DEPARTMENT OF FOOD AND RESOURCE ECONOMICS
UNIVERSITY OF COPENHAGEN

IFRO Working Paper

Estimating Causal Effects with Observational Data

Guidelines for Agricultural and Applied Economists

Arne Henningsen
Guy Low
David Wuepper
Tobias Dalhaus
Hugo Storm
Dagim Belay
Stefan Hirsch

2024 / 03

IFRO Working Paper 2024 / 03

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

Authors: Arne Henningsen[1], Guy Low, David Wuepper, Tobias Dalhaus, Hugo Storm, Dagim Belay, Stefan Hirsch

[1] arne@ifro.ku.dk

JEL-classification: C21, C23, C24, C26, C51, C52

Published: December 2024

See the full IFRO Working Paper series here:
https://ifro.ku.dk/english/publications/ifro_series/working_papers/

Department of Food and Resource Economics (IFRO)
University of Copenhagen
Rolighedsvej 23
DK 1958 Frederiksberg DENMARK
https://ifro.ku.dk/english/

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

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

[1] Department of Food and Resource Economics, University of Copenhagen, Denmark;
[2] Business Economics Group, Wageningen University & Research, The Netherlands;
[3] Institute for Food and Resource Economics, University of Bonn, Germany;
[4] Department of Management in Agribusiness, University of Hohenheim, Germany

Abstract
Most research questions in agricultural and applied economics are of a causal nature, i.e., how one or more variables (e.g., policies, prices, the weather) affect one or more other variables (e.g., the welfare of individuals or the society, the demanded or produced quantity, pollution). Only a small number of these research questions can be studied with economic experiments such as randomised controlled trials (RCTs), lab experiments or lab-in-the-field experiments. Hence, most empirical studies in agricultural and applied economics use observational data. However, estimating causal effects with observational data requires appropriate research designs and convincing identification strategies, which are usually very difficult or even impossible to devise. Likely as a consequence, in the applied economics literature, it can commonly be observed that results are interpreted as causal despite lacking a robust identification strategy, which has contributed to a credibility crisis in economics research. This paper provides an overview of various 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 who are intending to estimate causal effects with observational data, e.g., how to assess and discuss the chosen identification strategies in their publications.

Keywords: causal inference, observational data, instrumental variables, difference in differences, regression discontinuity

JEL codes: 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 economics research is that econometric designs are not suitable for identifying causal effects. Despite this, the resulting estimates are interpreted as such.[1] Therefore, reported results may often erroneously reflect the parameters that have been estimated (Gibson, 2019). Estimates obtained with Ordinary Least Squares (OLS), matching approaches, or difference-in-differences (DID) methods based on observational data may overstate the effect by 20–82% compared to causal estimates based on an experiment (McKenzie et al., 2010). For instance, Wuepper et al. (2021) find a robust positive association between family farming and rural employment that remains even in an instrumental-variable regression. However, the association disappears in a panel data regression with region- and year-fixed effects, suggesting that the entire cross-sectional association is spurious.

Therefore, thorough consideration of causality is of the utmost importance when conducting empirical economics research (Imbens, 2024). The misinterpretation of simple associations as causal effects, together with insufficient robustness and replicability of empirical analyses have led to a "credibility revolution" in quantitative economics research and a call for higher standards in statistical identification (Angrist and Pischke, 2010; Bellemare, 2012; Gibson, 2019).[2] While the "credibility revolution" has its origin in labour economics, it has also reached agricultural and applied economics, albeit with a delay (Bellemare, 2012). Here, it can still frequently be observed that empirical results that are used to test specific hypotheses on the relationship between economic variables are interpreted causally using terms such as "effect" or "impact" although the underlying research design and econometric framework are not based on a credible identification strategy, or at least not a sufficiently described and motivated identification strategy. For example, some studies use OLS or matching methods, which rely on a selection-on-observables assumption, in a context with strong selection-on-unobservables. 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 a method based on 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 method and the IVs. Therefore, the mere application of an IV approach without sufficient verification of the underlying assumptions, especially regard-

[1] A current additional issue that is contributing to the problem is p-hacking and the related p-value debate (e.g., Ioannidis and Doucouliagos, 2013; Heckelei et al., 2023) as well as other issues of statistical malpractice.

[2] More general factors related to the credibility of results include insufficient sample size, insufficient standardisation of variable definitions across studies, or exploratory research which lacks a motivation regarding the selection of tested relationships (Ioannidis and Doucouliagos, 2013). The latter is also known as Hypothesising After the Results are Known (HARKing), which can result in misleading conclusions or biased and less replicable results. See Ioannidis and Doucouliagos (2013) for a more detailed discussion of these drivers of incredibility.

ing the selected instrumental variables, is often falsely regarded as a sufficient condition for allowing 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 instrumental variable does not produce a strong first stage.

The correct identification of causal effects is particularly relevant for agricultural economics research because stakeholders such as policy makers, agribusinesses, or farmers often base real-world decisions on research results. Thus, incorrect or overestimated interpretations of results may lead to the misallocation of private or public funds (Finger et al., 2023). Hence, empirical agricultural economics papers that aim to identify causal effects should include a clear description and justification of the underlying "identification strategy". This refers to the identification of the exogenous variation in an endogenous covariate or treatment variable of interest, i.e., the part of the variation in this variable that is not related to unobservable factors (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 investigated.[3] If a method for addressing the non-experimental nature of a data set is used, the added value compared to classical approaches such as OLS should be pointed out. If the added value cannot be clearly highlighted, it may be preferable to stick with OLS estimation and interpret the results as associations.

The "gold standard" for analysing causal research questions is randomised controlled trials (RCTs) (Gibson, 2019). However, most of the (causal) research questions in agricultural and applied economics cannot be answered with experiments because they would be problematic or infeasible for various reasons. For example, randomly assigning import tariffs, randomly assigning different education levels 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, p. 36) would either be infeasible, impractical or unethical[4]. 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 and may suffer from non-compliance with treatment. In addition, it is difficult to identify the mechanisms behind the cause-

[3] In addition, the external validity of the results should be outlined and discussed, e.g., 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 out of the scope of this paper, and we only discuss internal validity.

[4] Note that even if such experiments were feasible, 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).

effect interplay (Quisumbing et al., 2020; Koppenberg et al., 2023; Todd and Wolpin, 2023). However, highly relevant research questions should not be neglected just because they cannot be answered by applying experimental methods. Instead, observational data needs to be used to answer these research questions as thoroughly as possible.

This paper discusses various 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 instrumental variables (e.g., Jiang, 2017; Young, 2022; Lal et al., 2024) for different disciplines, and tailor our guidelines to research questions and the commonly used econometric approaches in agricultural and applied economics.

The following section discusses the use of various methods that are based on the 'selection on observables' identification strategy such as ordinary-least squares (OLS) and matching methods (e.g., propensity score matching). The third section explores methods that are based on instrumental variables (or exclusion restrictions) such as 2SLS regression and endogenous switching regression. The fourth section discusses fixed-effects estimations and difference-in-differences (DiD) approaches, while the fifth section examines regression discontinuity designs.[5] Finally, the sixth section concludes the paper and provides some general guidelines for agricultural and applied economics research.

2 Selection on Observables

The selection-on-observables identification strategy is based on the assumption that we 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 conditional independence assumption (CIA), conditional ignorability, or conditional unconfoundedness.

Classical regression analyses (e.g., ordinary least squares (OLS), logit, probit, tobit, or Poisson regression) can be affected by three potential sources of statistical endogeneity:[6]

[5] The synthetic control method (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015) is a further method for estimating causal effects with observational data which has recently become very popular in the social sciences. As this method is very rarely used in agricultural economics, we do not provide guidelines for this method but instead refer to the excellent guidelines provided by Abadie (2021).

[6] In this paper, we focus on the endogeneity of explanatory variables. However, all other assumptions that are required for obtaining 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

(a) omitted variables / unobserved heterogeneity; (b) measurement error (any type of measurement error in the explanatory variable 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, e.g., 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, i.e., 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]

Whether a selection-on-observables identification strategy is feasible can, for example, be assessed by using Directed Acyclic Graphs (DAG). With DAGs, one can assess whether it is possible to find a set of (observed) control variables so that all "backdoor paths" between the treatment variable and the outcome variable are blocked (see, e.g., Morgan and Winship, 2014). A DAG can also be used to determine which variables should *not* be used as control variables, i.e., variables on the causal path from the treatment variable to the outcome variable ("bad controls"). If one or more of the "backdoor paths" cannot be blocked, i.e., there is at least one backdoor path that does not include any observed variable, one can use the DAG to consider whether the "front-door criterion" or an instrumental-variable approach (see following section) can be used to estimate a causal effect (see, e.g., Bellemare et al., 2024, for an example).[8]

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

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/`).

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.

Using matching methods such as propensity score matching (PSM)[9] or inverse probability weighting for estimating causal effects with observational data is basically 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. The same applies to the augmented inverse propensity weighted (AIPW) estimator which is 'doubly-robust' as it basically requires the same identification strategy as an OLS regression (e.g., Kurz, 2022, equation 1).

There are methods for assessing the sensitivity of the results to unobserved heterogeneity (e.g., Oster, 2019; Diegert et al., 2023), which have been used often 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.

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 relaxed by using nonparametric regression methods, most of the available matching methods, or machine learning approaches. While machine learning methods have rapidly advanced and are being increasingly used in agricultural and applied economics, it is important to point out 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 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 3.3). These methods are then called "causal machine learn-

[9] 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.

ing." 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.

The "Causal Forests" method (Wager and Athey, 2018), which is a special case of Generalised Random Forests (Athey et al., 2019), extends the DML approach allowing the estimation of heterogeneous treatment effects, i.e., 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; Stetter et al., 2022; Deines et al., 2023; Schulz et al., 2024).

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 6):

- 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.

- 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 methods for assessing the sensitivity of the results to unobserved heterogeneity.
- Consider using methods that do not rely on strict parametric assumptions.

3 Instrumental-Variable Methods

Instrumental-variable methods are often used in cases in which selection on observables cannot be justified (Lal et al., 2024). We define 'instrumental-variable (IV) methods' in a broad sense. The first part of this section refers to linear IV and 2-stage least squares (2SLS) regression (which is identical to IV-regression if the number of IVs[10] is equal to the number of endogenous regressors), while Section 3.1 refers to other estimators that also rely on IVs, including machine-learning IV methods. In Section 3.2, we present a brief overview of special types of instruments, while Section 3.3 provides practical advice on using IV methods.

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 instruments can exacerbate the problem, so that the bias in the 2SLS estimator even exceeds the OLS endogeneity bias (Lal et al., 2024). By construction, IV estimates are less precise than OLS estimates. Lal et al. (2024) show that 2SLS estimates have, on average, six times higher standard errors than OLS estimates although this decreases with instrument strength.[11]

Using an instrumental-variable approach to estimate a causal effect is possible if one has at least as many instrumental variables as endogenous regressors. These instrumental variables must fulfil the following two criteria: (a) they must be "relevant", i.e., strongly related to the endogenous regressors and (b) they must be statistically "exogenous", i.e., not related to the error term (exclusion restriction).

The first criterion can be empirically investigated with tests for weak instruments. Traditionally, an instrumental variable was considered to be 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 F-statistic of the first stage is below 20. They also demonstrate that in cases in which

[10] In this paper, we use the narrow definition of IVs, i.e., 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, while 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.

[11] This makes 2SLS more susceptible to p-hacking and publication bias (Lal et al., 2024).

there is only one instrument, the evaluation of instrument strength should be based on an F-statistic that exceeds 50. Moreover, estimation results (e.g., t-tests) are often unreliable even in cases in which there are much higher values for the F-statistic (e.g., Lee et al., 2022; Keane and Neal, 2023, 2024). Moreover, 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 (not weak).

Exclusion restrictions imply that the exogenous (excluded) instrument influences the dependent variable only via its effect on the endogenous explanatory variable and it is not correlated with the error term. Traditionally, the exogeneity of the instrumental variables cannot be empirically investigated unless more potential instrumental variables than endogenous regressors are available, and it is certain that there are at least as many exogenous instrumental variables as there are endogenous regressors. However, although the exclusion restrictions cannot be tested empirically, motivating their validity based on sound theoretical argumentation is of utmost importance (e.g., Lal et al., 2024). It is moreover helpful to think of placebo estimates that can be used to rule out specific violations of the exclusion restriction. For instance, the instrumental variable 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 instrumental variable for a (sub)sample of observations, where the outcome and the instrumental variable should be uncorrelated. If the endogenous regressor is a binary (treatment) variable, the falsification test suggested by Di Falco et al. (2011), in which the regression model is re-estimated with untreated observations only and with the endogenous regressor replaced by the instrumental variable, can be applied. Acemoglu et al. (2001) suggest estimating the outcome equation with both the endogenous regressor and the instrumental variable (and of course all relevant control variables). If the instrumental variable affects the dependent variable through the endogenous regressor only, the coefficient of the instrumental variable in this auxiliary regression should be close to zero. However, if the instrumental variable is highly relevant, it is highly correlated with the endogenous regressor, so that the coefficient of the instrumental variable is very imprecisely estimated in this auxiliary regression, and a statistical test on this coefficient has very little statistical power. If the main concern is that the instrumental variable 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 potential violation of the exclusion restriction by regressing this pathway on the instrumental variable. For example, if an instrumental variable 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 instrumental variable. One weakness of all these placebo tests is that they can never 'prove' that an IV is exogenous because a 'successful' placebo test, i.e., a statistically insignificant result may have many explanations, e.g., insufficient statistical power caused by a small number of observations, multicollinearity, or a large error variance. Hence, it is always necessary to critically discuss the assumption of statistical exogeneity for each instrumental variable used, e.g., by exploring potential (unobserved) variables that may be related to both the treatment variable and the outcome variable. This is very important as, for example, McKenzie et al. (2010) show that using instruments for which the exclusion restriction 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 makes things worse. As a general rule, the less specific the chosen instrumental variable, the less likely the exclusion restriction is valid (see, e.g., Mellon (2024) for a discussion of rainfall as an instrument).

In the case of a weak instrument or a violation of the exclusion restrictions, 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, e.g., 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. It is, therefore, advisable to also 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 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, which results 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 instruments 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).

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).[12] However, these assumptions are unlikely to be fulfilled in many

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

empirical analyses. Under less restrictive assumptions (e.g., monotonicity of the effect of the instrumental variable on the endogenous explanatory variable), 2SLS estimates indicate local average treatment effects (LATE), which indicates the effect of the part of the variation in the endogenous explanatory variable that is caused by variation in the instrumental variable (e.g., Imbens and Angrist, 1994). For instance, in the case of a binary instrumental variable and a binary endogenous explanatory variable, the LATE indicates the average treatment effect on those subjects that 'comply' with the instrumental variable, while the effects on the 'always takers' and the 'never takers' remain unidentified. 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).

3.1 Extended IV Methods

While the discussions above refer to IV and 2SLS regression, they are largely transferable to extended IV methods such as Wooldridge's 3-step IV method for binary endogenous regressors (Wooldridge, 2010, p. 937–942), 3-stage least squares (3SLS), 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), p. 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 additional regressor in the second-stage regression or with a one-step maximum likelihood estimation. In fact, these models can be estimated without instrumental variables (or exclusion restrictions) but in this case, the identification of the estimated parameters hinges solely on the distributional assumptions, e.g., 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 instrumental variables would very likely result in unreliable estimates. As strong instrumental variables render the distributional assumptions less relevant, it is imperative to use strong instrumental variables 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 instrumental variables because they basically need to fulfil the same criteria as instrumental variables 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 instrumental variables in the regression methods discussed in the beginning of this section.

A straight-foward 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 instrumental variables (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 (e.g., 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 et al., 2020a) or fractional logit models (e.g., Wuepper, 2020). As the identifying assumptions for the control function 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 instrumental-variable 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 instrumental variables, the validity of these IVs should be discussed as described above 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 instruments by using the lagged values of some variables as 'internal' instruments. 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 instruments. 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 3.2 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.

The availability of a valid instrument is a crucial requirement for obtaining unbiased treatment estimates using any IV approach. However, 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 instrument, 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 outcomes are the sum of an (unknown) nonlinear function of treatment and observed covariates (that are uncorrelated to unobservables) and an additive error term which may be correlated to unobservables (Newey and Powell, 2003). However, early nonparametric approaches based on basis functions/splines or kernel methods struggle with a larger number of covariates or instruments 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.

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 average treatment effect (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 treatment model. Athey et al. (2019) have developed Generalised Random Forests (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.

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.[13]

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 instruments (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 instruments under the assumption that the treatment assignment can be sufficiently predicted by a

[13] 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.

small subset of all the available instruments. However, in empirical settings, we very rarely face the (luxury) problem of having too many IVs.

3.2 Special types of Instruments

A special type of instrumental variable which is popular among applied economists is the so-called spatial instrumental variable or leave-one-out instrumental variable (e.g., Mason et al., 2013; Krishnan and Patnam, 2014; Smale and Mason, 2014; Magnan et al., 2015; Wuepper et al., 2018; Sellare et al., 2020b; Tabe-Ojong et al., 2022; Aïhounton and Henningsen, 2024). 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 instrumental variable 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 instrumental variable 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 instruments are Hausman-type instruments, 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 instrument 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 instrumental variables; 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 instrumental variables "is unlikely to lead to credible estimates."

Shift-share instruments, also known as Bartik-type instruments (Bartik, 1991; Borusyak et al., 2025), can be used in cases in which one wants to account for endogeneity of regional variables, e.g., when analysing the effect of a regional subsidy on farm performance. In this case, a shift-share instrument is based on the idea that 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 state of the regional economy, e.g., one year before the start of the sample that is used in the analysis. This remaining part of the variance in the regional subsidies is uncorrelated with the regional-level error term, which means it may serve as an IV (Bartik, 1991; Breuer, 2022; Zou et al., 2024). It is important to note that for shift-share instruments, 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).

3.3 Practical checks for IV approaches

When using IV-based methods, 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 6).

If various assessments indicate that an IV regression method may be suitable, it is important to assess whether it is indeed necessary to apply an IV-based method in the empirical analysis:

- 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 whether a potentially endogenous explanatory variable should indeed be instrumented. This discussion can partly be based on statistical tests such as the Durbin-Wu-Hausman test (sometimes called “Wu-Hausman test” or just “Hausman test”), the Davidson-MacKinnon test, or an analogous test for an extended IV method. If these tests reject the null hypothesis of exogeneity, we can conclude that it is necessary to use IV regression. However, if these tests do not reject the null hypothesis of exogeneity, we cannot conclude that a selection-on-observables identification strategy is suitable. In such cases, one could discuss the suitability of a selection-on-observables identification strategy and, depending on the conclusion of this discussion, decide whether to apply a potentially inefficient IV regression method or a potentially inconsistent method based on a selection-on-observables identification strategy. In all cases, it is advisable to provide and compare the results for both estimation strategies.

If one concludes that it is necessary to use an IV regression method, it is important to assess the strength of the instruments based on the following criteria:

- Always report full first-stage results including all model diagnostics.
- Only use IV-based methods when there is a sufficiently high correlation between the endogenous explanatory variable and the IV after controlling for exogenous control variables (i.e., in the first stage of an IV estimation).
- If the F-statistic of 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 instrument, 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 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 in, e.g., 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 deeper discussion of the options for investigating instrument strength.

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

- Use a Sargan-Hansen test / Sargan's J test / Hansen's J test to test for overidentifying restrictions if the model is overidentified (i.e., the number of IVs is larger than the number of endogenous explanatory variables) and it can be assumed that there are at least as many exogenous instruments as there are endogenous regressors.
- Use strong theoretical considerations to rule out any direct effect on the dependent variable or any relationship with omitted factors (error term), see, e.g., Mellon (2024), who discusses the use of weather as an instrument.
- Use placebo tests to assess the exclusion restriction(s) but be aware of their limitations.

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

If the exclusion restriction / independence 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 managed to address the OLS bias, which involves a discussion of the direction of the OLS bias and the extent to which 2SLS was able to 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 (AR) test (Keane and Neal, 2024) instead of standard t-tests.

4 Fixed Effects and Difference in Differences

Fixed effects are a useful tool to control for unobserved confounders that are constant at the fixed-effect level. In other words, 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. For instance, 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 can not quantify the effects of time invariant factors such as proximity to a city (Wooldridge, 2010). Similarly, fixed effects 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, fixed effects 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 instance, absolute profits into deviations from average profits in the observed time period per farm (Cunningham, 2021). Therefore, fixed effects 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 fixed effects 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 fixed-effect models are connected with reverse causality and time-variant confounders, which may still introduce simultaneity and omitted-variable biases (Cunningham, 2021).

While fixed effects help to control for biases arising from unobserved confounders, a common issue in fixed-effect applications is the temporal and spatial correlation in often heteroscedastic errors. The standard approaches to dealing with this are adjusting standard errors so that they are robust against heteroscedasticity and allowing for temporal and spatial autocorrelation through clustering (Cameron et al., 2011). Another issue that may arise as a result of using fixed effects to control for time-invariant heterogeneity can be seen from the above examples on time-invariant factors. In fact, 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 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 or twice first-differenced estimator and suggest a rolling first-differenced estimator (and others), which can either be used as alternatives to fixed effect estimators or at least to show sensitivity of the estimates to these different estimators and their underlying assumptions.

An alternative approach to estimating causal effects with panel data is the difference-in-differences (DID) method. In classic 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 average difference between the observed effects of a treatment 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.[14] 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 using a data set that 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 par-

[14] In certain cases, a simple double-difference (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 trend 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).

allel 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, on consumption, 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. On production, Belay and Jensen (2020) estimate the effect of information disclosure on antibiotic use and market survival among pig farms, while Belay and Jensen (2022) evaluate the impact of limiting antibiotic use on the economic performance of farms.

The basic DID set-up can be extended to situations 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 two-way fixed effects estimator offers a reliable estimation for inferring treatment effect causality (Huntington-Klein, 2021).

However, under heterogeneous treatment timing and treatment effect heterogeneity, the estimated average treatment effect of the two-way fixed effect estimator on the treated may be biased and causally interpreting the regression coefficient becomes problematic even if the parallel-trends assumption holds (Goodman-Bacon, 2021; Athey and Imbens, 2022). For instance, this may be the staggered adoption of an agricultural policy whose effect is time-varying, i.e., 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, the estimated average treatment effect on the treated may be negative even when the 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. Studies by 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 and 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.

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). Depending on the research design, one may select either the never-treated group or the not-yet-treated group as controls, for example, if there is a gradual policy rollout (Callaway and Sant'Anna, 2021;

de Chaisemartin and D'Haultfœuille, 2023a). A researcher can opt for efficient linear estimation (Borusyak et al., 2024), two-stage difference in differences (Gardner et al., 2024) or non-linear DID models such as exponential, logit, or probit models (Wooldridge, 2021). Moreover, heterogeneity-robust DiD designs exist for staggered (i.e., irreversible), continuous (i.e., non-binary and non-discrete), and multiple (i.e., reversible and re-treatable) treatments (de Chaisemartin and D'Haultfœuille, 2023a; Callaway et al., 2024). 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, 2023a). 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 (Liu et al., 2024; Sun and Abraham, 2021). 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. While hurricanes are transitory, their long-term impact persists, so (Deryugina, 2017) models the year of the first hurricane to capture these effects. By replacing the hurricane status (on/off) with an indicator for being previously hit by a hurricane, the treatment becomes absorbing, thereby enabling the use of staggered adoption designs (Sun and Abraham, 2021).

If the unconditional parallel trends assumption holds (without covariates), Table 1 provides an overview of the recommended estimation methods and their implementation in Stata and R for various DID model scenarios. From all methods listed in Table 1, 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.

Moreover, recently suggested DID estimators offer useful event-study-type plots which visualise aggregated effects from single group-time specific treatment effects, which can be used to evaluate both (dynamic) treatment effects and test pre-treatment parallel trends (e.g., Taylor, 2022; Li and Zhu, 2024). It is important to note that these conventional (event study[15]) pre-trend tests for parallel trends often lack power and therefore fail to detect biases from pre-existing trends (Roth, 2022). Researchers should assess the statistical power of these tests using tools such as the "pretrends" R package for nonlinear trends and consider alternatives such as visualisation tools (Freyaldenhoven et al., 2021) or magnitude-based pre-trend evaluation (Bilinski and Hatfield, 2020). To avoid pretesting biases, Freyaldenhoven et al. (2019) recommend adjusting for counterfactual trends with

[15] An event study is commonly used to examine the dynamic effects of a treatment and to test the parallel-trends assumption before the treatment.

unaffected covariates, while Rambachan and Roth (2023) offer confidence sets that address pre-trend uncertainty, which can be done using the "HonestDiD" package in R or Stata. Regardless of the approach, using economic knowledge to analyse potential parallel trend violations strengthens causal inferences over relying solely on the statistical significance of pre-trends tests (Roth, 2022).

An interesting extension to study staggered treatment problems is the matrix completion approach for causal panel data models, which allows the combination of two-way fixed-effects with synthetic controls in a data-driven manner (Athey et al., 2021). In an agricultural context, this approach is particularly appealing as it naturally deals with unbalanced panel data sets (Martinsson et al., 2024).

When using fixed-effect-based or DID-based methods, we suggest doing the following (in addition to following the general suggestions that we provide in Section 6):

- 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.

- Adjust standard errors to allow for spatial and/or temporal autocorrelation.

- Evaluate the validity of the parallel trends assumption by creating event-study plots in 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 pre-treatment parallel 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, in which the parallel-trends assumption only holds when conditioning on covariates.

- Provide reasoning based on economic theory on post-treatment parallel trends in DID settings.

- Choose a suitable DID method and substantiate the choice of method by providing convincing arguments (see, e.g., Table 1).

Table 1: Difference-in-Differences Methods

Number of time periods	**Treatment scenario**	**Dynamic TE**	**Specific scenario**	**Recommended estimation method**	**Implementation in Stata module or package**	**Implementation in R package**
Two	Single treatment group	Irrelevant		(Static) TWFE, AA, SZ	reghdfe, xtreg, absdid, drdid	plm, DRDID
Multiple (event study)		No	Average of the event-study coefficients ("Overall" ATT)	(Static) TWFE, AA, SZ	reghdfe, xtreg, absdid, drdid	plm, DRDID
		Yes	Baseline: average of all pre-treatment periods	BJS, W21, GT	did_imputation, did2s, xthdidregress, jwdid, wooldid	didimputation, did2s, etwfe
			Baseline: last pre-treatment period	(Dynamic) TWFE, CS, DH, SA	reghdfe, eventdd, xtevent, eventstudyinteract, csdid, did_multiplegt_dyn	plm, fixest, did, DIDmultiplegtDYN
	Staggered	Yes	Baseline: last pre-treatment period	CS, DH, SA	eventstudyinteract, csdid, did_multiplegt_dyn	fixest, did, DIDmultiplegtDYN
			Baseline: average of all pre-treatment periods	BJS, W21, GT	did_imputation, did2s, xthdidregress, jwdid, wooldid	didimputation, did2s, etwfe
			Control group: not-yet-treated	CS, DH, BJS	csdid, did_multiplegt_dyn, did_imputation	did, DIDmultiplegtDYN, didimputation
			Control group: last to be-treated or never-treated	CS, SA	csdid, eventstudyinteract	did, fixest
			Imputation methods	BJS, W21, GT	did_imputation, did2s, xthdidregress, jwdid, wooldid	didimputation, did2s, etwfe
			Fast estimation	BJS	did_imputation	didimputation
			Non-linear estimations	W23	jwdid, wooldid	etwfe
			Two-stage differences in differences	GT	did2s	did2s
			(Quasi-)random assignment of treatment	RS, AI	staggered	staggered
	Continuous	Yes		CGS, DH	did_multiplegt_dyn	DIDmultiplegtDYN
	Multiple (treatment-on-and-off)	No	No carryover	DH	did_multiplegt_dyn	DIDmultiplegtDYN
		Yes	(Arbitrary) carryover	Methods in staggered treatment scenario, LWX	fect	fect

Notes: AA: Abadie (2005) AI: Athey and Imbens (2022); BJS: Borusyak et al. (2024); CGS: Callaway et al. (2024); CS: Callaway and Sant'Anna (2021); DH: de Chaisemartin and D'Haultfœuille (2023a); 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.

5 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). All of them share a particular mechanism for identifying causal effects: If treatment assignment is triggered by a clearly-defined threshold in a continuously distributed variable, then—given a few falsifiable assumptions—discontinuity in the outcome right at this threshold quantifies the treatment effect (Cunningham, 2021; Huntington-Klein, 2021). Intuitively, this works especially well with arbitrarily set thresholds because it 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).

The fundamental requirement for Regression Discontinuity Designs (RDD) is the existence of a continuously distributed variable that has a threshold which triggers treatment assignment.[16] 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 et al., 2020a,b). When national borders are used, the triggered treatment is to which country a given area belongs. The most intuitive way of understanding how a national border can be used to identify the effect of an area that belongs to one country but not another is provided in Figure 1.

The following example uses data from Wuepper and Finger (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 (a) and crop yields can be quantified in high resolution from satellite imagery (b). When computing local averages of crop yields at equal distances from the border and plotting these as a function of border distance, a striking pattern emerges: Whereas crop yield is distributed rather smoothly on either side of the border, there is a stark jump right at the border, which cannot be explained by potential confounders such as rainfall or sunshine because these do *not* jump at the border: It is the countries

[16] 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 instrumental variable regression such as 2SLS with the threshold as the instrument.

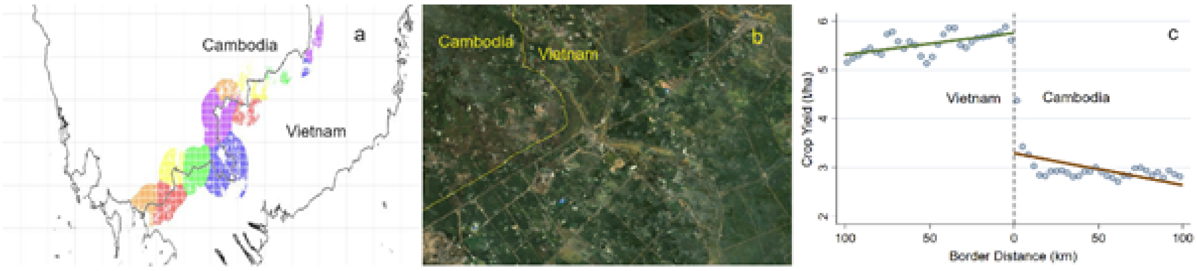

Figure 1: **a**) The border between Cambodia and Vietnam separates an otherwise comparable agricultural area into two countries. Colours distinguish different border segments. (**b**) Satellite data can be used to obtain a methodologically unified, high-resolution crop yield measure. (**c**) An important 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.

as political constructs that make the fields in Vietnam more productive than those in Cambodia (Wuepper and Finger, 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, e.g., 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.

Similarly, 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. 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. RDiT utilises flexible polynomial time trends and has been recently used in studies involving 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.

An increasingly popular research design is the Difference-in-Discontinuity Design (DiDD), which is a combination of RDD and Difference-in-Differences. It is set-up like a standard DID Design with the only difference being that it focuses 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 and Finger (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.

For the above-discussed research designs, simple procedures can be followed, which 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 is aided by off-the-shelf software especially that provided by Calonico et al. (2015) and Calonico et al. (2017).[17] 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 hectares was introduced, farmers who initially had 5.2 hectares quickly got rid of 0.3 hectares).

When using discontinuity-based methods, we suggest doing the following (in addition to following the general suggestions that we provide in Section 6):

- 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).

6 General Suggestions and Conclusions

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

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

- Start from the theoretical understanding of the problem (e.g., based on a DAG) to define an identification strategy and clearly discuss under what assumptions the

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

quantity of interest is identified, any potential explanations for violating the assumptions and their consequences for identification.

- Carefully consider the assumptions of various estimation approaches. Consider the extent to which these assumptions fit the theoretically motivated identification strategy.

- Clearly point out the added value of the chosen method compared to simpler approaches such as OLS. Unless a relevant added value can be clearly demonstrated, a simpler method may be preferable.

- Discuss the plausibility of the "Stable Unit Treatment Value Assumption" (SUTVA) in your 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.

- Simulate artificial data sets with known properties before using actual data to perform an empirical analysis. Known 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 exclusion restriction and IV strength (in the case of an IV-based method), the degree of serial correlation of observed and unobserved variables (in the case of panel data and/or the use of lagged variables), deviations from independently and identically distributed (iid) 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 under which conditions the estimation approach succeeds in recovering the effects used to create the artificial data. Using artificial data to test the code/inference is a integral part of the data-generating-process centric workflow proposed in Storm et al. (2024).

- Use multiple approaches and critically discuss what can be learnt from the results of different methods as, in most cases, there may not be a single best estimation approach as each approach has its advantages/drawbacks.

Even if these guidelines are followed, when investigating causal effects with observational data, except in very rare cases, one cannot be 100% certain that all the required assumptions are completely fulfilled. Therefore, as a precaution, one should refrain from 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", "A leads to a change in B", etc. Instead, one can write "A is positively related to B", "A is negatively related to B", "A is associated with

B", "A is conditionally associated with B", etc. It is important to use consistent language throughout the entire paper. 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, e.g., Aïhounton and Henningsen, 2024).

Acknowledgements

The authors would like to thank 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.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review*, 93(1):113–132.

Acemoglu, D., Johnson, S., and Robinson, J. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5):1369–1401.

Alix-Garcia, J., McIntosh, C., Sims, K. R., and Welch, J. R. (2013). The ecological footprint of poverty alleviation: evidence from Mexico's oportunidades program. *Review of Economics and Statistics*, 95(2):417–435.

Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2(2):105–110.

Angrist, J. (2014). The perils of peer effects. *Labour Economics*, 30:98–108.

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

Angrist, J. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30.

Arellano, M. and Bond, S. (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 Bover, O. (1995). Another look at the instrumental variables estimation of error components models. *Journal of Econometrics*, 68:29–51.

Aronow, P. M. and Carnegie, A. (2013). Beyond late: Estimation of the average treatment effect with an instrumental variable. *Political Analysis*, 21(4):492–506.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730.

Athey, S. and Imbens, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Auci, S., Barbieri, N., Coromaldi, M., and Michetti, M. (2021). Climate variability, innovation and firm performance: evidence from the European agricultural sector. *European Review of Agricultural Economics*, 48(5):1074–1108.

Aïhounton, G. and Henningsen, A. (2024). Does organic farming jeopardize food security of farm households in Benin? *Food Policy*, 124:102622.

Bartik, T. (1991). *Who Benefits from State and Local Economic Development Policies?* W.E. Upjohn Institute for Employment Research, Kalamazoo, MI.

Basu (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., Heckelei, T., and Storm, H. (2021). Machine learning in agricultural economics (chapter 83). In Barrett, C. B. and Just, D. R., editors, *Handbook of Agricultural Economics*, volume 5, pages 4551–4612. Elsevier.

Bazzi, S. and Clemens, M. A. (2013). Blunt instruments: Avoiding common pitfalls in identifying the causes of economic growth. *American Economic Journal: Macroeconomics*, 5(2):152–186.

Belay, D. G. and Jensen, J. D. (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 Jensen, J. D. (2022). Quantitative input restriction and farmers' economic performance: Evidence from Denmark's yellow card initiative on antibiotics. *Journal of Agricultural Economics*, 73(1):155–171.

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, March 15-17, 2012, `https://marcfbellemare.com/wordpress/wp-content/uploads/2012/03/BellemareSCCKeynote.pdf` (accessed in December 2024).

Bellemare, M. (2015). Metrics Monday: What to do with endogenous control variables? Blog Post, `https://marcfbellemare.com/wordpress/11057` (accessed in December 2024).

Bellemare, M., Bloem, J., and Wexler, N. (2024). The paper of how: Estimating treatment effects using the front-door criterion. *Oxford Bulletin of Economics and Statistics*, 86(4):951–993.

Bellemare, M., Masaki, T., and Pepinsky, T. (2017). Lagged explanatory variables and the estimation of causal effect. *The Journal of Politics*, 79(3):949–963.

Bellemare, M. and Novak, L. (2017). Contract farming and food security. *American Journal of Agricultural Economics*, 99:357–378.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80:2369–2429.

Betz, T., Cook, S., and Hollenbach, F. (2018). On the use and abuse of spatial instruments. *Political Analysis*, 26(4):474–479.

Bilinski, A. and Hatfield, L. A. (2020). Nothing to see here? non-inferiority approaches to parallel trends and other model assumptions. arXiv, `https://arxiv.org/abs/1805.03273` (accessed in December 2024).

Blanc, E. and Schlenker, W. (2017). The use of panel models in assessments of climate impacts on agriculture. *Review of Environmental Economics and Policy*, 11(2):258–279.

Blattman, C. (2010). The cardinal sin of matching. Blog Post, `https://chrisblattman.com/blog/2010/10/27/the-cardinal-sin-of-matching/` (accessed in December 2024).

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87:11–143.

Borusyak, K., Hull, P., and Jaravel, X. (2025). A practical guide to shift-share instruments. *Journal of Economic Perspectives*, forthcoming.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event study designs: Robust and efficient estimation. *The Review of Economic Studies*, 91(6):3253–3285.

Bovay, J. (2025). Shaming, stringency, and shirking: Evidence from food-safety inspections. *American Journal of Agricultural Economics*, 107(1):152–180.

Breuer, M. (2022). Bartik instruments: An applied introduction. *Journal of Financial Reporting*, 7(1):49–67.

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

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. C. (2024). Event studies with a continuous treatment. *AEA Papers and Proceedings*, 114:601–05.

Callaway, B. and Sant'Anna, P. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Calonico, S., Cattaneo, M., Farrell, M., and Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, 17(2):372–404.

Calonico, S., Cattaneo, M., and Titiunik, R. (2015). rdrobust: an R package for robust nonparametric inference in regression-discontinuity designs. *The R Journal*, 7(1):38.

Cameron, A. C., Gelbach, J., and Miller, D. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Cragg, J. and Donald, S. (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. and D'Haultfœuille, X. (2023a). Two-way fixed effects and difference-in-differences estimators with several treatments. *Journal of Econometrics*, 236(2):105480.

de Chaisemartin, C. and D'Haultfœuille, X. (2023b). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey. *The Econometrics Journal*, 26(3):C1–C30.

Deines, J., Guan, K., Lopez, B., Zhou, Q., White, C., Wang, S., and Lobell, D. (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., Wang, S., and Lobell, D. (2019). Satellites reveal a small positive yield effect from conservation tillage across the US corn belt. *Environmental Research Letters*, 14(12):124038.

Deryugina, T. (2017). The fiscal cost of hurricanes: Disaster aid versus social insurance. *American Economic Journal: Economic Policy*, 9(3):168–198.

Di Falco, S., Veronesi, M., and Yesuf, M. (2011). Does adaptation to climate change provide food security? A micro-perspective from Ethiopia. *American Journal of Agricultural Economics*, 93:825–842.

Diegert, P., Masten, M., and Poirier, A. (2023). Assessing omitted variable bias when the controls are endogenous. *arXiv, https://arxiv.org/abs/2206.02303 (accesses in December 2024).*

Fan, L., Stevens, A., and Thomas, B. (2022). Consumer purchasing response to mandatory genetically engineered labeling. *Food Policy*, 110:102296.

Finger, R., Grebitus, C., and Henningsen, A. (2023). Replications in agricultural economics. *Applied Economics Perspectives and Policy*, 45(3):1258–1274.

Freyaldenhoven, S., Hansen, C., Pérez Pérez, J., and Shapiro, J. M. (2021). Visualization, identification, and estimation in the linear panel event-study design. NBER Working Papers 29170, National Bureau of Economic Research.

Freyaldenhoven, S., Hansen, C., and Shapiro, J. M. (2019). Pre-event trends in the panel event-study design. *American Economic Review*, 109(9):3307–38.

Frölich, M. (2008). Parametric and nonparametric regression in the presence of endogenous control variables. *International Statistical Review*, 76:214–227.

Gardner, J., Thakral, N., Tô, L. T., and Yap, L. (2024). Two-stage differences in differences. Working Paper, `https://neilthakral.github.io/files/papers/2sdd.pdf` (acessed in December 2024).

Garg, T. and Shenoy, A. (2021). The ecological impact of place-based economic policies. *American Journal of Agricultural Economics*, 103(4):1239–1250.

Gibson, J. (2019). Are you estimating the right thing? An editor reflects. *Applied Economic Perspectives and Policy*, 41(3):329–350.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Groher, T., Heitkämper, K., Walter, A., Liebisch, F., and Umstätter, C. (2020). Status quo of adoption of precision agriculture enabling technologies in Swiss plant production. *Precision Agriculture*, 21:1327–1350.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*. PMLR.

Hausman, C. and Rapson, D. S. (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 Bresnahan, T. and Gordon, R., editors, *The Economics of New Goods*. University of Chicago Press.

Heckelei, T., Hüttel, S., Odening, M., and Rommel, J. (2023). The p-value debate and statistical (mal) practice – implications for the agricultural and food economics community. *German Journal of Agricultural Economics*, 72(1):47–67.

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. *The Journal of Human Resources*, 32(3):441–462.

Hirsch, S., Khalilov, M., Dalhaus, T., and Mishra, A. (2023). Firm names and profitability in German food processing. *European Review of Agricultural Economics*, 50:1103–1139.

Hirsch, S., Tiboldo, G., and Lopez, R. (2018). A tale of two Italian cities: brand-level milk demand and price competition. *Applied Economics*, 50(49):5239–5252.

Hoy, K. and Wrenn, D. (2020). The effectiveness of taxes in decreasing candy purchases. *Food Policy*, 97:101959.

Huntington-Klein, N. (2021). *The effect: An introduction to research design and causality.* Chapman and Hall/CRC.

Imbens, G. (2024). Causal inference in the social sciences. *Annual Review of Statistics and its Application*, 11:123–152.

Imbens, G. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62:467–475.

Ioannidis, J. and Doucouliagos, C. (2013). What's to know about the credibility of empirical economics? *Journal of Economic Surveys*, 27(5):997–1004.

Jafari, Y., Koppenberg, M., Hirsch, S., and Heckelei, T. (2023). Markups and export behavior: Firm-level evidence from the French food processing industry. *American Journal of Agricultural Economics*, 105(1):174–194.

Jiang, W. (2017). Have instrumental variables brought us closer to the truth. *The Review of Corporate Finance Studies*, 6(2):127–140.

Keane, M. and Neal, T. (2023). Instrument strength in IV estimation and inference: A guide to theory and practice. *Journal of Econometrics*, 235(2):1625–1653.

Keane, M. and Neal, T. (2024). A practical guide to weak instruments. *Annual Review of Economics*, 16:185–212.

King, G. and Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454.

Kleibergen, F. and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1):97–126.

Koppenberg, M., Mishra, A., and Hirsch, S. (2023). Food aid and violent conflict: A review and empiricist's companion. *Food Policy*, 121:102542.

Krishnan, P. and Patnam, M. (2014). Neighbors and extension agents in Ethiopia: Who matters more for technology adoption? *American Journal of Agricultural Economics*, 96(1):308–327.

Kurz, C. (2022). Augmented inverse probability weighting and the double robustness property. *Medical Decision Making*, 42(2):156–167.

Lal, A., Lockhart, M., Xu, Y., and Zu, Z. (2024). How much should we trust instrumental variable estimates in political science? Practical advice based on 67 replicated studies. *Political Analysis*, 32(4):521–540.

Lee, D., McCrary, J., Moreira, M., and Porter, P. (2022). Valid t-ratio inference for IV. *American Economic Review*, 112(10):3260–3290.

Li, H. and Zhu, J. (2024). Property rights and land quality. *American Journal of Agricultural Economics*, 106(5):1619–1647.

Liu, L., Wang, Y., and Xu, Y. (2024). A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 68(1):160–176.

Maggio, G., Mastrorillo, M., and Sitko, N. (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., Spielman, D., Lybbert, T., and Gulati, K. (2015). Leveling with friends: Social networks and indian farmers' demand for a technology with heterogeneous benefits. *Journal of Development Economics*, 116:223–251.

Martinsson, E., Hansson, H., Mittenzwei, K., and Storm, H. (2024). Evaluating environmental effects of adopting automatic milking systems on Norwegian dairy farms. *European Review of Agricultural Economics*, 51(1):128–156.

Mason, N., Jayne, T., and Mofya-Mukuka, R. (2013). Zambia's input subsidy programs. *Agricultural Economics*, 44(6):613–628.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

McKenzie, D. (2018). I'm not a fan of leave-one-out/spatial instruments. World Bank Blogs, `https://blogs.worldbank.org/en/impactevaluations/im-not-fan-leave-one-outspatial-instruments` (accessed in December 2024).

McKenzie, D., Gibson, J., and Stillman, S. (2010). How important is selection? Experimental vs. non-experimental measures of the income gains from migration. *Journal of the European Economic Association*, 8(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*, forthcoming.

Millimet, D. and Bellemare, M. (2023). Fixed effects and causal inference. Technical Report No. 16202, IZA Discussion Paper.

Morgan, S. and Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2nd ed. edition.

Mullally, C. and Chakravarty, S. (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(1):44–56.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy*, 9(4):513–548.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342.

Newey, W. and Powell, J. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578.

Noack, F., Larsen, A., Kamp, J., and Levers, C. (2022). A bird's eye view of farm size and biodiversity: The ecological legacy of the iron curtain. *American Journal of Agricultural Economics*, 104(4):1460–1484.

Okui, R., Small, D., Tan, Z., and Robins, J. (2012). Doubly robust instrumental variable regression. *Statistica Sinica*, pages 173–205.

Olden, A. and Møen, J. (2022). The triple difference estimator. *The Econometrics Journal*, 25(3):531–553.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37:187–204.

Pan, Y., Smith, S., and Sulaiman, M. (2018). Agricultural extension and technology adoption for food security: Evidence from Uganda. *American Journal of Agricultural Economics*, 100(4):1012–1031.

Quisumbing, A., Ahmed, A., Gilligan, D., Hoddinott, J., Kumar, N., Leroy, J., Menon, P., Olney, D., Roy, S., and Ruel, M. (2020). Randomized controlled trials of multi-sectoral programs: Lessons from development research. *World Development*, 127:104822.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *The Review of Economic Studies*, 90(5):2555–2591.

Rao, E. and Qaim, M. (2013). Supermarkets and agricultural labor demand in Kenya: A gendered perspective. *Food Policy*, 38:165–176.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel-trends. *American Economic Review: Insights*, 4(3):305–322.

Roth, J. and Sant'Anna, P. H. C. (2023). Efficient estimation for staggered rollout designs. *Journal of Political Economy Microeconomics*, 1(4):669–709.

Ruml, A. and Qaim, M. (2021). New evidence regarding the effects of contract farming on agricultural labor use. *Agricultural Economics*, 52:51–66.

Sant'Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.

Schulz, D., Stetter, C., Muro, J., Spekker, J., Börner, J., Cord, A., and Finger, R. (2024). Trade-offs between grassland plant biodiversity and yields are heterogenous across Germany. *Communications Earth & Environment*, 5(1):514.

Sellare, J., Meemken, E., and Qaim, M. (2020a). Fairtrade, agrochemical input use, and effects on human health and the environment. *Ecological Economics*, 176:106718.

Sellare, J., Meemken, E.-M., Kouamé, C., and Qaim, M. (2020b). Do sustainability standards benefit smallholder farmers also when accounting for cooperative effects? evidence from Côte d'Ivoire. *American Journal of Agricultural Economics*, 102(2):681–695.

Smale, M. and Mason, N. (2014). Hybrid seed and the economic well-being of smallholder maize farmers in Zambia. *Journal of Development Studies*, 50(5):680–695.

Staiger, D. and Stock, J. (1997). Instrumental variables with weak instruments. *Econometrica*, 65:557–586.

Stata Press (2023). *Stata Extended Regression Models Reference Manual, Release 18.* `https://www.stata.com/manuals/erm.pdf` (accessed in December 2024).

Stetter, C., Mennig, P., and Sauer, J. (2022). Using machine learning to identify heterogeneous impacts of agri-environment schemes in the eu: A case study. *European Review of Agricultural Economics*.

Storm, H., Baylis, K., and Heckelei, T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3):849–892.

Storm, H., Heckelei, T., and Baylis, K. (2024). Probabilistic programming for embedding theory and quantifying uncertainty in econometric analysis. *European Review of Agricultural Economics*, 51(3):589–616.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 255(2):175–199.

Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., and Lewis, G. (2019). Machine learning estimation of heterogeneous treatment effects with instruments. arXiv, `http://arxiv.org/abs/1905.10176` (accessed in December 2024).

Tabe-Ojong, M., Mausch, K., Woldeyohanes, T., and Heckelei, T. (2022). Three hurdles towards commercialisation: Integrating subsistence chickpea producers in the market economy. *European Review of Agricultural Economics*, 49(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(5):1563–1584.

Terza, J., Basu, A., and Rathouz, P. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3):531–543.

Todd, P. and Wolpin, K. (2023). The best of both worlds: Combining randomized controlled trials with structural modeling. *Journal of Economic Literature*, 61(1):41–85.

Verhofstadt, E. and Maertens, M. (2014). Smallholder cooperatives and agricultural performance in Rwanda: Do organizational differences matter? *Agricultural Economics*, 45:39–52.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, Y. and Bellemare, M. (2020). Lagged variables as instruments. Working Paper, `https://marcfbellemare.com/wordpress/wp-content/uploads/2020/09/WangBellemareLagIVsJuly2020.pdf` (accessed in December 2024).

Westreich, D. and Greenland, S. (2013). The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, 177(4):292–298.

Windmeijer, F. (2024). Testing underidentification in linear models, with applications to dynamic panel and asset pricing models. *Journal of Econometrics*, 240(2):105104.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd ed. edition.

Wooldridge, J. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(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, `http://dx.doi.org/10.2139/ssrn.3906345` (accessed in December 2024).

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal*, 26(3):C31–C66.

Wuepper, D. (2020). Does culture affect soil erosion? Empirical evidence from Europe. *European Review of Agricultural Economics*, 47(2):619–653.

Wuepper, D., Borrelli, P., and Finger, R. (2020a). Countries and the global rate of soil erosion. *Nature Sustainability*, 3(1):51–55.

Wuepper, D. and Finger, R. (2023). Regression discontinuity designs in agricultural and environmental economics. *European Review of Agricultural Economics*, 50(1):1–28.

Wuepper, D., Le Clech, S., Zilberman, D., Mueller, N., and Finger, R. (2020b). Countries influence the trade-off between crop yields and nitrogen pollution. *Nature Food*, 1(11):713–719.

Wuepper, D., Wimmer, S., and Sauer, J. (2021). Does family farming reduce rural unemployment? *European Review of Agricultural Economics*, 48(2):315–337.

Wuepper, D., Yesigat Ayenew, H., and Sauer, J. (2018). Social capital, income diversification and climate change adaptation: Panel data evidence from rural Ethiopia. *Journal of Agricultural Economics*, 69(2):458–475.

Young, A. (2022). Consistency without inference: Instrumental variables in practical application. *European Economic Review*, 147:104112.

Zou, B., Chen, Y., Mishra, A., and Hirsch, S. (2024). Agricultural mechanization and the performance of the local Chinese economy. *Food Policy*, 125:102648.