IVs as there are endogenous regressors, and thus the result of this test gives no practical guidance. Furthermore, it is important to note that this test relies on a correct model specification and does not investigate instrument relevance.
- Be aware of the limitations of statistical tests for the exogeneity of IVs, particularly that not rejecting the null hypothesis of exogeneity does not mean that the IVs are exogenous, particularly if the test has low statistical power.

For further discussion on how to assess the exogeneity assumption, we refer to previous parts of this section and the literature (e.g., Lal et al. 2024, section 2.2.2).

If the exogeneity assumption is considered to be appropriate, it is important to carefully assess and interpret the second-stage results and:

- Provide OLS estimates for comparison.
- Discuss whether 2SLS was able to address the bias of OLS estimates, which involves a discussion of the direction of the bias and the extent to which a 2SLS regression can attenuate this bias (see, e.g., Basu 2018).
- Interpret the results as LATE unless there is credible evidence that the chosen method and empirical specification provide an estimate of the ATE.
- Use the tF test (Lee et al. 2022) or the Anderson-Rubin test (Keane and Neal 2024) instead of standard t-tests.

4 | Fixed Effects and Difference in Differences

Fixed-Effects (FE) estimators and Difference-in-Differences (DID) research designs are useful tools to control for unobserved confounders when certain assumptions about these confounders are fulfilled.

4.1 | Fixed-Effects Estimators

FE estimators control for unobserved confounders that are constant at the fixed-effect level. For example, when using individual-fixed effects in a study with panel data, which in agricultural economics papers are often farm-fixed effects, one can control for all time-invariant unobserved heterogeneity at the individual (farm) level. The unobserved heterogeneity

may be differences in management skills, local climatic and soil conditions, infrastructure, or the remoteness of the area. Consequently, models with individual-fixed effects cannot quantify the effects of time-invariant factors such as proximity to a city (Wooldridge 2010). Similarly, FE can be set and combined at every level that reasonably groups the data. For instance, year-fixed effects control for all unobserved heterogeneity that affects all units in a given year in the same way, such as market conditions, the introduction of a certain policy, etc. Mathematically, FE are equal to a joint demeaning of the dependent variable and the independent variables, which is also called *within transformation*. For farm-fixed effects, this implies subtracting the farm average from each observation. This transforms, for example, farm profits into deviations from the average profit of the respective farm in the observed time period (Mundlak 1961). FE may be helpful for controlling for many unobserved factors, and they may also be combined with other methods such as IV or DID. However, there are only a few examples of cases in which FE are sufficient to fully establish causality in a model (Blanc and Schlenker 2017). One example is weather shock impact models that regress a measure of agricultural performance such as yields or productivity on a random and exogenous weather shock (Blanc and Schlenker 2017). Remaining caveats of FE models are related to reverse causality and time-variant confounders, which may still introduce simultaneity and omitted-variable biases.[18]

Taking a closer look at the above examples of time-invariant factors, climatic conditions, soil quality, and infrastructure may be reasonably considered time-invariant in the short run but they may change over longer time horizons. Therefore, Millimet and Bellemare (2023) follow Mundlak (1961, 1978) and argue that such a potential bias may be ignored in shorter panels due to negligible changes in these variables over time. However, in increasingly long panels, a trade-off arises between efficiency gains derived from more observations and potential biases and inconsistency resulting from not truly time-invariant factors accumulating to considerable unobserved confounders over time. Millimet and Bellemare (2023) highlight alternative estimators such as the First-Difference (FD), Twice First-Differenced (TFD), or Interactive Fixed Effects (IFE) estimator, and suggest Rolling FD (RFD), Rolling TFD (RTFD), and Rolling FE (RFE) estimators, which can either be used as alternatives to FE estimators or at least to explore the sensitivity of FE estimates to these alternative estimators.

Table 1 summarises panel data estimators that address unobserved heterogeneity. When using panel data sets with two time periods ($T = 2$), "rolling" estimators cannot be used, while (individual) FE estimates are equal to FD estimates without intercept and Two-Way Fixed Effects (TWFE) estimates are equal to FD estimates with intercept. In case of more than two time periods ($T > 2$) and mostly time-invariant unobserved heterogeneity, FE, TWFE, and FD estimators are recommended. In the presence of autocorrelation in the treatment variable, standard errors that are robust to clustering at the level of the treatment assignment yield valid estimates of the standard errors of any estimator in Table 1 (FD, FE, TWFE, etc.) (Bertrand et al. 2004; Cameron and Miller 2015; Chabé-Ferret 2025b, Chapter 9).[19] However, these estimators

TABLE 1 | Panel data estimators for unobserved heterogeneity ($N \gg T$).

Time T	Unit unobserved heterogeneity	Recommended estimator	Implementation in STATA	Implementation in R
$T = 2$	(Mostly) time-invariant	FE = RFE = FD without intercept; TWFE = FD with intercept	*xtreg, reghdfe, gen + by* (manual)	*fixest, plm*
$2 < T \ll \infty$	(Mostly) time-invariant	FE and TWFE (if no or little autocorrelation), FD (if strong positive autocorrelation)	*xtreg, reghdfe, gen + by* (manual)	*fixest, plm*
	Gradual time-varying	FD, RFD, RFE	*rolling, rangestat, reghdfe* in loop	*plm, fixest + rollapply*
	Rapid time-varying	TFD, RTFD, RFE	*rolling, rangestat, reghdfe* in loop	*rollapply, fixest* in loop
	Linear time-varying	FE and TWFE (with unit-specific trends), IFE for flexible form	*xtreg, reghdfe*	*fixest, phtt*
	Non-linear time-varying	IFE, RTFD	*regife, reghdfe*	*phtt, rollapply*, nonparametric rolling
$T \to \infty$	Strictly time-invariant (rare in practice)	FE, TWFE, FD	*xtreg, reghdfe*	*fixest, plm*
	Mostly time-varying (incl. formerly invariant)	RTFD, TFD, RFE, IFE	*rolling, rangestat, reghdfe* in loop	*rollapply, phtt, fixest* in loop

Note: This requires panel data with observations from units $i \in \{1, \ldots, N\}$ over time periods $t \in \{1, \ldots, T\}$, typically with $N \gg T$. All these recommendations assume homogenous treatment effects.
Abbreviations: FD = first differences; FE = fixed effects; IFE = interactive fixed effects; RFD = rolling FD; RFE = rolling FE; RTFD = rolling TFD; TFD = twice FD; TWFE = two-way fixed effects.

differ in efficiency depending on the degree of autocorrelation (McKenzie 2012). Following McKenzie (2012) and Millimet and Bellemare (2023), we recommend testing for autocorrelation and estimating both FD and FE/TWFE models, retaining the more precise specification. The FD estimator is preferable if there is strong positive temporal autocorrelation in the untransformed error term because the first differencing removes this strong temporal autocorrelation, while FE and TWFE estimators are preferable if there is no or only little temporal autocorrelation in the untransformed error term because, in this case, first differencing would introduce strong negative temporal autocorrelation.

Furthermore, the robustness of estimators that rely on the assumption of time-invariant confounders can be assessed by comparing their estimates to those of estimators that are more robust to time-varying unobserved heterogeneity. If discrepancies arise, Millimet and Bellemare (2023) recommend additionally reporting the results of alternative estimators. In the very long run ($T \to \infty$), most unobserved heterogeneity would change over time, making FE and TWFE unreliable and less relevant (Bellemare and Millimet 2025). In agricultural economics, some outcomes, such as crop or milk yield and farm profits, could exhibit substantial variability and low autocorrelation, making FE or TWFE estimators appropriate (McKenzie 2012). In these settings, using panel data with higher frequencies is often more beneficial than merely extending the time dimension of the panel (McKenzie 2012; Millimet and Bellemare 2023). For example, Belay and Olsen (2025) leverage monthly data to implement TWFE and IFE models in their analysis of milk yield.

4.2 | Difference-in-Differences

An alternative approach to estimating causal effects with panel data is the DID research design.[20] In classic (2×2) DID estimations, there are two groups and two time periods. There is a pre-treatment period, when no units are treated; and there is a post-treatment period, when some units are treated (the treated group) and others (the control group) remain untreated. By using the control group as the counterfactual in the post-treatment period, it is possible to calculate the difference in the changes in the average outcomes between the treatment group and the counterfactual: the "Average Treatment effects on the Treated" (ATT).

The underlying identifying assumption in DID is the parallel-trends assumption, which reasons that the treated units would have followed the same parallel trends as the untreated control units had the treated units gone from the pre-treatment period to the post-treatment period in the absence of treatment.[21] If this assumption is satisfied, then the control units can provide the counterfactual for the treated group in the post-treatment period. However, the parallel-trends assumption is purely hypothetical by definition since it is impossible to be certain that the trends of the treated units and the untreated control units would have followed parallel paths in the post-treatment period. When a data set includes multiple pre-treatment periods, one can verify that the pre-treatment trends of the two groups are parallel, though one should be cautious when inferring "true causality", as parallel trends in the pre-treatment periods may not necessarily imply parallel trends between the last pre-treatment period and the post-treatment period in the hypothetical situation in which the treatment group is not treated.

Multiple applications of DID in agricultural and food economics settings exist. For instance, in production economics, Belay and Jensen (2020) estimate the effect of information disclosure on antibiotic use and market survival among pig farms, while Belay and Ayalew (2020) examine the impact of reference market price disclosure on smallholders' crop choice. Similarly, Belay and Jensen (2022) evaluate the impact of limiting antibiotic use on the economic performance of pig farms. In consumption economics, Fan et al. (2022) estimate the impact of the introduction of a sugar tax on candy purchases and Hoy and Wrenn (2020) estimate the impact of GMO labelling on consumer choices. Other studies, such as Pufahl and Weiss (2009), Chabé-Ferret and Subervie (2013), and Wuepper and Huber (2022) apply DID design to evaluate alternative agri-environmental schemes.

The basic (2×2) DID set-up can be extended to situations with multiple time periods. In DID settings with multiple time periods, a key question is how treatment effects evolve with exposure duration, that is, do they increase, decrease, or remain stable over time? Investigating these dynamics is often the primary reason for using event-study (ES) regressions in DID designs (Callaway and Sant'Anna 2021; Miller 2023). The DID estimations with multiple periods can also be extended to scenarios in which different units of the treatment group receive the treatment at different times, which is known as heterogeneous treatment timing. Under conditions in which the size of the treatment effect is: (a) constant over time; and (b) independent of the time period of the treatment, a standard TWFE estimator offers a reliable estimation for inferring treatment effect causality (Roth et al. 2023).

However, under heterogeneous treatment timing and treatment effect heterogeneity, the TWFE estimator may result in a biased estimate of the average treatment effect on the treated and, thus, causally interpreting the regression coefficient becomes problematic even if the parallel-trends assumption holds (de Chaisemartin and D'Haultfœuille 2020; Goodman-Bacon 2021; Athey and Imbens 2022). For instance, this may be the staggered[22] adoption of an agricultural policy whose effect is time-varying, that is, the magnitude of the effect depends on the time when a farm faced the treatment (e.g., policy) for the first time, the number of years that the farm has already faced the treatment (e.g., due to adjustments, learning, and/or accumulating effects over time), and/or the specific year (e.g., on the weather or market conditions in the year). By making so-called "forbidden comparisons" between groups that received the treatment at earlier and later times, standard DID methods may give negative estimates of the average treatment effect on the treated even when the "true" effect is, in fact, positive, which is known as the *negative weights problem* (Goodman-Bacon 2021; de Chaisemartin and D'Haultfœuille 2023b; Borusyak et al. 2024). Recent developments in DID have identified solutions to this issue. Callaway and Sant'Anna (2021), Sun and Abraham (2021), Wooldridge (2021), de Chaisemartin and D'Haultfœuille (2023a), and Borusyak et al. (2024) have overcome the negative weights problem by restricting the types of comparisons that can be made, ensuring that appropriate counterfactuals are used to

causally infer effects under heterogeneous treatment timing and treatment effect heterogeneity under various conditions of the parallel-trends assumption. For example, one may condition the parallel-trends assumption on additional covariates, such as weather or growing conditions, or on anticipatory behaviour such as in the event of an upcoming policy change (Callaway and Sant'Anna 2021).

Researchers could choose from alternative DID estimators summarised in Table 2, depending on the treatment design, data structure, number of groups, causal estimand of interest (e.g., overall/static ATT or event-study/dynamic ATT), choice of baseline period (in event studies), control group definition and cohort size (in staggered designs), nature of parallel-trends violation, computational speed, and other criteria. The table includes several recent heterogeneity-robust DID estimators for staggered treatment designs. For example, in the context of gradual policy rollout, one can select either the never-treated group or the not-yet-treated group as controls (Callaway and Sant'Anna 2021; de Chaisemartin and D'Haultfœuille 2023a). A researcher can opt for estimators that construct counterfactuals using imputation methods based on efficient and fast linear estimation (Borusyak et al. 2024), two-stage difference in differences (Gardner et al. 2024), or non-linear DID models such as exponential, Poisson, logit, or probit models (Wooldridge 2023). Moreover, heterogeneity-robust DID designs have also been developed for continuous (i.e., non-binary and non-discrete) treatments (Callaway et al. 2024a, 2024b; de Chaisemartin and D'Haultfœuille 2024; de Chaisemartin et al. 2025) as well as for multiple (i.e., reversible and retreatable) (de Chaisemartin and D'Haultfœuille 2024) and several treatments (de Chaisemartin and D'Haultfœuille 2023b). In the case of multiple treatments (sometimes also called treatment-on-and-off scenario), it is important to distinguish between "no-carryover" and "(arbitrary) carryover." In the no-carryover case, only the current treatment status affects outcomes with no lasting impact from past treatment (de Chaisemartin and D'Haultfœuille 2024). In contrast, (arbitrary) carryover means that the treatment history influences outcomes, making it resemble the staggered treatment scenario. In this case, "intent-to-treat" effects can be estimated by defining treatment as "has ever been treated" in a staggered treatment fashion, thereby ensuring that the treatment status is absorbing and accounts for any potential carryover effects (Sun and Abraham 2021; Liu et al. 2024). In many cases, the effect of having previously received the treatment is of interest as it reflects the long-term impact of the treatment, even if the treatment itself is temporary. For instance, Deryugina (2017) studies the fiscal cost for counties hit by hurricanes. Although hurricanes are transitory, their long-term impact persists, so the author models the year of the first hurricane to capture these effects. Deryugina (2017) then adopts what de Chaisemartin and D'Haultfœuille (2023a) refer to as a "binarise and staggerise" approach, that is, by replacing the hurricane status (on/off) with a binary indicator of having been previously hit by a hurricane, the treatment becomes absorbing, allowing the use of staggered adoption designs (Sun and Abraham 2021; de Chaisemartin and D'Haultfœuille 2023a).

It is important to note that the estimation methods recommended for various DID model scenarios in Table 2, along with their implementation in Stata and R, are based on the assumption that the parallel-trends assumption holds unconditionally (i.e., without covariates). Of all the methods listed in Table 2, the method suggested by Callaway and Sant'Anna (2021) is the most suitable for cases where the parallel-trends assumption holds only after conditioning on covariates. This method is applicable for treatments that are both binary and staggered.[23]

When violations of the parallel-trends assumption arise from long-run discrepancies in outcome trends between groups, estimators such as the one suggested by Borusyak et al. (2024), which leverage the full set of pre-treatment periods to construct counterfactuals, can be particularly effective, especially compared to methods that rely solely on the last pre-treatment period as a baseline. However, if the violation stems from a known anticipation effect, approaches such as those suggested by Callaway and Sant'Anna (2021) and Sun and Abraham (2021), can be adapted to use the anticipation period as a baseline to produce reliable estimates (de Chaisemartin and D'Haultfœuille 2023a).

Moreover, DID in an ES framework (including recent generalised DID estimators) offers plots that visually present both dynamic treatment effects and pretreatment trends, allowing the evaluation and testing of parallel trends before treatment (e.g., Taylor 2022; Li and Zhu 2024). However, it is important to note that failure to detect a non-parallel pre-treatment trend does not necessarily imply its absence, as conventional ES tests for parallel pre-treatment trends often lack power and therefore fail to detect non-parallel pre-treatment trends (Freyaldenhoven et al. 2019, 2021; Roth 2022). Researchers should assess the statistical power of these tests using tools such as the R package *pretrends* (Roth 2025) for nonlinear trends and consider alternative visualisation tools such as the *xtevent* package in Stata (Freyaldenhoven et al. 2025), the *eventstudyr* package in R (Freyaldenhoven et al. 2023), or magnitude-based pre-treatment trend evaluation (Bilinski and Hatfield 2020). If the (unconditional) trends during and after the treatment cannot be considered to be parallel (e.g., if pre-treatment trends do not seem to be parallel), Freyaldenhoven et al. (2019) recommend using a 2SLS framework (available in *xtevent* or *eventstudyr*) with one or more covariates that are affected by the confounding (non-parallel) trends but are not related to the treatment. Rambachan and Roth et al. (2023) propose confidence sets that are robust to violations of the parallel-trends assumption, which can be obtained using the *HonestDiD* package (Rambachan and Roth 2024) in R or Stata, as applied by Wuepper and Huber (2022). Regardless of the approach, using economic knowledge to analyse potential parallel-trends violations strengthens causal inferences over relying solely on the statistical significance of tests of parallel pre-treatment trends (Roth 2022). Furthermore, it is important to emphasise that the parallel-trends assumption cannot be tested, as even perfect parallel pre-treatment trends do not guarantee that the trends during and after the treatment period would also be parallel.[24]

An interesting extension to study staggered treatment problems is the matrix completion approach for causal panel data models, which allows the combination of TWFE with synthetic controls in a data-driven manner (Athey et al. 2021). In an agricultural context, this approach is particularly appealing as it naturally

TABLE 2 | Difference-in-differences methods.

Time periods	Treatment design		TE/Estimand	Specific design		Recommended estimator	Implementation in STATA	Implementation in R
Two	Binary	Single treated group	Static (ATT)			(Static) TWFE, AA, SZ	*reghdfe*, *xtreg*, *absdid*, *drdid*	*plm*, *fixest*, *DRDID*
	Continuous					CGS		*contdid*
Multiple [Event Study (ES)]	Binary	Single treated group	Static (Average of ES coefficients: "Overall" ATT)			(Static) TWFE, AA, SZ, AAHIW, Methods in Dynamic/ES ATT	*reghdfe*, *xtreg*, *absdid*, *drdid*, *sdid*, Packages in Dynamic/ES ATT	*plm*, *fixest*, *DRDID*, *synthdid*, Packages in Dynamic/ES ATT
			Dynamic/ ES ATT	Baseline	Average of all pre-treatment periods	BJS, W21, GT, AAHIW	*did_imputation*, *did2s*, *xthdidregress*, *jwdid*, *wooldid*, *sdid_event*	*didimputation*, *did2s*, *etwfe*, *synthdid*
					Last pre-treatment period	(Dynamic/ES) TWFE, CS, DH, SA	*reghdfe*, *eventdd*, *xtevent*, *eventstudyinteract*, *csdid*, *did_multiplegt_dyn*	*plm*, *fixest*, *did*, *DIDmultiplegtDYN*
		Staggered	Dynamic/ ES ATT	Baseline	Last pre-treatment period	CS, DH, SA	*eventstudyinteract*, *csdid*, *did_multiplegt_dyn*	*fixest*, *did*, *DIDmultiplegtDYN*
					Average of all pre-treatment periods	BJS, GT, W21, W23, AAHIW	*did_imputation*, *did2s*, *xthdidregress*, *jwdid*, *wooldid*, *sdid_event*	*didimputation*, *did2s*, *etwfe*, *synthdid*
				Controls	Not-yet-treated group	CS, DH	*csdid*, *did_multiplegt_dyn*	*did*, *DIDmultiplegtDYN*
					Last-treated or never-treated group	CS, SA, AAHIW	*csdid*, *eventstudyinteract*, *sdid_event*	*did*, *fixest*, *synthdid*
					Imputation (regression-based)	BJS, GT, W21, W23	*did_imputation*, *did2s*, *xthdidregress*, *jwdid*, *wooldid*	*didimputation*, *did2s*, *etwfe*
					Imputation (synthetic control-based)	AAHIW	*sdid_event*	*synthdid*

(Continues)

TABLE 2 | (Continued)

Time periods	Treatment design		TE/Estimand	Specific design		Recommended estimator	Implementation in STATA	Implementation in R
				Fast estimation		BJS, GT	*did_imputation*, *did2s*	*didimputation*, *did2s*
				Non-linear estimations		W23	*jwdid*, *wooldid*	*etwfe*
				(Quasi-)random assignment of treatment		RS, AI	*staggered*	*staggered*
		Multiple (Treatment on-and-off)	Static ("Overall") ATT	No carryover		DH	*did_multiplegt_stat*, *did_multiplegt_dyn*	*DIDmultiplegtDYN*, *didmultiplegtstat*
			Dynamic/ ES ATT	(Arbitrary) carryover		Methods in Staggered design, LWX	Packages in Staggered design, *fect*	Packages in Staggered design, *fect*
	Continuous	Single treated group	Static ("Overall") ATT	Stayers	Yes	CGS, DHPSV	*did_multiplegt_stat*	*contdid*, *didmultiplegtstat*
					No	CGS, DHV	*did_had*	*contdid*, *did_had*
			Dynamic/ ES ATT	Baseline	Last pre-treatment period	CGS, DH	*did_multiplegt_dyn*	*contdid*, *DIDmultiplegtDYN*
		Staggered	Dynamic/ ES ATT	Controls	Not-yet treated group	CGS, DH	*did_multiplegt_dyn*	*contdid*, *DIDmultiplegtDYN*
				Baseline	Last pre-treatment period	CGS, DH	*did_multiplegt_dyn*	*contdid*, *DIDmultiplegtDYN*

Note: AA: Abadie (2005); AAHIW: Arkhangelsky et al. (2021); AI: Athey and Imbens (2022); BJS: Borusyak et al. (2024); CGS: Callaway et al. (2024a); CS: Callaway and Sant'Anna (2021); DH: de Chaisemartin and D'Haultfœuille (2024); DHPSV: de Chaisemartin et al. (2025); DHV: de Chaisemartin, D'Haultfœuille, and Vazquez-Bare (2024); GT: Gardner et al. (2024); LWX: Liu et al. (2024); RS: Roth and Sant'Anna (2023); SA: Sun and Abraham (2021); SZ: Sant'Anna and Zhao (2020); TWFE: two-way fixed effects; W21: Wooldridge (2021); W23: Wooldridge (2023). All these recommendations are explicitly made under the assumption that the unconditional parallel-trends assumption is fulfilled (without covariates). However, some of these methods are also suitable in DID settings, in which the parallel-trends assumption only holds when conditioning on covariates (e.g., AA, CS, and SZ).

deals with unbalanced panel data sets (Martinsson et al. 2024). Similarly, Arkhangelsky et al. (2021) developed the Synthetic Difference-in-Differences (SDID) method, which combines elements of the synthetic control approach (discussed below) with DID. While SDID requires a relatively longer pre-treatment period to construct credible counterfactuals, it does more than merely testing for parallel pre-treatment trends using past outcomes (Roth 2022); it leverages the pre-treatment information to compute weights that ensure pre-treatment trends are parallel by construction.

Another relevant impact estimator related to DID design is the Changes-in-Changes (CIC) estimator (Athey and Imbens 2006), which serves as an alternative to DID by focusing on the Quantile Treatment Effect on the Treated (QTT) rather than on the ATT. The QTT approach helps policymakers understand how the benefits and/or costs of a treatment are distributed across subgroups, particularly when decisions depend on distributional effects. For example, Mayr and Agnolucci (2023) apply CIC to estimate heterogeneous impacts of voluntary climate agreements in the UK on business electricity consumption and employment.

In applications of DID in the field of agricultural and applied economics, functional form issues often arise when outcome variables such as shares, areas, incomes, spending, or yields contain many zeros or are non-continuous, such as technology adoption. For outcomes with many zeros, researchers are advised to follow the approaches proposed by Bellemare and Wichman (2020) and Chen and Roth (2024). For non-continuous outcome variables, nonlinear DID specifications such as exponential, logit, or probit models are recommended (Wooldridge 2023).

4.3 | Suggestions

When using fixed-effect-based or DID-based methods, we suggest doing the following (in addition to the general suggestions that we provide in Section 7):

- Provide reasoning based on economic theory on unobserved confounders that potentially bias estimates and that can be addressed by the use of fixed effects.
- Provide reasoning on the time invariance of potential unobserved confounders with respect to the covered time horizon when using individual-fixed effects.
- Select an appropriate estimator to account for unobserved heterogeneity in panel data, and justify the choice with compelling arguments (see, e.g., Table 1).
- When using FE or TWFE estimators (e.g., in case of low temporal autocorrelation), increasing data frequency is more beneficial than lengthening the panel duration.
- Adjust standard errors to make them robust to heteroscedasticity, clustering, and spatial and temporal autocorrelation (if necessary).
- Choose a suitable DID method and substantiate the choice of method by providing convincing arguments (see, e.g., Table 2).
- Evaluate if pre-treatment trends are parallel by creating parallel-trends plots in static DID analyses and event-study plots in dynamic DID settings.
- Empirically investigate the extent to which pre-treatment trends are parallel in DID settings. This investigation should include supplementing event-study plots with diagnostic tests that assess the statistical power of tests for parallel pre-treatment trends.
- Consider using methods such as those suggested by Abadie (2005), Sant'Anna and Zhao (2020), and Callaway and Sant'Anna (2021) in DID settings where the parallel-trends assumption only holds when conditioning on covariates. However, Freyaldenhoven et al. (2019) emphasise that this conditioning approach may often be inadequate in real-world economic applications because it requires the conditioning covariate to be a perfect proxy for the confounding trend—an assumption that may not always hold. To address this, Freyaldenhoven et al. (2019) propose a generalised 2SLS framework for an event study that allows conditioning on covariates that are not necessarily perfect proxies for the confounding trend.
- In simple DID (i.e., classic (2×2) or multiple-period single-treated), if the unconditional pre-treatment trends are not parallel and the researcher wishes to specifically control for lagged outcome due to potential policy/program anticipation effects (Ashenfelter 1978; Ashenfelter and Card 1985; Heckman and Smith 1999), selecting the appropriate estimator requires testing for unit roots and assessing parallel pre-treatment trends (Chabé-Ferret 2025a). If there is a unit root in the outcome and pre-treatment trends are parallel only when conditioning on lagged outcome, use DID with lagged outcomes. If neither unconditional nor conditional pre-treatment trends are parallel, choose the method with the lower pre-treatment bias (Chabé-Ferret 2025a).
- Provide reasoning based on economic theory on parallel post-treatment trends in DID settings.
- In DID with staggered treatment, consider using the Bacon Decomposition to explicitly diagnose and interpret static TWFE estimates as a weighted average of all possible pairwise 2×2 DID comparisons (Goodman-Bacon 2021). This decomposition can be conducted with or without time-varying covariates, and implemented using the *bacondecomp* package in Stata (Goodman-Bacon et al. 2022) or R (Flack and Jee 2020). Alternatively, one can use the *twowayfeweights* package in Stata (de Chaisemartin, D'Haultfoeuille, and Deeb 2024) or the *TwoWayFEWeights* package in R (Ciccia et al. 2024).
- Consider supplementing DID estimates using falsification or placebo tests on outcomes arguably unrelated to the treatment/intervention.
- In DID applications with many zeros or non-continuous outcome variables, researchers should follow Bellemare and Wichman (2020) and Chen and Roth (2024) for zero-inflated outcomes and use nonlinear DID models such as exponential, Poisson, logit, or probit specifications (Wooldridge 2023).
- If SUTVA violations are plausible (e.g., due to spillovers) in DID or other panel-data settings, see Butts (2023) and

Chabé-Ferret et al. (2021) for methods to test and relax this assumption, including machine-learning approaches for detecting social networks in panel data (Manresa 2013).

5 | Synthetic Control Method

The Synthetic Control Method (SCM) was introduced by Abadie and Gardeazabal (2003) and later formalised by Abadie et al. (2010, 2015). SCM is a combination of DID and matching. Synthetic control units are selected as the weighted average of all potential comparison units based on how closely they resemble the treated unit(s) in the pre-treatment periods (Abadie 2021). According to Athey and Imbens (2017), "the synthetic control approach [...] is arguably the most important innovation in the policy evaluation literature in the last 15 years." However, despite the increasing availability of long panel datasets, this method has not yet been widely applied in agricultural economics. The few examples for the use of this method in agricultural economics include, for example, Grogger (2017), Mohan (2017), Opatrny (2020) and Kim (2023).

5.1 | Prerequisites for Using SCM

SCM is particularly valuable when the parallel-trends assumption required for DID does not hold, provided that sufficiently long pre-treatment panel data are available. It is especially well-suited for evaluating the impact of interventions affecting a single or small number of large units such as cities, regions, or countries, making it a useful tool in agricultural and applied economics, where national or state-level agricultural, food, and environmental policies can be assessed by constructing a synthetic control group of comparable nations or states. For example, Grogger (2017) estimates the impact of the soda tax implemented in Mexico in 2014 on soda prices by comparing them to those of other untaxed non-substitute goods, creating a synthetic control group. Alternatively, researchers can construct the synthetic control using soda price data from other countries not subject to the tax, offering another way to estimate the causal effect of the policy.

Furthermore, a balanced panel data set must be available that includes the treated unit(s) and a reasonably large number of potential comparison units ("donor pool"), while it includes a reasonably large number of pre-treatment periods and at least one post-treatment period.[25] Although the SCM usually cannot give unbiased estimates of the treatment effect, Abadie et al. (2010) show that—under certain assumptions—the bias is bounded and approaches zero with an increasing number of pre-treatment periods. Hence, it is important to have a sufficiently large number of pre-treatment units.

One of the most basic assumptions of the standard SCM is that the data generating process corresponds to a "factor model" (Abadie et al. 2010, equation 1), which assumes, for example, that unobserved differences between units are constant over time and that the effects of observed and unobserved differences between units on the outcome are identical across all units (but these effects can change over time). Thus, empirical applications must clearly discuss the appropriateness of these assumptions, for example, if the treatment could potentially affect the effects of observed and unobserved variables on the outcome so that these effects differ between the treatment unit and the control unit in the post-treatment period. Furthermore, in order to avoid overfitting, the number of potential comparison units should not be too large, which can be achieved by restricting potential comparison units to those that are sufficiently similar to the treatment unit (Abadie et al. 2015). Recent studies further clarified the theoretical links between SCM and latent factor models, highlighting both its strengths and limitations. Specifically, Gobillon and Magnac (2016) demonstrate that SCM is a special case of interactive FE estimators, while Liu et al. (2024) unify SCM, interactive fixed effects, and other counterfactual estimators within a common framework for panel data analysis.

5.2 | Criticisms of SCM

Although SCM enhances transparency by revealing each comparison unit's contribution to the counterfactual and, thus, enables clear interpretation and expert evaluation of potential biases, researcher discretion in selecting donor pools, predictors, and weights may raise concerns about robustness and replicability. These concerns can be mitigated by designing the study (e.g., selecting donor units and predictors) without access to post-treatment data, thereby reducing risks of specification searches and *p*-value hacking (Abadie 2021). Furthermore, pre-registration of synthetic control weights before the intervention, similar to pre-registration of RCTs, further strengthens transparency and credibility (Abadie 2021).

5.3 | Suggestions

When using SCM, we suggest doing the following (in addition to the general suggestions provided in Section 7):

- Clearly state the assumptions that the chosen SCM requires and discuss how credible these assumptions are in the presented empirical analysis
- Make sure that there is a sufficiently long pre-treatment period
- Ensure that there is a sufficiently large but not too large number of comparison units.
- Visualise the SCM estimation results using graphs.
- Present the contributions of each unit to the synthetic control.
- Conduct inference using the permutation method (Abadie et al. 2015).
- To support the internal validity of causality using SCM, researchers are advised to conduct validity tests, such as leaving out units of the donor pool (with non-zero weights), placebo tests, using fake treatment dates, and other outcomes not related to the treatment.
- When appropriate and beneficial for reliability, consider combining SCM with DID, using SDID (Arkhangelsky et al. 2021).
- Note that: (a) inference in SCM is limited to the data used to construct the synthetic control; (b) SCM does not allow

predictions or inferences outside the range of the observed data; and (c) extrapolation or generalisation outside the supporting data and context is invalid

- For better transparency, replicability, and credibility, researchers are recommended to preregister synthetic control weights before the analysis or the intervention, or select donor units and predictors without access to post-treatment data (Abadie 2021)

6 | Regression Discontinuity and Difference-in-Discontinuity Designs

Regression Discontinuity Designs (RDDs) and Difference-in-Discontinuity Designs (DiDDs) can be set up in multiple ways (as discussed below and in Wuepper and Finger 2023, in more detail) but they all share a particular mechanism for identifying causal effects: if treatment assignment is triggered by a clearly-defined threshold in a continuously distributed variable,[26] then—given a few falsifiable assumptions—discontinuity in the outcome right at this threshold quantifies the treatment effect (Thistlewaite and Campbell 1960; Imbens and Lemieux 2008). Intuitively, this works especially well with arbitrarily set thresholds because this minimises the risk that, besides the treatment assignment, something else "jumps" exactly at the threshold. Another important condition is that observations (usually people) cannot choose which side of the threshold they are on (e.g., if it is well known that a subsidy is available to farms below a certain size, farmers whose farms are just above the threshold may be able to take measures that ensure that their farms fall just below, which might make the treatment endogenous).

6.1 | Regression Discontinuity Design

The fundamental requirement for RDDs is the existence of a continuously distributed variable that has a threshold which triggers treatment assignment.[27] For instance, public extension services may only visit farms within an arbitrarily defined maximum distance-to-branch (Pan et al. 2018), and governments might target villages with an anti-poverty programme if they are above an arbitrarily defined poverty threshold (Alix-Garcia et al. 2013). Also, geographical borders can be used such as historical borders within a country (Noack et al. 2022), or national borders dividing countries (Wuepper, Borrelli, and Finger 2020; Wuepper, Le Clech, et al. 2020). When geographic borders are being used, the most general treatment one can define is "belonging to one side of the border or the other." For example, one might ask how much agricultural or environmental outcomes are simply the result of an area belonging to one country and not another (see, e.g., figure 1 or Wuepper, Borrelli, and Finger 2020). When the border triggers mainly one specific mechanism, one might also be able to focus more narrowly on this mechanism directly. For example, Noack et al. (2022) use the historical border between East and West Germany to identify the effect of agricultural structures (small-scale vs. large-scale farming) on bird diversity, and Gupta et al. (2024) use Indian state borders to identify the negative impact of language barriers on the effectiveness of agricultural extension services. Sometimes the treatment is introduced spatially with a clear boundary, for example, in the case of protected areas (Neal 2024) or World Heritage sites (Rodríguez et al. 2025). In this case, the effect of "belonging to one side and not the other" is a narrow treatment in and of itself.

The most intuitive way to understand how a national border can be used to identify the effect of an area belonging to one country but not another is provided in Figure 1. This figure is based on data from Wuepper et al. (2023). Their starting point is to quantify for each of many years how much countries matter for local crop yields. Here, we only focus on two countries: Vietnam and Cambodia. The border can be divided into small segments (panel a), and crop yields can be quantified in high resolution from satellite imagery (panel b) (Wuepper et al. 2025). When computing local averages of crop yields at equal distances from the border and plotting these as a function of border distance, a pattern emerges: whereas crop yield is distributed rather smoothly on either side of the border, there is a stark jump at the border (panel c), which cannot be explained by potential confounders such as rainfall or sunshine because these do *not* jump at the border: it is the countries as political constructs that make the fields in Vietnam more productive than those in Cambodia (Wuepper et al. 2023). The most important assumption here is that no potential confounding factors also show a discontinuity right at the border. For example, if this border was located right on top of a natural barrier such as a major mountain range, the sudden change in agricultural conditions could also explain a jump in crop yields. This can be tested, for example, by replacing the outcome variable, in this case crop yields, with elevation, rainfall, temperature, or sunshine, which would reveal whether these are also discontinuously distributed. Wuepper et al. (2023) analyse first the role of the institutions of

FIGURE 1 | (a) The border between Cambodia and Vietnam separates an otherwise comparable agricultural area into two countries. Colours distinguish different border segments. (b) Satellite data can be used to obtain a methodologically unified, high-resolution crop yield measure. (c) An important step: Before the actual RDD is estimated, the data should be plotted, so that it is possible to visually inspect whether the discontinuity that is to be estimated is visible. It is usually helpful to aggregate the data points in small bins and fit regression lines separately on both sides of the threshold. The actual RDD estimates the size of the discontinuity at the threshold. *Sources*: Wuepper et al. (2023) (a + c), Google Earth (b).

these countries in differences in crop yields, and second how much agricultural technology (mechanisation and irrigation) is the channel. For these further analyses, they move on to panel data, as discussed in the following section.

6.2 | Difference-in-Discontinuity Design and Regression Discontinuity in Time

An increasingly popular research design is the DiDD, which is a combination of RDD and DID. It is set up like a standard DID design with the only difference being that it focusses on the change in a discontinuity from before to after treatment. This built-in extra step improves the chance of a valid parallel-trends assumption because the estimated discontinuity already helps to avoid confounding factors as discussed above. In the best-case scenario, a researcher finds a situation in which the threshold is newly created at some point in time (e.g., an existing state is split into two), which means that demonstrating that there was no discontinuity prior to treatment is straightforward, and afterwards the discontinuity shows the causal treatment effect (Garg and Shenoy 2021). Alternatively, in the study by Wuepper et al. (2023), the leveraged country borders do not change, but they show that the discontinuities in crop yields are stable before treatment and change in response to countries' institutional changes.

Finally, Regression Discontinuity in Time (RDiT) tackles endogeneity by examining a narrow time window around the implementation of a policy, where time is used as the running variable and the treatment date acts as the threshold.[28] This approach assumes that unobserved factors remain similar within the window, which allows pre-treatment observations to be used as a comparison for post-treatment observations. RDiT utilises flexible polynomial time trends and has been recently used in studies involving so-called "sin taxes", sugar and fat taxes, air quality, fisheries, and food safety (Hausman and Rapson 2018; Bovay 2025). The growing availability of high-frequency data further enhances its utility for researchers evaluating national agricultural and environmental policies and interventions.

6.3 | Assessing the Discontinuity

For the research designs discussed above, simple procedures can be followed. These include performing various tests and analytics in a chronological order, which allows readers to easily follow and judge the credibility of the analysis (Wuepper and Finger 2023). This procedure is facilitated by off-the-shelf software packages, especially the Python, R, and Stata packages provided by Calonico et al. (2015) and Calonico et al. (2017).[29] The two main assumptions of RDD are exogenous thresholds and no endogenous sorting. The simplest way of examining the assumption of no endogenous sorting is to look for bunching near the threshold (McCrary 2008). The simple logic is that if there is a striking dip in observations on one side of the threshold, and these "missing" observations all bunch together on the other side of the threshold, it is likely that it is the result of optimising behaviour (e.g., if a regulation that only applies to farms above 5 ha was introduced, farmers who initially had 5.2 ha quickly got rid of 0.3 ha).

6.4 | Technical Aspects

There are a few important technical aspects to consider. First, for any kind of discontinuity analysis, one needs to restrict the dataset to observations within an "optimal" bandwidth near the cut-off (Cattaneo and Titiunik 2022). This can have an important impact on the estimates as it involves a variance-bias trade-off. The cleanest comparison is possible just next to the threshold (assuming the absence of spillovers). However, using only observations that are directly at the threshold will make the sample size small and specific; keeping only one observation on each side of the threshold would even make it impossible to fit a regression line. Thus, in order to obtain precise and meaningful estimates, one must allow for some maximum distance to the border, while still having two sides that are sufficiently comparable to each other. Over the years, various algorithms have been developed that aim to find the statistically optimal bandwidth (Wuepper and Finger 2023). It is generally a good idea to demonstrate the sensitivity of one's findings to small (or large) deviations from the chosen bandwidth. For example, if the running variable is farm size and the optimal bandwidth (e.g., according to the Mean Squared Error) is 30 ha, it is good to additionally report the findings for a bandwidth of 25 and 35 ha. Second, in addition to choosing the optimal bandwidth, one must decide how to fit the regression to the observations. The simplest approach is to use a linear regression with a dummy variable identifying the threshold and then two variables reflecting the continuous running variable, separately on each side of the threshold. A more sophisticated way to do it is to use local polynomial functions (Cattaneo and Titiunik 2022). These can be based on a continuity assumption as discussed above, that is, a smooth distribution of potential outcomes across the threshold; or a local randomisation assumption similar to common experimental set-ups, that is, potential outcomes are statistically the same on either side of the threshold (Cattaneo and Titiunik 2022; Wuepper and Finger 2023). A limitation of the local polynomial approach is its relative complexity and computational demand compared to a linear regression framework. For example, in the local polynomial framework, it is not straightforward how to handle panel data with fixed effects. Furthermore, with increasingly available, very large datasets, such as high-resolution satellite data (Wuepper et al. 2025), the simpler linear regression approach is clearly faster than the local polynomial approach.

6.5 | Suggestions

When using discontinuity-based methods, we suggest doing the following (in addition to following the general suggestions that we provide in Section 7):

- Clearly describe the running variable, explain the reason for a discontinuity at the threshold, and discuss all variables that might discontinuously "jump" at the threshold, especially potential confounders.
- Visually assess the discontinuity (or the change in discontinuity) and the data distributions around the discontinuity.
- Conduct placebo tests to probe the exogeneity of the threshold (see, e.g., Wuepper and Finger 2023).

- Use alternative algorithms to compute the optimal statistical bandwidth for robustness checks.
- Test for endogenous sorting across the threshold (McCrary 2008).
- In a discrete running variable with mass points, consider local randomisation or redefining the running variable by aggregating observations at the mass points to handle the discreteness (Cattaneo and Titiunik 2022).
- Discuss the generalisability of the results as the effects are identified very locally at the threshold. Sometimes, a threshold might be found in unrepresentative places (e.g., places with especially high or low agricultural productivity) or among unrepresentative units (e.g., among farms that are especially large or small).

7 | General Suggestions

In addition to the method-specific guidelines provided in previous sections of this paper, we suggest doing the following irrespective of the chosen method:

- Before pursuing causal inference, it is important to determine whether the question at hand concerns the "effect of a cause" or the "cause of an effect." Plausible and policy-relevant causal inference can typically be made only in relation to the "effect of a cause," not the "cause of an effect." The latter is only meaningful to the extent that it helps identify which cause to study when estimating the "effect of a cause" (Gelman and Imbens 2013)
- As the mere use of big data sets, by itself, does not resolve causal identification challenges, avoid the "big data fallacy", where a large sample size is erroneously perceived as a substitute for a proper identification strategy (Vosgerau et al. 2025).
- Start from the theoretical understanding of the problem (e.g., based on a DAG) to define an identification strategy and clearly discuss the assumptions under which the quantity of interest is identified, any potential explanations for the assumptions being violated and their consequences for identification.
- Before jumping into the econometric analyses, begin by examining the data, which should involve the computation of various descriptive statistics and plotting and mapping the data in various ways. For instance, for panel data, it is important to understand how the data varies over units and over time. Fully understanding the data is essential for making appropriate data preparation and modelling choices.
- Carefully consider the assumptions of the chosen estimation approach(es). Consider the extent to which these assumptions fit the theoretically motivated identification strategy and the data at hand.
- Clearly point out the added value of the chosen method compared to simpler approaches such as OLS. Unless added value can be clearly demonstrated, a simpler method may be preferable.
- Check if the methods used require a "common support" and, if they do, the extent to which the common support assumption is fulfilled, that is, the treated units indeed have comparable counterparts in the control group (see, e.g., Heckman et al. 1998)[30]
- Discuss the plausibility of the SUTVA in the specific empirical analysis. Under this assumption, the potential outcomes of each observation only depend on the treatment of this observation and not on the treatment of other observations. All methods discussed in previous sections require this assumption unless spillovers between observations are explicitly and appropriately accounted for in the empirical analysis.
- It can be informative to simulate artificial data sets with known properties before using actual data to perform an empirical analysis. These properties may include the functional form of the analysed relationship, the magnitude of the treatment effect and its heterogeneity between observations, correlations between observed variables and between observed and unobserved variables, potential endogeneity issues, validity of the exogeneity assumption and IV strength (in the case of an IV-based method), the degree of autocorrelation of observed and unobserved variables (in the case of panel data and/or the use of lagged variables), deviations from independently and identically distributed (i.i.d.) error terms (e.g., heteroscedasticity, clustering), and other assumptions. Use these data sets to test the estimation approach (as well as the code used to implement it). Test the conditions under which the estimation approach succeeds in recovering the effects used to create the artificial data. Using artificial data to test the code/inference is an integral part of the data-generating-process centric workflow (McElreath 2018; Gelman et al. 2020; Storm et al. 2024).
- If feasible, consider using multiple identification approaches and critically discuss what can be learnt from the different estimates, as they are based on different assumptions and have different advantages and disadvantages. Recent textbooks on causal inference, such as Cunningham (2021) and Huntington-Klein (2025), provide more detailed information about and code examples for several of the methods mentioned in this paper, which are helpful sources of information for robustness checks and sensitivity analyses.

8 | Conclusions

We do not recommend one particular method over another as the most suitable method is case-dependent. Therefore, our aim is to provide clear guidelines that should be followed when applying these methods.

Even if these guidelines for investigating causal effects with observational data are followed, there is always uncertainty about whether all the required assumptions are completely fulfilled. Therefore, one should be very careful when using causal language such as "the effect of A on B", "the impact of A on B", "A affects B", "A reduces B", "A increases B", or "A leads to a change in B". As a precaution, one could use statements about associations

such as "A is positively related to B", "A is negatively related to B", "A is associated with B", or "A is conditionally associated with B". In any case, it is important to use consistent language throughout the entire paper.[31] If causal statements are made, it is crucial to clearly point out that these statements are conditional on the appropriateness of the identifying assumptions, the model specification implemented, and the data used for estimation.

Finally, it is important to acknowledge that our article cannot provide an exhaustive overview of all available approaches for causal identification in applied economics research, as the field continues to advance and benefit from ongoing methodological developments. One particularly active area of development is mediation analysis (e.g., Imai, Keele, and Tingley 2010; Imai, Keele, and Yamamoto 2010; Deuchert et al. 2019; Chabé-Ferret 2025b, section 15). Another noteworthy line of research, inspired by the classical work of LaLonde (1986), involves assessing the reliability of observational methods (including those discussed in this article) in estimating causal effects (e.g., Glazerman et al. 2003; Chaplin et al. 2018; Gordon et al. 2019; Gechter 2024; Bernard et al. 2024).

Acknowledgements

The authors would like to thank Marc F. Bellemare, Irene Villalba, and three anonymous reviewers for their feedback on earlier versions of this manuscript and Jonathan Roth and Pedro Sant'Anna for their valuable feedback on DID methods. Arne Henningsen extends his gratitude to Marc F. Bellemare and Fabio G. Santeramo for their invaluable discussions and inspiration, especially during the joint preparation and teaching of PhD courses on the topic of this paper. Arne Henningsen is also grateful for inspiring discussions with other members of the research project "CliFT—Climate-Smart Futures in Rural Tanzania" and to the Ministry of Foreign Affairs of Denmark for financially supporting this project (Grant: 23-04-KU). David Wuepper and Hugo Storm acknowledge the financial support provided by the German Research Foundation (DFG) Cluster of Excellence "PhenoRob" (EXC 2070 Grant No. 390732324). Guy Low and Tobias Dalhaus acknowledge the financial support provided by the European Union's Horizon 2020 research and innovation programme (grant agreement No. 862357 (MIXED)); this work only reflects the authors' views; the Commission is not responsible for any use of the information contained in this article. Dagim Belay acknowledges the financial support provided by Innovation Fund Denmark (grant No. JPIAMR2021-182 SEFASI) under the umbrella of the Joint Programming Initiative on Antimicrobial Resistance (JPIAMR) and the Independent Research Fund Denmark (DFF) (grant No. 2029-00021B). Of course, the authors take full responsibility for any remaining errors.

Funding

This work was supported by Ministry of Foreign Affairs of Denmark (Grant: 23-04-KU), German Research Foundation (DFG) (EXC 2070 Grant No. 390732324), European Union (Horizon 2020 research and innovation programme, grant agreement No. 862357), Innovation Fund Denmark (grant No. JPIAMR2021-182 SEFASI), Independent Research Fund Denmark (DFF) (grant No. 2029-00021B).

Ethics Statement

The authors have nothing to report.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Endnotes

[1] More general factors undermining the credibility of published results include publication bias, insufficient sample size, insufficient standardisation of variable definitions across studies, and various types of statistical malpractice such as p-value hacking or exploratory research that is incorrectly framed as confirmatory research, a practice known as Hypothesising After the Results are Known (HARKing). All of these factors and practices can lead to biased and less replicable results and misleading conclusions. See Ioannidis and Doucouliagos (2013) for a more detailed discussion of these drivers of a loss of credibility.

[2] While recent research documents that the uptake of scientific evidence by policymakers may be selective (see, for example, Vivalt and Coville (2023) who show that policymakers update their beliefs more strongly on positive evaluation outcomes than on negative evaluation outcomes, or Rao (2025) who shows that policy decisions are only in very specific situations related to results of programme evaluations), we are unaware of studies that investigate the use of results from economics research by other decision makers. However, given that governments and other stakeholders frequently spend money on independent research studies and that private businesses can benefit from reliable and unbiased scientific evidence, we contend that many decision makers are genuinely committed to evidence-based decision making. For example, in many resource-constrained settings, credible estimates of programme effectiveness can help guide more efficient resource allocation and support better-targeted interventions. Even where evidence uptake may be delayed or selective, credible (causal) impact evaluation solidifies the evidence base and can inform policy discourse and decisions.

[3] In addition, the external validity of the results should be outlined and discussed, for example whether the results that are based on a specific group of economic agents such as farmers or consumers in a specific region or country may also be valid for other groups of economic agents such as farmers or consumers in other regions or countries. However, the discussion of external validity is outside the scope of this paper, which is limited to the issue of internal validity.

[4] See, for example, El Benni et al. (2025), who provide a review of studies using experimental methods to assess how nudge interventions influence farmers' behaviour.

[5] Even in the relatively rare cases in which experimental methods can be applied, their results often have important limitations. For example, RCTs are usually restricted to narrow cases, the results are rarely directly generalisable, and there are often additional complications, such as non-compliance with the treatment or uncontrollable external influences. Furthermore, it may be hard to prevent the non-treated group from becoming informed about the treatment of the intervention group (Buchanan-Smith et al. 2016; Koppenberg et al. 2023). In addition, it is difficult to identify the mechanisms behind the cause-effect interplay (Quisumbing et al. 2020; Koppenberg et al. 2023; Todd and Wolpin 2023). Even when using experiments, only relationships with randomised variables can indicate causal effects, while relationships with non-randomised variables (e.g., personal characteristics) usually cannot be interpreted as causal effects (see, e.g., Nigus et al. 2024). Deaton (2010) and Bulte et al. (2020) provide overviews of the limitations of RCTs.

[6] In this paper, we focus on the endogeneity of explanatory variables. However, all other assumptions that are required to obtain unbiased and/or consistent estimates should also be fulfilled and discussed when presenting econometric analyses. For instance, the functional form used in the econometric analysis should resemble the relationship between the explanatory variables and the dependent variable in the population. Furthermore, the observations used for the estimation should be a random sample of the relevant population, while deviations from random sampling, e.g., non-proportional stratified random sampling, should be appropriately addressed in the econometric analysis. Furthermore, what the used data actually measure and what the results really imply should also be correctly interpreted (Gibson 2019).

[7] Regarding the interpretation of the coefficients of covariates, see Westreich and Greenland (2013).

[8] Several online and offline software tools for visualising and analysing DAGs exist. One of these tools is the open-source software DAGitty (https://www.dagitty.net/).

[9] It is important to note that a DAG indicates whether a causal effect is *non-parametrically* identified, that is, the identification does not rely on parametric assumptions, for example, about the functional form of the modelled relationships or the distribution of the error term. Even when using parametric empirical methods, in most cases it is desirable to identify causal effects non-parametrically so that approximately reliable results are obtained if parametric assumptions are not 100% fulfilled.

[10] This estimator is called 'doubly-robust' as only one of its two regression equations needs to be correctly specified, but otherwise this estimator requires the same identification assumptions as an OLS regression (e.g., Kurz 2022, equation 1).

[11] King and Nielsen (2019) point out that "propensity scores should not be used for matching" and that other matching methods are more suitable than PSM.

[12] In this paper, we use the narrow definition of IVs, that is, we only consider the variables that are used to explain the endogenous regressor but that are not used to explain the outcome variable as IVs. The broad definition of IVs additionally includes the variables that are used to explain the outcome variable because these variables are also used to explain the endogenous regressor.

[13] See figure 3 of Lal et al. (2024). Lal et al. (2024) also point out that this makes 2SLS estimations more susceptible to p-value hacking and publication bias.

[14] Some empirical researchers (e.g., Acemoglu et al. 2001) aim to test the exogeneity of IVs by estimating the outcome equation with both the endogenous regressor and the IV (and of course all relevant control variables). If the IV affects the dependent variable only through the endogenous regressor, the coefficient of the IV in this auxiliary regression should be close to zero. However, if the endogenous explanatory variable is indeed endogenous, the coefficient of this variable and the coefficient of the IV are not jointly identified (Conley et al. 2012). Hence, this auxiliary regression does not provide useful information.

[15] The 'falsification test' for IV-regression with a binary endogenous treatment variable, originally applied by Di Falco et al. (2011), has been used in several empirical studies in the agricultural and applied economics literature. However, this test is invalid because it relies exclusively on untreated units, thereby introducing sample-selection bias. If the IV is valid, meaning it is independent of the potential outcomes with treatment and without treatment, it does not remain independent of the (observed) outcome without treatment when conditioning on the sample of untreated units. This is because, if the IV is relevant, conditioning on untreated units induces a correlation between the IV and unobserved factors that affect the treatment assignment and potentially also the outcome. We illustrate this with a simple treatment assignment model. A unit i is treated if $\alpha_0 + \alpha' x_i + \gamma z_i + \varepsilon_i > 0$, where x_i is a vector of control variables, z_i is the IV, and ε_i is the error term capturing the influence of unobserved factors on treatment assignment. Restricting the sample to untreated units implies: $\varepsilon_i \leq -\alpha_0 - \alpha' x_i - \gamma z_i$, which means the expected value of ε_i becomes a function of z_i (unless the IV is irrelevant, i.e., $\gamma = 0$). As a result, the IV z_i becomes correlated with the error term in the outcome equation, violating the exclusion restriction (unless the selection-on-observables assumption holds, i.e., there are no unobserved factors that affect both treatment and outcome so that ε_i is uncorrelated with the error term in the outcome equation). The authors thank an anonymous reviewer for pointing this out.

[16] Aronow and Carnegie (2013) suggest a method that requires either homogeneity of the treatment effect or homogeneity of compliance (i.e., IVs have the same effect on the treatment assignment across all observations).

[17] Appendix A provides a more detailed discussion of machine learning IV methods.

[18] While fixed effects help to control for biases arising from unobserved time-invariant confounders, common issues in fixed-effect applications are temporal and spatial correlation, clustering, and heteroscedasticity in the error term. The standard approach to dealing with this is to obtain standard errors that are robust to these deviations from independently and identically distributed errors (see, e.g., Low et al. 2025, for an example).

[19] Making standard errors robust to clustering accounts for autocorrelation in the treatment variable but not for autocorrelation in the outcome variable, which may require dynamic specifications for consistent estimation (Arellano and Bond 1991; Arellano and Bover 1995; Chabé-Ferret 2025a, 2025b).

[20] It is important to note that DID is a *research design*, while FE, TWFE, FD, etc. are *estimation methods*. Depending on the data structure and assumptions about the data generating process, different estimators are suitable for DID research designs.

[21] In certain cases, a simple DID design may not yield reliable causal inference. For instance, if a policy targets farmers younger than 40 years in a specific state, comparing this group of farmers to either farmers aged 40–49 years in the same state or to farmers younger than 40 years in other states may lead to biased estimates because it does not account for age-related or state-specific trends, respectively. To address this, a triple-DID estimator uses differences in three dimensions (state, age group, and time) to isolate the causal effect of the policy change. The triple DID estimator, which can also be calculated as the difference between two DID estimators, may only require one parallel-trends assumption as long as the bias is the same in both estimators, in which case the bias cancels out when differenced (Olden and Møen 2022).

[22] Staggered treatment is a setting where different units adopt/implement the treatment at different times with no reversal to the unit's treatment status, that is, if a unit is treated once, it remains always treated (Callaway and Sant'Anna 2021).

[23] For classic (2×2) and multiple-period (single treated group) DID, see Chabé-Ferret (2015) and Chabé-Ferret (2025a) for tests and guidance on when to condition on lagged outcomes. In particular, Chabé-Ferret (2025a) provides a detailed practical checklist to support these decisions.

[24] Although parallel pre-treatment trends are neither necessary nor sufficient for obtaining unbiased estimates, it is highly recommended to test for parallel pre-treatment trends because if there are parallel pre-treatment trends, it is more likely that the parallel-trends assumption is fulfilled, and if pre-treatment trends are not parallel, it is less likely that the parallel-trends assumption is fulfilled.

[25] Some Generalised SCM methods can also be applied to unbalanced panel data, for example, the method implemented in the R package *gsynth* (Xu and Liu 2021) that adds the capability to use unbalanced panel data to the method suggested by Xu (2017).

[26] Under certain conditions, it is also possible to apply the RDD framework if the running variable is discrete (e.g., food safety inspection score based on restaurant hygiene inspections). However, the empirical analysis must take into account the discreteness of the running variable. Details are available, for example, in Kolesár and Rothe (2018) and Cattaneo and Titiunik (2022), while software packages such as *rdhonest* for Stata (Armstrong et al. 2023) or *RDHonest* for R (Kolesár 2025) can be used.

[27] The threshold does not have to deterministically trigger the treatment as it does in the standard model. If the threshold only changes the probability of treatment, one moves from the sharp RDD to the fuzzy RDD, which involves estimating an IV regression such as 2SLS with the threshold as the IV.

[28] RDiT is related to Interrupted Time Series (ITS), which is another method that also leverages temporal variation. However, RDiT requires discontinuity at the cut-off, bandwidth selection, and strong RDD assumptions that can be empirically tested (e.g., via the density test suggested by McCrary 2008). While ITS can identify changes in a trend without these requirements, it typically requires longer time series and lacks formal tests for violations of key identifying assumptions, such as manipulation or anticipation.

[29] All available at: https://rdpackages.github.io/rdrobust/.

[30] The concept of common support is based on a binary treatment variable, but similar criteria can be made for continuous explanatory variables of interest. For instance, when investigating the effect of farm size and farm size is strongly correlated with confounders such as farm type, soil characteristics, climate, and ownership type, there may not be farms available in the data that differ only in size while sharing similar values of other characteristics.

[31] One minor exception to this rule would be to write that a study "aims to estimate the effect of A on B", to explain why the estimates may not indicate causal effects, and to interpret all estimates as conditional associations (as done in, for example, Aïhounton and Henningsen 2024).

References

Abadie, A. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72, no. 1: 1–19.

Abadie, A. 2021. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature* 59, no. 2: 391–425.

Abadie, A., A. Diamond, and J. Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105, no. 490: 493–505.

Abadie, A., A. Diamond, and J. Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59, no. 2: 495–510.

Abadie, A., and J. Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93, no. 1: 113–132.

Acemoglu, D., S. Johnson, and J. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91, no. 5: 1369–1401.

Aïhounton, G., and A. Henningsen. 2024. "Does Organic Farming Jeopardize Food Security of Farm Households in Benin?" *Food Policy* 124: 102622.

Alix-Garcia, J., C. McIntosh, K. R. Sims, and J. R. Welch. 2013. "The Ecological Footprint of Poverty Alleviation: Evidence From Mexico's Oportunidades Program." *Review of Economics and Statistics* 95, no. 2: 417–435.

Amemiya, T. 1974. "The Nonlinear Two-Stage Least-Squares Estimator." *Journal of Econometrics* 2, no. 2: 105–110.

Angrist, J. 2014. "The Perils of Peer Effects." *Labour Economics* 30: 98–108.

Angrist, J., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, J., and J.-S. Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24, no. 2: 3–30.

Arellano, M., and S. Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58: 277–297.

Arellano, M., and O. Bover. 1995. "Another Look at the Instrumental Variables Estimation of Error Components Models." *Journal of Econometrics* 68: 29–51.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. 2021. "Synthetic Difference-in-Differences." *American Economic Review* 111, no. 12: 4088–4118.

Armstrong, T., M. Kolesár, Y. Chen, S. Zhang, and K.-H. Lee. 2023. "rdhonest: Honest Confidence Intervals for Regression Discontinuity Designs." Stata Package. Accessed July 2025. Development Version. https://raw.githubusercontent.com/tbarmstr/RDHonest-vStata/master/current/.

Aronow, P. M., and A. Carnegie. 2013. "Beyond Late: Estimation of the Average Treatment Effect With an Instrumental Variable." *Political Analysis* 21, no. 4: 492–506.

Ashenfelter, O. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 60, no. 1: 47–57.

Ashenfelter, O., and D. Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics* 67, no. 4: 648–660.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. 2021. "Matrix Completion Methods for Causal Panel Data Models." *Journal of the American Statistical Association* 116, no. 536: 1716–1730.

Athey, S., and G. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31, no. 2: 3–32.

Athey, S., and G. W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-In-Differences Models." *Econometrica* 74, no. 2: 431–497.

Athey, S., and G. W. Imbens. 2022. "Design-Based Analysis in Difference-In-Differences Settings With Staggered Adoption." *Journal of Econometrics* 226, no. 1: 62–79.

Athey, S., J. Tibshirani, and S. Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47, no. 2: 1148–1178.

Auci, S., N. Barbieri, M. Coromaldi, and M. Michetti. 2021. "Climate Variability, Innovation and Firm Performance: Evidence From the European Agricultural Sector." *European Review of Agricultural Economics* 48, no. 5: 1074–1108.

Barrett, C. B., and M. R. Carter. 2010. "The Power and Pitfalls of Experiments in Development Economics: Some Non-Random Reflections." *Applied Economic Perspectives and Policy* 32, no. 4: 515–548.

Bartik, T. 1991. *Who Benefits From State and Local Economic Development Policies?* W.E. Upjohn Institute for Employment Research.

Basu, D.. 2018. "When Can We Determine the Direction of Omitted Variable Bias of OLS Estimators?" Technical Report No. 2018-16. Department of Economics, University of Massachusetts, Amherst, MA.

Baylis, K., T. Heckelei, and H. Storm. 2021. "Machine Learning in Agricultural Economics (Chapter 83)." In *Handbook of Agricultural Economics*, edited by C. B. Barrett and D. R. Just, vol. 5, 4551–4612. Elsevier.

Bazzi, S., and M. A. Clemens. 2013. "Blunt Instruments: Avoiding Common Pitfalls in Identifying the Causes of Economic Growth." *American Economic Journal: Macroeconomics* 5, no. 2: 152–186.

Belay, D. G., and H. Ayalew. 2020. "Nudging Farmers in Crop Choice Using Price Information: Evidence From Ethiopian Commodity Exchange." *Agricultural Economics* 51, no. 5: 793–808.

Belay, D. G., and J. D. Jensen. 2020. "'The Scarlet Letters': Information Disclosure and Self-Regulation: Evidence From Antibiotic Use in Denmark." *Journal of Environmental Economics and Management* 104: 102385.

Belay, D. G., and J. D. Jensen. 2022. "Quantitative Input Restriction and Farmers' Economic Performance: Evidence From Denmark's Yellow Card Initiative on Antibiotics." *Journal of Agricultural Economics* 73, no. 1: 155–171.

Belay, D., and J. V. Olsen. 2025. "Economic Impacts of Salmonella Dublin in Dairy Farms: Panel Evidence From Denmark." *Agricultural Economics* 56, no. 4: 666–693.

Bellemare, M. 2012. "The 'Credibility Revolution' in Economics: Agricultural and Applied Economists, Take Note." Keynote Lecture at the Annual Meeting of the SCC-76 'Economics and Management of Risk in Agriculture and Natural Resources' Group, Pensacola Beach (FL), USA. Accessed December 2024. https://marcfbellemare.com/wordpress/wp-content/uploads/2012/03/BellemareSCCKeynote.pdf.

Bellemare, M. 2015. "Metrics Monday: What to Do With Endogenous Control Variables?" Blog Post, Accessed December 2024. https://marcfbellemare.com/wordpress/11057.

Bellemare, M., J. Bloem, and N. Wexler. 2024. "The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion." *Oxford Bulletin of Economics and Statistics* 86, no. 4: 951–993.

Bellemare, M., T. Masaki, and T. Pepinsky. 2017. "Lagged Explanatory Variables and the Estimation of Causal Effect." *Journal of Politics* 79, no. 3: 949–963.

Bellemare, M. F., and D. L. Millimet. 2025. "Retrospectives: Yair Mundlak and the Fixed Effects Estimator." *Journal of Economic Perspectives* 39, no. 2: 261–274.

Bellemare, M., and L. Novak. 2017. "Contract Farming and Food Security." *American Journal of Agricultural Economics* 99: 357–378.

Bellemare, M. F., and C. J. Wichman. 2020. "Elasticities and the Inverse Hyperbolic Sine Transformation." *Oxford Bulletin of Economics and Statistics* 82, no. 1: 50–61.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen. 2012. "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain." *Econometrica* 80: 2369–2429.

Bernard, D., G. Bryan, S. Chabé-Ferret, J. de Quidt, J. Fliegner, and R. Rathelot. 2024. "How Much Should We Trust Observational Estimates? Accumulating Evidence Using RCTs With Imperfect Compliance." Technical Report 18794. CEPR Press. Accessed September 2025. https://cepr.org/publications/dp18794.

Bertrand, M., E. Duflo, and S. Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics* 119, no. 1: 249–275.

Betz, T., S. Cook, and F. Hollenbach. 2018. "On the Use and Abuse of Spatial Instruments." *Political Analysis* 26, no. 4: 474–479.

Bilinski, A., and L. A. Hatfield. 2020. "Nothing to See Here? Non-Inferiority Approaches to Parallel Trends and Other Model Assumptions." arXiv. Accessed December 2024. https://arxiv.org/abs/1805.03273.

Blanc, E., and W. Schlenker. 2017. "The Use of Panel Models in Assessments of Climate Impacts on Agriculture." *Review of Environmental Economics and Policy* 11, no. 2: 258–279.

Blattman, C. 2010. "The Cardinal Sin of Matching." Blog Post. Accessed December 2024. https://chrisblattman.com/blog/2010/10/27/the-cardinal-sin-of-matching/.

Blundell, R., and S. Bond. 1998. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics* 87: 11–143.

Borusyak, K., P. Hull, and X. Jaravel. 2025. "A Practical Guide to Shift-Share Instruments." *Journal of Economic Perspectives* 39: 181–204.

Borusyak, K., X. Jaravel, and J. Spiess. 2024. "Revisiting Event Study Designs: Robust and Efficient Estimation." *Review of Economic Studies* 91, no. 6: 3253–3285.

Bovay, J. 2025. "Shaming, Stringency, and Shirking: Evidence From Food-Safety Inspections." *American Journal of Agricultural Economics* 107, no. 1: 152–180.

Breuer, M. 2022. "Bartik Instruments: An Applied Introduction." *Journal of Financial Reporting* 7, no. 1: 49–67.

Brignoli, P. L., Y. de Mey, and C. Gardebroek. 2024. "Everything Under Control: Comparing Machine Learning and Classical Econometric Impact Assessment Methods Using FADN Data." *European Review of Agricultural Economics* 51, no. 5: 1410–1441.

Buchanan-Smith, M., J. Cosgrave, and A. Warner. 2016. "Evaluation of Humanitarian Action Guide." ALNAP (Active Learning Network for Accountability and Performance in Humanitarian Action).

Bulte, E., G. Beekman, S. Di Falco, J. Hella, and P. Lei. 2014. "Behavioral Responses and the Impact of New Agricultural Technologies: Evidence From a Double-Blind Field Experiment in Tanzania." *American Journal of Agricultural Economics* 96, no. 3: 813–830.

Bulte, E., S. Di Falco, and R. Lensink. 2020. "Randomized Interventions and 'Real' Treatment Effects: A Cautionary Tale and an Example." *World Development* 127: 104790.

Butts, K. 2023. "Difference-in-Differences Estimation With Spatial Spillovers." Accessed September 2025. arXiv. https://arxiv.org/abs/2105.03737.

Caetano, C., G. Caetano, L. Goff, and E. Nielsen. 2025. "Identification of Causal Effects With a Bunching Design." Accessed November 2025. arXiv. https://arxiv.org/pdf/2507.05210.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna. 2024a. *Difference-in-Differences With a Continuous Treatment*. NBER Working Paper 32117, National Bureau of Economic Research.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna. 2024b. "Event Studies With a Continuous Treatment." *AEA Papers and Proceedings* 114: 601–605.

Callaway, B., and P. Sant'Anna. 2021. "Difference-in-Differences With Multiple Time Periods." *Journal of Econometrics* 225, no. 2: 200–230.

Calonico, S., M. Cattaneo, M. Farrell, and R. Titiunik. 2017. "Rdrobust: Software for Regression-Discontinuity Designs." *Stata Journal: Promoting Communications on Statistics and Stata* 17, no. 2: 372–404.

Calonico, S., M. Cattaneo, and R. Titiunik. 2015. "Rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs." *R Journal* 7, no. 1: 38.

Cameron, C. A., and D. L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50, no. 2: 317–372.

Cattaneo, M. D., and R. Titiunik. 2022. "Regression Discontinuity Designs." *Annual Review of Economics* 14, no. 1: 821–851.

Chabé-Ferret, S. 2015. "Analysis of the Bias of Matching and Difference-In-Difference Under Alternative Earnings and Selection Processes." *Journal of Econometrics* 185, no. 1: 110–123.

Chabé-Ferret, S. 2025a. "Should We Combine Difference in Differences With Conditioning on Pre-Treatment Outcomes?" Acccessed September 2025. https://drive.google.com/file/d/1JjHRdca65yCDyKtrf13S-23kRQuubVix/view.

Chabé-Ferret, S. 2025b. "Statistical Tools for Causal Inference." Accessed September 2025. https://chabefer.github.io/STCI/.

Chabé-Ferret, S., A. Reynaud, and E. Tène. 2021. "Water Quality, Policy Diffusion Effects and Farmers' Behavior." Accessed September 2025. https://drive.google.com/file/d/1wJYHWkfBz9LII_9yPAXdl-ppXpVd7dr8/view.

Chabé-Ferret, S., and J. Subervie. 2013. "How Much Green for the Buck? Estimating Additional and Windfall Effects of French Agro-Environmental Schemes by DID-Matching." *Journal of Environmental Economics and Management* 65, no. 1: 12–27.

Chaplin, D. D., T. D. Cook, J. Zurovac, et al. 2018. "The Internal and External Validity of the Regression Discontinuity Design: A Meta-Analysis of 15 Within-Study Comparisons." *Journal of Policy Analysis and Management* 37, no. 2: 403–429.

Chen, J., and J. Roth. 2024. "Logs With Zeros? Some Problems and Solutions." *Quarterly Journal of Economics* 139, no. 2: 891–936.

Chernozhukov, V., D. Chetverikov, M. Demirer, et al. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21, no. 1: C1–C68.

Ciccia, D., F. Knau, M. Malezieux, D. Sow, S. Zhang, and C. de Chaisemartin. 2024. "TwoWayFEWeights: Estimation of the Weights Attached to the Two-Way Fixed Effects Regressions." R Package Version 2.0.4. Accessed July 2025. https://CRAN.R-project.org/package=TwoWayFEWeights.

Cinelli, C., and C. Hazlett. 2020. "Making Sense of Sensitivity: Extending Omitted Variable Bias." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 82, no. 1: 39–67.

Conley, T. G., C. B. Hansen, and P. E. Rossi. 2012. "Plausibly Exogenous." *Review of Economics and Statistics* 94, no. 1: 260–272.

Cragg, J., and S. Donald. 1993. "Testing Identifiability and Specification in Instrumental Variable Models." *Econometric Theory* 9: 222–240.

Cunningham, S. 2021. *Causal Inference: The Mixtape*. Yale University Press.

de Chaisemartin, C., X. D'Haultfoeuille, and A. Deeb. 2024. "TWOWAYFEWEIGHTS: Stata Module to Estimate the Weights and Measure of Robustness to Treatment Effect Heterogeneity Attached to Two-Way Fixed Effects Regressions." Statistical Software Components, Boston College Department of Economics. Accessed July 2025. https://EconPapers.repec.org/RePEc:boc:bocode:s458611.

de Chaisemartin, C., X. d'Haultfoeuille, F. Pasquier, D. Sow, and G. Vazquez-Bare. 2025. "Difference-in-Differences for Continuous Treatments and Instruments With Stayers."

de Chaisemartin, C., and X. D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators With Heterogeneous Treatment Effects." *American Economic Review* 110, no. 9: 2964–2996.

de Chaisemartin, C., and X. D'Haultfœuille. 2023a. "Two-Way Fixed Effects and Difference-in-Differences Estimators With Several Treatments." *Journal of Econometrics* 236, no. 2: 105480.

de Chaisemartin, C., and X. D'Haultfœuille. 2023b. "Two-Way Fixed Effects and Differences-in-Differences With Heterogeneous Treatment Effects: A Survey." *Econometrics Journal* 26, no. 3: C1–C30.

de Chaisemartin, C., and X. D'Haultfœuille. 2024. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." *Review of Economics and Statistics* 1: 1–45.

de Chaisemartin, C., X. D'Haultfœuille, and G. Vazquez-Bare. 2024. "Difference-in-Difference Estimators With Continuous Treatments and no Stayers." *AEA Papers and Proceedings* 114: 610–613.

Deaton, A. 2010. "Instruments, Randomization, and Learning About Development." *Journal of Economic Literature* 48, no. 2: 424–455.

Deines, J., K. Guan, B. Lopez, et al. 2023. "Recent Cover Crop Adoption Is Associated With Small Maize and Soybean Yield Losses in the United States." *Global Change Biology* 29: 794–807.

Deines, J., S. Wang, and D. Lobell. 2019. "Satellites Reveal a Small Positive Yield Effect From Conservation Tillage Across the US Corn Belt." *Environmental Research Letters* 14, no. 12: 124038.

Deryugina, T. 2017. "The Fiscal Cost of Hurricanes: Disaster Aid Versus Social Insurance." *American Economic Journal: Economic Policy* 9, no. 3: 168–198.

Deuchert, E., M. Huber, and M. Schelker. 2019. "Direct and Indirect Effects Based on Difference-In-Differences With an Application to Political Preferences Following the Vietnam Draft Lottery." *Journal of Business & Economic Statistics* 37, no. 4: 710–720.

Di Falco, S., M. Veronesi, and M. Yesuf. 2011. "Does Adaptation to Climate Change Provide Food Security? A Micro-Perspective From Ethiopia." *American Journal of Agricultural Economics* 93: 825–842.

Diegert, P., M. Masten, and A. Poirier. 2023. "Assessing Omitted Variable Bias When the Controls Are Endogenous." arXiv. Accessed December 2024. https://arxiv.org/abs/2206.02303.

El Benni, N., Y. Wang, J. Ammann, G. Mack, and R. Finger. 2025. "A Scoping Review of Nudges for Enhancing Agricultural Policy." *Applied Economic Perspectives and Policy*.

Fan, L., A. Stevens, and B. Thomas. 2022. "Consumer Purchasing Response to Mandatory Genetically Engineered Labeling." *Food Policy* 110: 102296.

Finger, R., C. Grebitus, and A. Henningsen. 2023. "Replications in Agricultural Economics." *Applied Economic Perspectives and Policy* 45, no. 3: 1258–1274.

Flack, E., and E. Jee. 2020. "bacondecomp: Goodman-Bacon Decomposition." R Package Version 0.1.1. Accessed July 2025. https://CRAN.R-project.org/package=bacondecomp.

Freyaldenhoven, S., C. Hansen, J. P. Pérez, et al. 2023. "eventstudyr Package." Accessed July 2025. https://github.com/JMSLab/eventstudyr.

Freyaldenhoven, S., C. B. Hansen, J. P. Pérez, J. M. Shapiro, and C. Carreto. 2025. "Xtevent: Estimation and Visualization in the Linear Panel Event-Study Design." *Stata Journal* 25, no. 1: 97–135.

Freyaldenhoven, S., C. Hansen, J. Pérez Pérez, and J. M. Shapiro. 2021. *Visualization, Identification, and Estimation in the Linear Panel Event-Study Design. NBER Working Papers 29170*. National Bureau of Economic Research.

Freyaldenhoven, S., C. Hansen, and J. M. Shapiro. 2019. "Pre-Event Trends in the Panel Event-Study Design." *American Economic Review* 109, no. 9: 3307–3338.

Frölich, M. 2008. "Parametric and Nonparametric Regression in the Presence of Endogenous Control Variables." *International Statistical Review* 76: 214–227.

Gardner, J., N. Thakral, L. T. Tô, and L. Yap. 2024. "Two-Stage Differences in Differences." Working Paper. Accessed December 2024. https://neilthakral.github.io/files/papers/2sdd.pdf.

Garg, T., and A. Shenoy. 2021. "The Ecological Impact of Place-Based Economic Policies." *American Journal of Agricultural Economics* 103, no. 4: 1239–1250.

Gechter, M. 2024. "Combining Experimental and Observational Studies in Meta-Analysis: A Debiasing Approach." Accessed September 2025. https://www.dropbox.com/scl/fi/yqa5xevi5hf72lb274apm/MGRM_Combining_Experimental_and_Observational_Studies.pdf?rlkey=jwhtduzk2ixoz9svdc3jo665c&e=1&dl=0.

Gelman, A., and G. Imbens. 2013. "Why Ask Why? Forward Causal Inference and Reverse Causal Questions." National Bureau of Economic Research Working Paper 19614. Accessed July 2025. https://www.nber.org/papers/w19614.

Gelman, A., A. Vehtari, D. Simpson, et al. 2020. "Bayesian Workflow." arXiv. Accessed November 2025. https://doi.org/10.48550/arXiv.2011.01808.

Gibson, J. 2019. "Are You Estimating the Right Thing? An Editor Reflects." *Applied Economic Perspectives and Policy* 41, no. 3: 329–350.

Glazerman, S., D. M. Levy, and D. Myers. 2003. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *Annals of the American Academy of Political and Social Science* 589, no. 1: 63–93.

Gobillon, L., and T. Magnac. 2016. "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls." *Review of Economics and Statistics* 98, no. 3: 535–551.

Gollin, D., C. W. Hansen, and A. M. Wingender. 2021. "Two Blades of Grass: The Impact of the Green Revolution." *Journal of Political Economy* 129, no. 8: 2344–2384.

Goodman-Bacon, A. 2021. "Difference-in-Differences With Variation in Treatment Timing." *Journal of Econometrics* 225, no. 2: 254–277.

Goodman-Bacon, A., T. Goldring, and A. Nichols. 2022. "BACONDECOMP: Stata Module to Perform a Bacon Decomposition of Difference-in-Differences Estimation." Statistical Software Components, Boston College Department of Economics. Accessed July 2025. https://EconPapers.repec.org/RePEc:boc:bocode:s458676.

Gordon, B. R., F. Zettelmeyer, N. Bhargava, and D. Chapsky. 2019. "A Comparison of Approaches to Advertising Measurement: Evidence From Big Field Experiments at Facebook." *Marketing Science* 38, no. 2: 193–225.

Grogger, J. 2017. "Soda Taxes and the Prices of Sodas and Other Drinks: Evidence From Mexico." *American Journal of Agricultural Economics* 99, no. 2: 481–498.

Groher, T., K. Heitkämper, A. Walter, F. Liebisch, and C. Umstätter. 2020. "Status Quo of Adoption of Precision Agriculture Enabling Technologies in Swiss Plant Production." *Precision Agriculture* 21: 1327–1350.

Gunsilius, F. F. 2021. "Nontestability of Instrument Validity Under Continuous Treatments." *Biometrika* 108, no. 4: 989–995.

Gupta, A., J. Ponticelli, and A. Tesei. 2024. "Language Barriers, Technology Adoption and Productivity: Evidence From Agriculture in India." *Review of Economics and Statistics*, forthcoming.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy. 2017. "Deep IV: A Flexible Approach for Counterfactual Prediction." In *Proceedings of the 34th International Conference on Machine Learning*, edited by D. Precup and Y. W. Teh. PMLR.

Hausman, C., and D. S. Rapson. 2018. "Regression Discontinuity in Time: Considerations for Empirical Applications." *Annual Review of Resource Economics* 10: 533–552.

Hausman, J. 1996. "Valuation of New Goods Under Perfect and Imperfect Competition." In *The Economics of New Goods*, edited by T. Bresnahan and R. Gordon. University of Chicago Press.

Heckman, J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5: 475–492.

Heckman, J. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources* 32, no. 3: 441–462.

Heckman, J., H. Ichimura, J. Smith, and P. Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66, no. 5: 1017–1098.

Heckman, J. J., and J. A. Smith. 1999. "The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies." *Economic Journal* 109, no. 457: 313–348.

Hirsch, S., M. Khalilov, T. Dalhaus, and A. Mishra. 2023. "Firm Names and Profitability in German Food Processing." *European Review of Agricultural Economics* 50: 1103–1139.

Hirsch, S., G. Tiboldo, and R. Lopez. 2018. "A Tale of Two Italian Cities: Brand-Level Milk Demand and Price Competition." *Applied Economics* 50, no. 49: 5239–5252.

Hoy, K., and D. Wrenn. 2020. "The Effectiveness of Taxes in Decreasing Candy Purchases." *Food Policy* 97: 101959.

Hünermund, P., and E. Bareinboim. 2025. "Causal Inference and Data Fusion in Econometrics." *Econometrics Journal* 28, no. 1: 41–82.

Huntington-Klein, N. 2025. *The Effect: An Introduction to Research Design and Causality*. Chapman and Hall/CRC.

Imai, K., L. Keele, and D. Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15, no. 4: 309–334.

Imai, K., L. Keele, and T. Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science* 25, no. 1: 51–71.

Imbens, G. 2024. "Causal Inference in the Social Sciences." *Annual Review of Statistics and Its Application* 11: 123–152.

Imbens, G. W. 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58, no. 4: 1129–1179.

Imbens, G., and J. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62: 467–475.

Imbens, G. W., and T. Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142, no. 2: 615–635.

Imbens, G. W., and J. M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47, no. 1: 5–86.

Ioannidis, J., and C. Doucouliagos. 2013. "What's to Know About the Credibility of Empirical Economics?" *Journal of Economic Surveys* 27, no. 5: 997–1004.

Jafari, Y., M. Koppenberg, S. Hirsch, and T. Heckelei. 2023. "Markups and Export Behavior: Firm-Level Evidence From the French Food Processing Industry." *American Journal of Agricultural Economics* 105, no. 1: 174–194.

Jiang, W. 2017. "Have Instrumental Variables Brought Us Closer to the Truth." *Review of Corporate Finance Studies* 6, no. 2: 127–140.

Keane, M., and T. Neal. 2023. "Instrument Strength in IV Estimation and Inference: A Guide to Theory and Practice." *Journal of Econometrics* 235, no. 2: 1625–1653.

Keane, M., and T. Neal. 2024. "A Practical Guide to Weak Instruments." *Annual Review of Economics* 16: 185–212.

Kédagni, D., and I. Mourifié. 2020. "Generalized Instrumental Inequalities: Testing the Instrumental Variable Independence Assumption." *Biometrika* 107, no. 3: 661–675.

Kim, Y. 2023. *Payments for Ecosystem Services Programs and Climate Change Adaptation in Agriculture. Technical Report*. Department of Economics, University of Oxford.

King, G., and R. Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27, no. 4: 435–454.

Kitagawa, T. 2015. "A Test for Instrument Validity." *Econometrica* 83, no. 5: 2043–2063.

Kleibergen, F., and R. Paap. 2006. "Generalized Reduced Rank Tests Using the Singular Value Decomposition." *Journal of Econometrics* 133, no. 1: 97–126.

Kolesár, M. 2025. "RDHonest: Honest Inference in Regression Discontinuity Designs." R Package Version 1.0.1.9000. Accessed July 2025. https://github.com/kolesarm/rdhonest.

Kolesár, M., and C. Rothe. 2018. "Inference in Regression Discontinuity Designs With a Discrete Running Variable." *American Economic Review* 108, no. 8: 2277–2304.

Koppenberg, M., A. Mishra, and S. Hirsch. 2023. "Food Aid and Violent Conflict: A Review and Empiricist's Companion." *Food Policy* 121: 102542.

Krishnan, P., and M. Patnam. 2014. "Neighbors and Extension Agents in Ethiopia: Who Matters More for Technology Adoption?" *American Journal of Agricultural Economics* 96, no. 1: 308–327.

Kurz, C. 2022. "Augmented Inverse Probability Weighting and the Double Robustness Property." *Medical Decision Making* 42, no. 2: 156–167.

Lal, A., M. Lockhart, Y. Xu, and Z. Zu. 2024. "How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on 67 Replicated Studies." *Political Analysis* 32, no. 4: 521–540.

LaLonde, R. J. 1986. "Evaluating the Econometric Evaluations of Training Programs With Experimental Data." *American Economic Review* 76, no. 4: 604–620.

Lee, D., J. McCrary, M. Moreira, and P. Porter. 2022. "Valid t-Ratio Inference for IV." *American Economic Review* 112, no. 10: 3260–3290.

Li, H., and J. Zhu. 2024. "Property Rights and Land Quality." *American Journal of Agricultural Economics* 106, no. 5: 1619–1647.

Liu, L., Y. Wang, and Y. Xu. 2024. "A Practical Guide to Counterfactual Estimators for Causal Inference With Time-Series Cross-Sectional Data." *American Journal of Political Science* 68, no. 1: 160–176.

Low, G., M. P. Meuwissen, R. Finger, and T. Dalhaus. 2025. "Frost Shocks Negatively Impact the Supply of High-Stem Juicing Apples in Switzerland." *Agricultural Systems* 228: 104363.

Maggio, G., M. Mastrorillo, and N. Sitko. 2022. "Adapting to High Temperatures: Effect of Farm Practices and Their Adoption Duration on Total Value of Crop Production in Uganda." *American Journal of Agricultural Economics* 104: 385–403.

Magnan, N., D. Spielman, T. Lybbert, and K. Gulati. 2015. "Leveling With Friends: Social Networks and Indian Farmers' Demand for a Technology With Heterogeneous Benefits." *Journal of Development Economics* 116: 223–251.

Manresa, E. 2013. "Estimating the Structure of Social Interactions Using Panel Data." Accessed September 2025. https://igier.unibocconi.eu/sites/default/files/media/attach/JMP_EM_1st_Dec-Manresa20140117095117.pdf.

Martinsson, E., H. Hansson, K. Mittenzwei, and H. Storm. 2024. "Evaluating Environmental Effects of Adopting Automatic Milking Systems on Norwegian Dairy Farms." *European Review of Agricultural Economics* 51, no. 1: 128–156.

Mason, N., T. Jayne, and R. Mofya-Mukuka. 2013. "Zambia's Input Subsidy Programs." *Agricultural Economics* 44, no. 6: 613–628.

Mayr, K. F., and P. Agnolucci. 2023. "Heterogeneous Impacts in Voluntary Agreements: A Changes-In-Changes Approach to the UK Climate Change Agreements." *Environmental and Resource Economics* 86, no. 3: 345–379.

McCrary, J. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142, no. 2: 698–714.

McElreath, R. 2018. *Statistical Rethinking: A Bayesian Course With Examples in R and Stan*. Chapman and Hall/CRC.

McKenzie, D. 2012. "Beyond Baseline and Follow-Up: The Case for More T in Experiments." *Journal of Development Economics* 99, no. 2: 210–221.

McKenzie, D. 2018. *I'm Not a Fan of Leave-One-Out/Spatial Instruments*. World Bank Blogs. Accessed December 2024. https://blogs.worldbank.org/en/impactevaluations/im-not-fan-leave-one-outspatial-instruments.

McKenzie, D., J. Gibson, and S. Stillman. 2010. "How Important Is Selection? Experimental vs. Non-Experimental Measures of the Income Gains From Migration." *Journal of the European Economic Association* 8, no. 4: 913–945.

Mellon, J. 2024. "Rain, Rain, Go Away: 194 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable." *American Journal of Political Science* 69, no. 3: 881–898.

Miller, D. L. 2023. "An Introductory Guide to Event Study Models." *Journal of Economic Perspectives* 37, no. 2: 203–230.

Millimet, D., and M. Bellemare. 2023. "Fixed Effects and Causal Inference." Technical Report No. 16202, IZA Discussion Paper.

Mohan, P. 2017. "The Economic Impact of Hurricanes on Bananas: A Case Study of Dominica Using Synthetic Control Methods." *Food Policy* 68: 21–30.

Morgan, S., and C. Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

Mullally, C., and S. Chakravarty. 2018. "Are Matching Funds for Smallholder Irrigation Money Well Spent?" *Food Policy* 76: 70–80.

Mundlak, Y. 1961. "Empirical Production Function Free of Management Bias." *Journal of Farm Economics* 43, no. 1: 44–56.

Mundlak, Y. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46, no. 1: 69–85.

Neal, T. 2024. "Estimating the Effectiveness of Forest Protection Using Regression Discontinuity." *Journal of Environmental Economics and Management* 127: 103021.

Nevo, A. 2000. "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand." *Journal of Economics and Management Strategy* 9, no. 4: 513–548.

Nevo, A. 2001. "Measuring Market Power in the Ready-to-Eat Cereal Industry." *Econometrica* 69, no. 2: 307–342.

Newey, W., and J. Powell. 2003. "Instrumental Variable Estimation of Nonparametric Models." *Econometrica* 71: 1565–1578.

Nigus, H. Y., P. Mohnen, and E. Nillesen. 2024. "Market Experience and Agricultural Technology Adoption: The Role of Risk Aversion and Locus of Control." *European Review of Agricultural Economics* 51, no. 5: 1312–1347.

Noack, F., A. Larsen, J. Kamp, and C. Levers. 2022. "A Bird's Eye View of Farm Size and Biodiversity: The Ecological Legacy of the Iron Curtain." *American Journal of Agricultural Economics* 104, no. 4: 1460–1484.

Okui, R., D. Small, Z. Tan, and J. Robins. 2012. "Doubly Robust Instrumental Variable Regression." *Statistica Sinica* 22: 173–205.

Olden, A., and J. Møen. 2022. "The Triple Difference Estimator." *Econometrics Journal* 25, no. 3: 531–553.

Opatrny, M. 2020. "The Impact of EU Accession on Farm Production in The Czech Republic: A Synthetic Control Method Approach." *Czech Journal of Economics and Finance* 30, no. 3: 281–307.

Oster, E. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37: 187–204.

Pan, Y., S. Smith, and M. Sulaiman. 2018. "Agricultural Extension and Technology Adoption for Food Security: Evidence From Uganda." *American Journal of Agricultural Economics* 100, no. 4: 1012–1031.

Pearl, J. 1995a. "Causal Inference From Indirect Experiments." *Artificial Intelligence in Medicine* 7, no. 6: 561–582.

Pearl, J. 1995b. "On the Testability of Causal Models With Latent and Instrumental Variables." In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, edited by P. Besnard and S. Hanks, 435–443. ACM Press.

Pearl, J., and D. Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pufahl, A., and C. R. Weiss. 2009. "Evaluating the Effects of Farm Programmes: Results From Propensity Score Matching." *European Review of Agricultural Economics* 36, no. 1: 79–101.

Quisumbing, A., A. Ahmed, D. Gilligan, et al. 2020. "Randomized Controlled Trials of Multi-Sectoral Programs: Lessons From Development Research." *World Development* 127: 104822.

Rambachan, A., and J. Roth. 2023. "A More Credible Approach to Parallel Trends." *Review of Economic Studies* 90, no. 5: 2555–2591.

Rambachan, A., and J. Roth. 2024. "HonestDiD: Robust Inference in Difference-in-Differences and Event Study Designs." R Package Version 0.2.6. Accessed July 2025. https://CRAN.R-project.org/package=HonestDiD.

Rao, E., and M. Qaim. 2013. "Supermarkets and Agricultural Labor Demand in Kenya: A Gendered Perspective." *Food Policy* 38: 165–176.

Rao, M. 2025. "Program Evaluations and Policy Spending." Accessed July 2025. https://michelle-rao.github.io/website_papers/01_papers/Rao_policyCCTs.pdf.

Rodríguez, O., M. Vrachioli, D. Wuepper, and J. Sauer. 2025. "World Heritage Status and Farmers' Income: Evidence From a Regression Discontinuity Design in Colombia." *Agricultural Economics* 56, no. 5: 728–748.

Roth, J. 2022. "Pretest With Caution: Event-Study Estimates After Testing for Parallel-Trends." *American Economic Review: Insights* 4, no. 3: 305–322.

Roth, J. 2025. "pretrends: Power Calculations and Visualization for Pre-Trends." R Package Version 0.1.0. Accessed July 2025. https://github.com/jonathandroth/pretrends.

Roth, J., and P. H. C. Sant'Anna. 2023. "Efficient Estimation for Staggered Rollout Designs." *Journal of Political Economy: Microeconomics* 1, no. 4: 669–709.

Roth, J., P. H. C. Sant'Anna, A. Bilinski, and J. Poe. 2023. "What's Trending in Difference-In-Differences? A Synthesis of the Recent Econometrics Literature." *Journal of Econometrics* 235, no. 2: 2218–2244.

Ruml, A., and M. Qaim. 2021. "New Evidence Regarding the Effects of Contract Farming on Agricultural Labor Use." *Agricultural Economics* 52: 51–66.

Sant'Anna, P. H., and J. Zhao. 2020. "Doubly Robust Difference-in-Differences Estimators." *Journal of Econometrics* 219, no. 1: 101–122.

Schulz, D., C. Stetter, J. Muro, et al. 2024. "Trade-Offs Between Grassland Plant Biodiversity and Yields Are Heterogenous Across Germany." *Communications Earth & Environment* 5, no. 1: 514.

Sellare, J., E. Meemken, and M. Qaim. 2020. "Fairtrade, Agrochemical Input Use, and Effects on Human Health and the Environment." *Ecological Economics* 176: 106718.

Sellare, J., E.-M. Meemken, C. Kouamé, and M. Qaim. 2020. "Do Sustainability Standards Benefit Smallholder Farmers Also When Accounting for Cooperative Effects? Evidence From Côte d'Ivoire." *American Journal of Agricultural Economics* 102, no. 2: 681–695.

Smale, M., and N. Mason. 2014. "Hybrid Seed and the Economic Well-Being of Smallholder Maize Farmers in Zambia." *Journal of Development Studies* 50, no. 5: 680–695.

Staiger, D., and J. Stock. 1997. "Instrumental Variables With Weak Instruments." *Econometrica* 65: 557–586.

Stata Press. 2023. "Stata Extended Regression Models Reference Manual, Release 18." Accessed December 2024. https://www.stata.com/manuals/erm.pdf.

Stetter, C., P. Mennig, and J. Sauer. 2022. "Using Machine Learning to Identify Heterogeneous Impacts of Agri-Environment Schemes in the EU: A Case Study." *European Review of Agricultural Economics* 49: 723–759.

Storm, H., K. Baylis, and T. Heckelei. 2020. "Machine Learning in Agricultural and Applied Economics." *European Review of Agricultural Economics* 47, no. 3: 849–892.

Storm, H., T. Heckelei, and K. Baylis. 2024. "Probabilistic Programming for Embedding Theory and Quantifying Uncertainty in Econometric Analysis." *European Review of Agricultural Economics* 51, no. 3: 589–616.

Sun, L., and S. Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects." *Journal of Econometrics* 255, no. 2: 175–199.

Syrgkanis, V., V. Lei, M. Oprescu, M. Hei, K. Battocchi, and G. Lewis. 2019. "Machine Learning Estimation of Heterogeneous Treatment Effects With Instruments." arXiv. Accessed December 2024. http://arxiv.org/abs/1905.10176.

Tabe-Ojong, M., K. Mausch, T. Woldeyohanes, and T. Heckelei. 2022. "Three Hurdles Towards Commercialisation: Integrating Subsistence Chickpea Producers in the Market Economy." *European Review of Agricultural Economics* 49, no. 3: 668–695.

Taylor, R. 2022. "It's in the Bag? The Effect of Plastic Carryout Bags Bans on Where and What People Purchase to Eat." *American Journal of Agricultural Economics* 104, no. 5: 1563–1584.

Terza, J., A. Basu, and P. Rathouz. 2008. "Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling." *Journal of Health Economics* 27, no. 3: 531–543.

Thistlewaite, D. L., and D. T. Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment." *Journal of Educational Psychology* 51: 309–317.

Todd, P., and K. Wolpin. 2023. "The Best of Both Worlds: Combining Randomized Controlled Trials With Structural Modeling." *Journal of Economic Literature* 61, no. 1: 41–85.

Verhofstadt, E., and M. Maertens. 2014. "Smallholder Cooperatives and Agricultural Performance in Rwanda: Do Organizational Differences Matter?" *Agricultural Economics* 45: 39–52.

Vivalt, E., and A. Coville. 2023. "How Do Policymakers Update Their Beliefs?" *Journal of Development Economics* 165: 103121.

Vosgerau, J., G. Giambastiani, and I. Scopelliti. 2025. "The Big Data Fallacy: When N is Large, Correlation = Causation." Available at SSRN. Accessed November 2025. https://doi.org/10.2139/ssrn.5553621.

Wager, S., and S. Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113, no. 523: 1228–1242.

Wang, Y., and M. Bellemare. 2020. "Lagged Variables as Instruments." Working Paper. Accessed December 2024. https://marcfbellemare.com/wordpress/wp-content/uploads/2020/09/WangBellemareLagIVsJuly2020.pdf.

Westreich, D., and S. Greenland. 2013. "The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients." *American Journal of Epidemiology* 177, no. 4: 292–298.

Wilebore, B., M. Voors, E. H. Bulte, D. Coomes, and A. Kontoleon. 2019. "Unconditional Transfers and Tropical Forest Conservation: Evidence From a Randomized Control Trial in Sierra Leone." *American Journal of Agricultural Economics* 101, no. 3: 894–918.

Windmeijer, F. 2024. "Testing Underidentification in Linear Models, With Applications to Dynamic Panel and Asset Pricing Models." *Journal of Econometrics* 240, no. 2: 105104.

Wooldridge, J. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wooldridge, J. 2015. "Control Function Methods in Applied Econometrics." *Journal of Human Resources* 50, no. 2: 420–445.

Wooldridge, J. 2021. "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators." Working Paper, Available at SSRN. Accessed December 2024. https://doi.org/10.2139/ssrn.3906345.

Wooldridge, J. M. 2023. "Simple Approaches to Nonlinear Difference-in-Differences With Panel Data." *Econometrics Journal* 26, no. 3: C31–C66.

Wuepper, D. 2020. "Does Culture Affect Soil Erosion? Empirical Evidence From Europe." *European Review of Agricultural Economics* 47, no. 2: 619–653.

Wuepper, D., P. Borrelli, and R. Finger. 2020. "Countries and the Global Rate of Soil Erosion." *Nature Sustainability* 3, no. 1: 51–55.

Wuepper, D., and R. Finger. 2023. "Regression Discontinuity Designs in Agricultural and Environmental Economics." *European Review of Agricultural Economics* 50, no. 1: 1–28.

Wuepper, D., and R. Huber. 2022. "Comparing Effectiveness and Return on Investment of Action- and Results-Based Agri-Environmental Payments in Switzerland." *American Journal of Agricultural Economics* 104, no. 5: 1585–1604.

Wuepper, D., S. Le Clech, D. Zilberman, N. Mueller, and R. Finger. 2020. "Countries Influence the Trade-Off Between Crop Yields and Nitrogen Pollution." *Nature Food* 1, no. 11: 713–719.

Wuepper, D., W. A. Oluoch, and H. Hadi. 2025. "Satellite Data in Agricultural and Environmental Economics: Theory and Practice." *Agricultural Economics* 56: e70006.

Wuepper, D., H. Wang, W. Schlenker, M. Jain, and R. Finger. 2023. *Institutions and Global Crop Yields. Technical Report 31426*. National Bureau of Economic Research.

Wuepper, D., H. Yesigat Ayenew, and J. Sauer. 2018. "Social Capital, Income Diversification and Climate Change Adaptation: Panel Data Evidence From Rural Ethiopia." *Journal of Agricultural Economics* 69, no. 2: 458–475.

Xu, Y. 2017. "Generalized Synthetic Control Method: Causal Inference With Interactive Fixed Effects Models." *Political Analysis* 25, no. 1: 57–76.

Xu, Y., and L. Liu. 2021. "gsynth: Generalized Synthetic Control Method." R Package Version 1.2.1. Accessed July 2025. https://CRAN.R-project.org/package=gsynth.

Young, A. 2022. "Consistency Without Inference: Instrumental Variables in Practical Application." *European Economic Review* 147: 104112.

Zou, B., Y. Chen, A. Mishra, and S. Hirsch. 2024. "Agricultural Mechanization and the Performance of the Local Chinese Economy." *Food Policy* 125: 102648.

Appendix A

Extended IV Methods

While the discussions in Section 3 refer to IV and 2SLS regression, they are largely transferable to other methods that rely on IVs such as 3-Stage Least Squares (3SLS) regression, extended IV methods for binary endogenous regressors (Angrist and Pischke 2009, 142–144; Wooldridge 2010, 937–942; Wooldridge 2015), and more recent estimators that are particularly suited to handling binary and ordinal endogenous variables such as the extended regression IV approaches in Stata, which estimate the parameters using maximum likelihood (see Jafari et al. 2023, for an example and Stata Press 2023, 183, for a technical description). These discussions are also largely transferable to estimators that are based on distributional assumptions of error terms as suggested by Heckman (1976) such as the endogenous treatment effect model and the endogenous switching regression model. These models can be estimated with a two-stage approach that uses an inverse Mills ratio as an additional regressor in the second-stage regression or with a one-step maximum likelihood estimation. In fact, these models can be estimated without IVs (or exclusion restrictions) but in this case, the identification of the estimated parameters hinges solely on the distributional assumptions, for example, a bivariate normal distribution of the two error terms. As it is very unlikely that the distributional assumptions will be fulfilled exactly in a real-world application, using these estimators without IVs would very likely result in unreliable estimates. As strong IVs render the distributional assumptions less relevant, it is imperative to use strong IVs when using these estimators. Thus, at least one variable that strongly affects the selection outcome (i.e., whether an observation is treated in an endogenous treatment effect model or whether an observation is in the first or second outcome regime of an endogenous switching regression model) but does not affect the dependent variable of the outcome equation and is not related to the error term(s) of the outcome equation(s) is needed (see, e.g., Auci et al. 2021, for an example). These variables are frequently called IVs because they basically need to fulfil the same criteria as IVs in the regression methods discussed in the beginning of this section. Hence, the validity of the exclusion restrictions must be investigated and critically discussed in similar ways to the validity of IVs in the regression methods discussed in the beginning of this section.

A straightforward extension of a 2SLS estimation to non-linear regression models would be to regress each endogenous explanatory variable on the exogenous explanatory variables and the IVs (using linear or non-linear regression) and to obtain the predicted values of the endogenous explanatory variables. One can then estimate the non-linear regression model with the endogenous explanatory variables replaced by the predicted values obtained in the first stage. However, caution is advised here to avoid falling into what Angrist and Pischke (2009) refer to as the "forbidden regression" trap and directly applying the 2SLS argument to a non-linear case, for example, using the predicted values from a probit first stage in the second stage. Another mistake that must be avoided in this context is, when dealing with both a linear and quadratic form of the endogenous variable, simply using the square of the predicted values from the first stage instead of estimating two separate first-stage regressions (Angrist and Pischke 2009).

In the case of non-linear least-squares regression, the Non-linear Two-Stage Least Squares (N2SLS) estimator has similar properties to the 2SLS estimator (Amemiya 1974). However, in many other non-linear regression models (e.g., logit, probit, count-data models), this approach, which is sometimes called Two-Stage Predictor Substitution (2SPS), results in inconsistent estimates (e.g., Terza et al. 2008). An alternative to this approach is a slightly different procedure: the first stage is identical to the first-stage regression of 2SLS, N2SLS and 2SPS estimators, but in the second stage, the residuals that were obtained in the first stage are added as additional regressors (while the endogenous explanatory variables are used as regressors). This approach is called Two-Stage Residual Inclusion (2SRI) in biostatistics and health economics Terza et al. (2008), while it is called the Control-Function (CF) approach in the econometrics literature (e.g., Wooldridge 2015). In the case of linear regression models, this approach provides the same estimates as a 2SLS estimation, while the consistency of this approach has been demonstrated for many non-linear estimators. Hence, it is frequently used to address the endogeneity of regressors in non-linear regression models such as double hurdle models (e.g., Rao and Qaim 2013; Sellare, Meemken, and Qaim 2020) or fractional logit models (e.g., Wuepper 2020). As the identifying assumptions for the 2SRI/CF approach are similar to those of IV

and 2SLS estimations, the identification strategy should be based on the same evaluation criteria as for other estimations with IVs.

A further regression framework that can be used in an IV setting is the Generalised Method of Moments (GMM), which identifies the regression coefficients by assuming moment conditions in the population and then imposing these moment conditions in the sample. The number of assumed moment conditions must be equal to or larger than the number of regression coefficients to be estimated. Given that a myriad of different moment conditions can be assumed, the GMM framework is very flexible and many well-known estimators such as OLS regression and 2SLS regression are special cases. If a GMM approach is used to estimate causal effects, the appropriateness of the assumed moment conditions must be thoroughly and critically discussed. If a GMM estimation uses IVs, the validity of these IVs should be discussed in a similar way as for other methods that use IVs. If we have more moment conditions available than we have regression coefficients, a Sargan-Hansen test (also known as Sargan's *J* test or Hansen's *J* test) can be used to empirically assess the validity of the moment conditions.

In the case of panel data, the GMM framework can address the endogeneity of explanatory variables even without external IVs by using the lagged values of some variables as "internal" IVs. The "Difference GMM" estimator suggested by Arellano and Bond (1991) and the "System GMM" estimator suggested by Arellano and Bover (1995) and Blundell and Bond (1998) are frequently used GMM estimators that use internal IVs. The moment conditions assumed by these types of estimators can be complex. Similar to using lagged values of endogenous regressors as IVs in 2SLS estimations (see Section B below and Wang and Bellemare 2020), these types of estimators usually require restrictive assumptions about unobserved factors, which may be unrealistic in most empirical applications.

Finally, frequently used Structural Equation Modelling (SEM) and Partial Least Squares SEM (PLS-SEM) rely on similar identifying assumptions as outlined in Sections 2 and 3 (e.g., regarding unobserved confounders and exclusion restrictions) and are often based on additional assumptions. Hence, like other methods, a causal interpretation of the SEM results requires a careful and critical discussion of the assumptions that the SEM relies on.

Even if one uses a valid (i.e., exogenous and highly relevant) IV, IV regression can result in substantially biased estimates if parametric assumptions, e.g., about the functional form, are not fulfilled (Okui et al. 2012). Hence, it might be worthwhile to consider using nonparametric IV regression methods. Chernozhukov et al. (2018) show that Double Machine Learning (see Section 2) can also be applied to an IV setting, which means the linearity assumption of 2SLS regression can be relaxed. Their approach allows both the outcome equation and the treatment equation to be unknown nonlinear equations that can be approximated by any flexible machine learning algorithm. However, it still requires assuming either homogeneity of treatment or homogeneity of treatment assignment. Under these conditions, the approach provides a consistent estimate of an ATE. Going further, multiple approaches also relax the homogeneity assumptions and allow the estimation of treatment effects that vary depending on the observed characteristics. Hartford et al. (2017) have developed an approach called DeepIV, which uses deep neural networks in both the outcome and the treatment model. Athey et al. (2019) have developed Generalised RFs as a nonparametric estimator that can be used to estimate any quantity identified by a set of (local) moment conditions. They demonstrate that this approach can be used to estimate treatment effects under the unconfoundedness assumption (leading to an approach called Causal Forests, see Section 2) but also in an IV setting. Generalised RFs can basically be understood as a more flexible alternative to GMM estimation methods. Importantly, Generalised RFs are able to learn treatment heterogeneity in a data-driven manner. Additionally, it is possible to obtain asymptotic uncertainty intervals for the estimated treatment effect, allowing the user to assess uncertainty in the estimates and perform hypothesis testing. While DeepIV and Generalised RFs are specifically designed around deep neural networks and RFs, respectively, Syrgkanis et al. (2019) provide a generalised framework (Orthogonal IV) for nonparametric IV estimations that allows the use of any machine learning approach in the outcome and treatment model. They also develop methods that allow the projection of treatment heterogeneity to a simpler (potentially linear) lower dimensional space. This means asymptotic confidence intervals can be derived and machine learning interpretability methods (e.g., SHAP values) can be used to illustrate and inspect treatment heterogeneity.

Another relatively specialised case of machine learning in the context of IV estimation is to deal with a situation in which there is a large number of potential IVs (potentially larger than the number of observations). Belloni et al. (2012) demonstrate that simple machine learning methods such as LASSO can be used to select IVs under the assumption that the treatment assignment can be sufficiently predicted by a small subset of all the available IVs. However, in empirical settings, we very rarely face the (luxury) problem of having too many IVs.

Most of the machine-learning approaches that are relevant for applied economists (Double Machine Learning, DeepIV, Causal Forest, Generalised RFs for IV, Orthogonal IV) are available in the Python package *EconML* (https://econml.azurewebsites.net/index.html), which provides a unified API for all these approaches and represents a relatively simple application for applied researchers.

Appendix B

Special Types of IVs

This section discusses some special types of IVs that are frequently used in agricultural and applied economics. One of these special types of IVs is the so-called spatial IV or leave-one-out IV (e.g., Mason et al. 2013; Krishnan and Patnam 2014; Smale and Mason 2014; Magnan et al. 2015; Wuepper et al. 2018; Sellare, Meemken, Kouamé, and Qaim 2020; Tabe-Ojong et al. 2022). In this case, an endogenous explanatory (treatment) variable is instrumented by the average or proportion within a peer group leaving out the respective observation. For example, a farmer's adoption of a technology is instrumented by the proportion of farmers in the village who adopted this technology, leaving out the respective farmer. However, while this type of IV is usually highly relevant, its exogeneity requires strict assumptions that are not fulfilled in many empirical applications (Angrist 2014; Betz et al. 2018; McKenzie 2018). In some empirical analyses, it may be reasonable to use such a spatial IV or a variant thereof, potentially combined with other tools, but authors must provide clear reasoning as to why this identification strategy is valid in their study (e.g., Maggio et al. 2022).

Closely related to spatial IVs are Hausman-type IVs, which are frequently used in food product demand analyses to account for the endogeneity of product prices (see, e.g., Nevo 2001). The idea is that the price of a product in other regions can be used as an IV since the same product has similar marginal costs across regions but different demand shifters (Hausman 1996; Nevo 2000; Hirsch et al. 2018). However, this assumption may be violated in the case of a nationwide shock in demand; for example, if a nationwide advertising campaign that influences the demand of a product across regional borders is launched (Nevo 2000, 2001).

Similar to using lagged values of explanatory variables to address endogeneity in an identification-on-observables identification strategy (see Section 2), lagged values can also be used as IVs; an identification strategy that is popular among applied economists. However, Wang and Bellemare (2020) show that IVs of this type require specific assumptions. For instance, even if the exclusion restriction is fulfilled, the estimates are biased (although consistent), and the likelihood of making Type-1 errors is high if there is first-order autocorrelation in unobserved factors because this leads to a correlation between the lagged IV and the error term (Wang and Bellemare 2020). As this cannot be ruled out in most empirical applications, Wang and Bellemare (2020) conclude that using lagged values of endogenous explanatory variables as IVs "is unlikely to lead to credible estimates."

Shift-share IVs, also known as Bartik-type IVs (Bartik 1991; Borusyak et al. 2025), can be used in cases where panel data is available and the

intensity of a unit's treatment is affected by an initial share that affects the exposure to a trend. Either the trend or the share needs to be exogenous for this approach to be valid. Then, the interaction of the shift and the share provides an IV, conditional on the standard IV assumptions being valid. For example, when analysing the effect of a regional subsidy on farm performance, a shift-share IV can be constructed based on the idea that the nationwide values of subsidies "shift" the regional (endogenous) subsidies according to a predetermined out-of-sample economic state of the region (share) (see, e.g., Zou et al. 2024, for an example). More precisely, in this case, the Bartik IV is the product of a variable that captures the national subsidy level and a variable with information on the initial state of the regional economy, e.g., 1 year before the start of the sample period that is used in the analysis. This reflects the exogenous variation in regional subsidies which is uncorrelated with the regional-level error term, which means that it may serve as a valid IV (Bartik 1991; Breuer 2022; Zou et al. 2024). It is important to note that for shift-share IVs, valid identification can be achieved when either the shift component or the share component of the IV is exogenous. For additional guidance, we refer to Borusyak et al. (2025). Another illustrative example of a shift-share IV analysis is the paper of Gollin et al. (2021) who estimate the impact of the Green Revolution with a shift-share IV.