Understanding Shift-Share Instruments: an Application

Seung Chul Lee

Advisor(s): Dr. Aaron Schein

Approved _____

Date _____

February 10, 2023

Abstract

In this paper, I explore shift-share instrumental variables (SSIVs), which have recently become a popular causal inference tool in the economics literature. Shift-share instruments are a special case of instrumental variables (IVs), which arise in applications where the outcome of interest is at a local level and the treatment of interest can be partitioned into both local and group levels. I apply SSIVs to an empirical problem of whether increase in immigration induces more voter registration for the Republican party using county-level US data. I find that the estimated effect is not statistically significant. I further conduct a series of simulations to check whether it is plausible to have a true positive effect but an insignificant result in the presence of large random errors.

Acknowledgements

In retrospect, the countless hours of effort that now culminates in the completion of this program has instilled a great sense of confidence in myself as a researcher. I feel extremely lucky to have had the privilege of learning from the best, to the extent that I now understand why people write this stuff in their dissertations. First, I would like to express my utmost gratitude to my academic advisor, Professor Aaron Schein, for the invaluable guidance that he provided, without which this thesis would not have been possible. I would also like to thank our Departmental Master's Advisor and committee member, Professor Mei Wang, for the immense support she provided throughout my tenure in the program. I am also deeply indebted to Professor Emanuele Colonnelli at the Booth School of Business, the research assistantship with whom introduced me to the topic of this work. Furthermore, I convey the most heartfelt appreciation to all my friends in the Department of Statistics for the dearest moments along this rigorous path. Last but not least, I would like to thank my family for the unconditional support that made this journey feasible.

Contents

1 Introduction

A crow flies, a pear falls.

- a Korean proverb[1]

Causal inference is a difficult task. Indeed, most of the common methods in a statistician's arsenal to learn causality are grounded upon assumptions that can never be known. The popular Neyman-Rubin Causal Model, or the potential outcomes framework, posits the existence of hypothetical outcomes: one under treatment and the other under the absence of treatment. In reality, a researcher can only observe either one, but never both. This setting creates a specific missing value problem, where 50% of the potential outcomes is always missing. In observational studies, often without the benefit of randomized control trials (RCTs), such missing values must be substituted to estimate a causal effect. Each causal inference method is a unique way of imputing the missing outcomes, which requires different sets of untestable assumptions. For instance, we have the parallel trends assumption for difference-in-differences (DID) and exclusion restriction for instrumental variables (IVs).

Despite such challenges, scientific studies of causality are inevitable. Notably, IVs have played a pivotal role across multiple academic disciplines in this process. Instruments are desirable tools that allows one to learn the causal effect by removing unobserved confounding effects. In order to qualify as an IV, the variable must meet the assumptions of relevance and exogeneity, the meanings of which are discussed in more depth in Section 2. Again, these conditions are not fundamentally testable and must be believed. Most literature using IVs goes at length to justify the validity of their proposed instrument, often relying on intuition, logic or prior literature. Naturally, there have been efforts to partially relax the assumptions that are more plausible to argue for. The shift-share instrumental variable (SSIVs) is a special instance. SSIVs impose a certain structure to the treatment *intensity*, whenever the treatment and response can be measured at some local units and treatment can be subdivided further into group-specific effects. A canonical example is looking at geographical units as local units and industries as groups.

The aim of this paper is twofold: (1) obtain a coherent understanding of shift-share instrumental variables; and (2) apply it to a real-world data set. For the empirical analysis, I use an SSIV to study the effect of immigration on voter support for the Republican party in

[1]It is a phrase used to warn against interpreting two independent, concurrent events as causal. The saying dates back to the reign of King Injo of Joseon Dynasty in early 17$^{\text{th}}$ century.

the decade leading up to the Trump presidency. This is a particularly interesting question given the outcome of the 2016 Presidential Elections and how the Republican campaign fueled anti-immigrant sentiment. Such political retaliations against immigration and swings to conservative parties have been well documented for many other countries outside the US (Halla et al., 2017; Dustmann et al., 2019; Becker and Fetzer, 2016). My initial analysis contends that there is no significant evidence for such a trend in the United States, but Monte Carlo simulations suggest that it may be the fault of data granularity and noise in calculating the SSIV.

The remainder of the paper is organized as follows. Section 2 provides a thorough discussion of shift-share instruments. Section 3 presents the empirical application. Section 4 concludes.

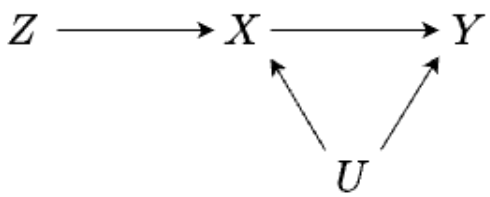
2 Shift-Share Instruments

Prior to a formal discussion of shift-share instruments, I briefly discuss instrumental variables (IVs) and its role in studying causality. Normally, a researcher wishes to learn about the effect of some treatment X on some response Y. A naïve approach would be to estimate a simple linear regression of the following form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{1}$$

However, is it appropriate to interpret the coefficient β_1 as a "causal effect"? The answer would be yes if X were assigned randomly to each individual as in an RCT, *i.e.*, $X \perp\!\!\!\perp \varepsilon$. In other cases, this will not necessarily be true. To illustrate, consider a chain of causal relationships depicted in Figure 1. In this graph, a researcher is interested in the directional

Figure 1: Graph of Causal Relationship for Instrumental Variables



effect of X on Y. Suppose there is another variable U that affects both X and Y. Under the model defined by Equation 1, the impact of U lies in the error term ε and one cannot

consistently estimate the effect β_1. In such settings, we call the variable U to be a *confounding variable* or a *confounder*. It is also often the case that confounders are unobservable.

One method that tackles this problem of unobservable confounding is the IV. Consider a variable Z in Figure 1 that is unrelated to the unobservable U but affects the treatment of interest X. If the statistician were to acquire such a variable, one could learn the variation in X that is caused purely by Z and not driven by U. In other words, if Z is orthogonal to U and we project X onto Z, we should be able to isolate the effect of X that is uncorrelated with U. Such a variable Z is called an *instrumental variable* or an *instrument*.

To have a valid Z, we would require that Z is correlated with X, so that we are not considering a completely irrelevant variable. This is referred to as the *relevance* condition. However, how to formalize the idea that a variable Z is irrelevant to U is quite different with respect to discipline, namely statistics and econometrics. Here, I try to provide a succinct summary of the two approaches. For a more detailed treatment of the subject, I direct the reader to Angrist et al. (1996) and Imbens (2014). First, a statistician's approach is to invoke (conditional) independence of potential outcomes in the context of the Neyman-Rubin Causal Model. That is,

$$Y(Z, X) = Y(Z', X) = Y(X), \ \forall Z, Z' \text{ and } \forall X, \tag{2}$$

where $Y(Z, X)$ denotes the potential outcome given an instrument value of Z and treatment value of X. To elaborate, the potential outcomes (*"what could have been observed"*) are completely determined by the treatment and independent of the instrument Z. To briefly discuss the concept of potential outcomes, it is a fictitious value that would have been observed for a person had they received a different treatment. In the canonical setting of binary treatments[2], the researcher will only observe either the treated condition (denoted $Y(1)$) or the control condition (denoted $Y(0)$) for an individual. As an individual will either receive or not receive the treatment, a statistician can never observe both outcomes. Hence, the assumption given by Equation (2) is completely philosophical in nature and cannot be tested with a *post hoc* analysis of observed data.

On the other hand, an econometrician would prefer the use of structural equations. For instance, in addition to Equation (1), one can specify the following:

$$X_i = \gamma_0 + \gamma_1 Z_i + \nu_i. \tag{3}$$

[2]This would be when $X \in \{0, 1\}$, where $X = 1$ if the person receives a certain treatment condition and $X = 0$ otherwise.

This is what is commonly referred to as a *first stage*, named after the two-stage least squares (2SLS) that is often used to estimate IV structural models (Angrist and Imbens, 1995). The assumption of irrelevance or the *exogeneity* condition is formulated as

$$\mathbb{E}[Z_i \cdot \nu_i] = \mathbb{E}[Z_i \cdot \varepsilon_i] = 0,$$

which is again something that has to be believed about the error terms. There is also a well-developed literature, such as Kleibergen and Zivot (2003), that takes a Bayesian approach to this setting, treating the first stage as a prior. The advantage of such structural models is that it is relatively easier to extend, *e.g.*, continuous X and Z, adding additional covariates for control, using a multidimensional Z, *etc.* However, it comes at the cost of interpretability of model assumptions and restrictions regarding heterogeneous treatment effects. Although the model assumptions are specified in a different manner, the estimator that both statisticians and econometricians use boils down to the 2SLS and are equivalent.

Even though instruments can be helpful in learning causality, searching for an instrument suffers from the same problem that U is unobservable. As briefly mentioned above, the underlying assumptions in both traditions are ultimately not testable. As a result, the literature employing IVs often resorts to a verbose discussion about the justifiability of the instrument. Hence, there have been efforts to come to a plausible setting. Shift-share (or Bartik) instrument is a good example of such an effort, which have recently become popular in the economics literature. Its application now spans a wide variety of topics[3]. The strategy takes advantage of the fact that each unit of observation may have a heterogeneous exposure to a common shock, while the shock itself may not be exogenous. It is often applied in settings where the outcome variable of interest can be subdivided into intuitive *local* units and the proposed cause of the said outcome can be subdivided into both *local* and *group-specific* units. A canonical example often invoked in the literature is defining *local* subdivisions as geographical units (*e.g.*, states or counties in the United States) and *group-specific* subdivisions as industries.

Goldsmith-Pinkham et al. (2020) and Borusyak et al. (2021) elaborate on two different approaches in validating the Bartik instrument. In the end, both yield in an equivalent

[3]To name some examples, Autor et al. (2013, 2019) and Borusyak et al. (2021) look at the effect of competition from Chinese products on economic outcome in the US. Jaravel (2019) study a traditional question on the effect of demand shocks on the price of goods. Acemoglu and Restrepo (2020) use shift-share instruments to learn the effect of robot adoptions on the outcome of the labor market. Greenstone et al. (2020) look at the causal relationship between local bank lending and employment growth in small firms. Nunn and Qian (2014) study whether US food aid induces a civil conflict in the receiving country. Finally, Tabellini (2020) studies the effect of immigration on local political and economic outcome in the US.

result in terms of estimation and inference. However, the intrinsic argument required to justify the instrument becomes very different. Following the notations used in Goldsmith-Pinkham et al. (2020), the Bartik instrument takes advantage of two identities:

$$X_{lt} = \sum_k z_{lkt} g_{lkt}, \tag{4}$$

$$g_{lkt} = g_{kt} + \tilde{g}_{lkt}. \tag{5}$$

Equation (4) defines the growth of our variable of interest X_{lt} in a location l at time t as the weighted average of a set of local- and subgroup-specific shocks g_{lkt}, with weights z_{lkt} reflecting a heterogeneous exposure to this shock. In other words, it is the inner product of the "shares" (z) at some local level and the "shifts" (g) that are a determinant of X. Equation (5) further decomposes the local level variable into two parts, namely an overall, global level component g_{kt}, and an idiosyncratic, location-specific component $\tilde{g}_{lkt}$. Thus, a Bartik instrument can be formulated as:

$$B_{lt} = \sum_k z_{lk0} g_{kt}, \tag{6}$$

which is essentially a predictor of X_{lt} based on local shares and global shifts.

Let Y_{lt} be the outcome at time t for location l, X_{lt} be the regressor of our interest at time t for location l, and $\mathbf{D}_{lt}$ be a vector of control variables including fixed effects at time t for location l. Then, a valid structural model for learning the causal effect of X on Y is:

$$\begin{aligned} Y_{lt} &= \mathbf{D}_{lt}\beta_0 + X_{lt}\beta + \varepsilon_{lt}, \\ X_{lt} &= \mathbf{D}_{lt}\tau + B_{lt}\gamma + \eta_{lt}, \end{aligned}$$

where ϵ_{lt} and η_{lt} are random errors.

As with the usual IV, two assumptions are required for the Bartik instrument to be valid:

$$\begin{aligned} \mathbb{E}\left[B_{lt}\varepsilon_{lt} \mid \mathbf{D}_{lt}\right] = 0, \quad &\text{(exogeneity)} \\ \text{Cov}\left(B_{lt}, X_{lt} \mid \mathbf{D}_{lt}\right) \neq 0. \quad &\text{(relevance)} \end{aligned}$$

Of course, these assumptions can also be formulated using potential outcomes model, a version of which is done by Adão et al. (2019). There are largely two different approaches in achieving the above conditions. The first is exogeneity of shocks (g_{kt}) *à la* Borusyak et al. (2021), which relies on having a large number of subdivisions (k). According to their paper,

this approach is design-based in spirit and can be interpreted as a form of an instrument. The other is exogeneity in shares (z_{lkt}) *à la* Goldsmith-Pinkham et al. (2020). This formulation is similar to a difference-in-differences (DID) design, with differing degrees of treatment on each unit instead of a binary treatment indicator.

The models are equivalent under both approaches, relying on the heterogeneous exposure to a common shock to be exogenous. However, the arguments required to justify the method are vastly different. The former demands the statistician to argue that shocks (g) are plausibly unrelated to the outcome variable (Y) other than through its effect on X. The latter entails an argument that the initial allocation of shares (z) is plausibly unrelated to the outcome variable (Y) other than through its effect on X. In the end, the statistician obtains a valid instrument in B_{lt} as defined in Equation (6).

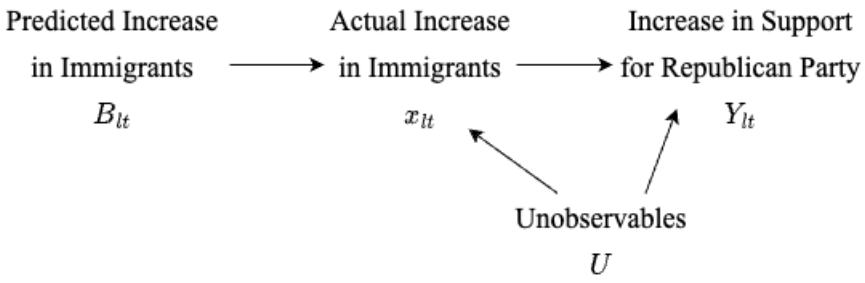
3 Empirical Application

To further explore the properties of shift-share instruments, I try using it on a real-world problem. For this application, I choose to study the causal relationship between immigration and political support for right-wing parties. To explicitly state the question, I ask: *"Does an increase in immigration cause an increase in support for the Republican party?"* This question is particularly interesting given the extensive discussions of the relationship following former President Donald Trump's victory in the 2016 US Presidential Elections. A similar phenomenon has been widely documented outside the United States by works such as Halla et al. (2017) for Austria, Dustmann et al. (2019) for Denmark, and Becker and Fetzer (2016) for the United Kingdom. Thus, the causal question that I pose is one that would nicely complement this growing literature on the political backlash of globalization.

3.1 Research Design

My design and formulation of the shift-share instruments is closely related to that of Tabellini (2020). Tabellini (2020) also looks at the effect of immigration on economic and political outcomes in the 1910s to 1930s with the IPUMS US Census data. He uses metropolitan areas as the local level and immigrant origin as the subgroup level. A key finding in this paper is that an increase in immigration leads to a decrease in support for the Democratic party in the 1932 and the 1936 Presidential elections, compared to the 1912 and 1916 elections. The underlying logic of employing the shift-share instruments is that migration tends to flow toward locations with a large ethnic community of the migrants' origin. That is, for instance, new Italian immigrants to the United States are more likely to settle in areas where

Figure 2: Graph of Causal Relationship for Voter Registration



there is a sizable Italian community. The critical assumption is that such ethnic clusters are formulated in the early days of settlement in America and are not meaningfully correlated in the concurrent economic and political environment of the area. To ensure this, the study focuses only on immigrants from select European countries. Hence, Tabellini (2020) relies on the exogeneity of shares to invoke the instrument. The resulting identification strategy can be depicted as the graph in Figure 2.

Although this study is a close relative, there are key departures that clearly differentiates it from what has been done. First, the period under scrutiny is 2005 to 2015, a decade leading up to a definitive event of Donald Trump's election. Thus, it attempts to shed light on the important claim that the outcome of 2016 elections was largely caused by anti-immigrant sentiments as suggested by many. Second, I use counties as my local level of shares as opposed to cities in Tabellini (2020). By extending beyond urban areas, I am able to test whether the findings also hold in rural areas. Since it has often been pointed out that the rural population responded more favorably toward Trump's rhetoric, it would be an important contribution to check whether this was true. Third, I also look at a broader scope of ethnic groups from all origins. This is also particularly meaningful, given the diversification of immigrant origins in the recent era and the negative light in which the Republican campaign portrayed Hispanic and Asian immigrants.

Due to this last distinction, the original setting in which Tabellini (2020) constructed his shift-share instrument may not be exactly replicable. The ethnic groups that formed in more recent years could have gravitated toward areas where there is more economic opportunities or less adverse attitudes toward foreign immigrants. Hence, although I form a similar instrument, the exogeneity of shifts is perhaps more reasonable. In fact, I posit that there was a negative exogenous shock in global immigration shifts toward the United States during this period. This would be due to the economic downturn caused by rising competition from Chinese manufacturers that became more pronounced during this time and the sub-

prime mortgage crisis in 2007. The adverse shock would have affected immigration from each country to differing degrees, creating exogenous variations in the number of migrants from each origin. Thus, based on these assumptions, I formulate the following structural model:

$$\Delta Repub_{lt} = \beta_0 + \beta_1 \Delta Imm_{lt} + \varepsilon_{lt}, \tag{7}$$

$$\Delta Imm_{lt} = \gamma_0 + \gamma_1 B_{lt} + \eta_{lt}. \tag{8}$$

B_{lt} denotes the shift-share instrument at county l at time t calculated as $\sum_k z_{lk0} g_{kt}$, where z_{lk0} is the initial share of immigrant population for ethnic group k at location l and g_{kt} is the global increase in immigration for ethnic group k at time t. ΔImm_{lt} is the actual increase of immigrants at location l at time t. $\Delta Repub_{lt}$ is the increase in the share of registered voters for the Republican party at location l at time t. ε_{lt} and η_{lt} are random errors. As discussed in Section 2, I have identification if

$$\mathbb{E}\left[B_{lt}\varepsilon_{lt}\right] = 0, \quad \text{(exogeneity)}$$

$$\text{Cov}\left(B_{lt}, \Delta Imm_{lt}\right) \neq 0, \quad \text{(relevance)}$$

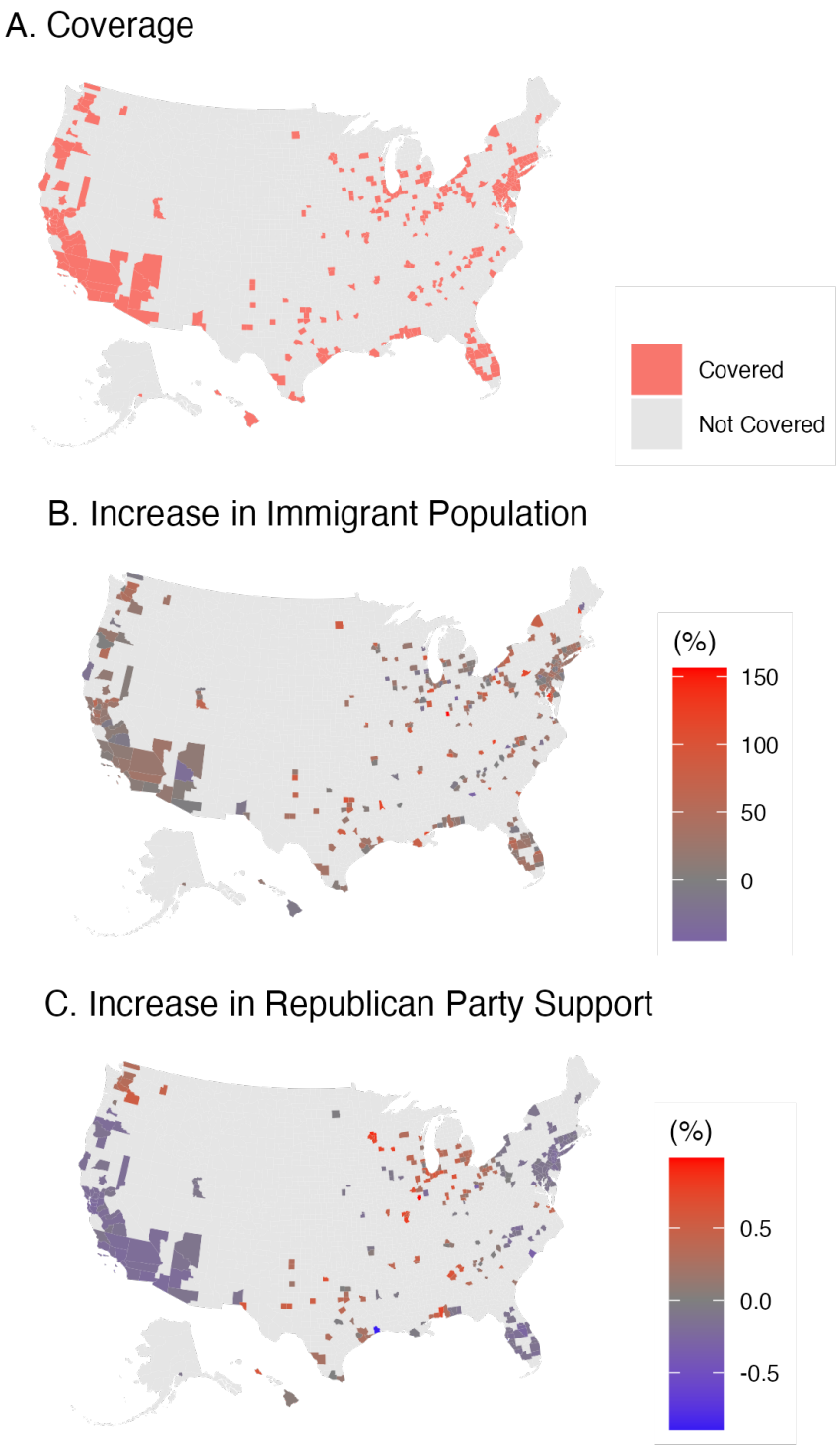
where exogeneity of B_{lt} is obtained through g_{kt} being exogenous.

3.2 Data

I obtain the data on immigrant growth and shares of ethnic groups for years 2005 and 2015 from the American Community Survey (ACS) provided by IPUMS USA (Ruggles et al., 2022). The original data set contains a total of 46,373,936 observations at the individual level. I resort to the ACS as the full count census data is only publicly available up until 1940. This creates major hurdles in terms of the granularity of data. The most salient feature would be the fact that not all people from all counties are observed due to sampling. There is also no guarantee that the collected data properly represents even those counties that are covered. The expectation is that the survey respondents are not subject to a selection bias. In this case, the proportions and increases in immigrant groups may be susceptible to noise but remain unbiased of the true changes in immigration. Obtaining a more granular measure of patterns in migration is one major potential avenue to which this analysis can be improved.

With the immigration information in ACS data set, I first compute the actual changes in immigrant proportions in each county from 2005 to 2015 (ΔImm_{lt}). To construct the shift-

Figure 3: Geographical View of the Data



share instrument, I use the same leave-one-out technique employed by Tabellini (2020). To elaborate, for county l, I exclude its observations when calculating the global increase in immigration g_{kt}. Thus, in reality, each county has a slightly different value of global growth applied to the computation of B_{lt}. This is often done in practice to further mitigate concerns about endogeneity (Card, 2001; Burchardi et al., 2019; Tabellini, 2020, among others).

As for my response variable, I rely on the voter file data provided by Professor Aaron Schein at The University of Chicago. This data set is identical to that used by Brown and Enos (2021), who explore the geographic distribution of partisanship with it. I aggregate individual voter registration in 2005 and 2015 based on the county of the person's residence to obtain the proportion of voters registered for the Republican party in each county for the two years. I then calculate the percentage increase over this ten-year period as my dependent variable.

Finally, I match the immigration and voter file data using the county FIPS code available in both data sets. I end up with a sample of 332 counties, which represents approximately 10% of the population of all counties in the United States. Panel A of Figure 3 illustrates the counties that are covered. Quite naturally, I find that the sample mostly contains counties that are relatively populous. Note that much of the Rocky Mountains region is excluded, whereas both coastal regions are better represented. Panel B shows the change in immigration population for the counties in sample. Panel C indicates the change in voter registration for the Republican party. The patterns observed in the plots are largely consistent with the idea that there was a surge in Republican support in the Midwest. Table 1 provides some summary statistics of the resulting data set.

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
ΔImm_{lt}	332	0.269	0.304	−0.444	1.562
B_{lt}	332	0.093	0.060	−0.082	0.279
$\Delta Repub_{lt}$	332	0.071	0.283	−0.895	0.983

3.3 Results

Now, I run the model defined by Equations (7) and (8) and present the resulting estimates using two-stage least squares (2SLS). Panel A of Table 2 provides the structural equation, *i.e.*, estimated values for Equation (7), and Panel B gives the values for Equation (8). I

find that the coefficient estimate $\hat{\beta}_1$, which would be my estimate for the causal effect, is negative at -5.208 and not statistically significant. The first stage reported in Panel B further suggests that the correlation between the instrument and endogenous regressor (*i.e.*, increase in immigrants) is fairly small. The F statistic of the first stage regression (not reported on table) is 0.2787, which is equivalent to the squared value of the t statistic. This value suggests weak identification, *i.e.*, violation of the relevance condition discussed

Table 2: Voter Registration Model Results

Panel A. Structural Equation

	Estimate	Std Error	t	p
$\hat{\beta}_0$	1.470	3.127	0.470	0.641
$\hat{\beta}_1$	−5.208	11.614	−0.448	0.656

Panel B. First Stage

	Estimate	Std Error	t	p
$\hat{\gamma}_0$	0.255	0.031	8.267	0.000
$\hat{\gamma}_1$	0.147	0.278	0.528	0.598

in Section 2. A commonly used heuristic to measure weak instruments is whether the first-stage F statistic falls below 10 (Angrist and Pischke, 2009). Thankfully, there is an extensive literature that provide some guidance to dealing with weak instruments (Staiger and Stock, 1997; Andrews and Stock, 2005). I choose to implement the weak identification robust inference technique suggested by Chernozhukov and Hansen (2008) to obtain a robust 95% confidence interval. This method requires a prespecified interval of values to test β_1 on. For this study, I choose $\mathcal{B} = [-1, 1]$ as my set. I find that the significance of β cannot be rejected on the entirety of this interval, which yields the entire set $\mathcal{B}$ as my confidence interval. That is, I have no evidence of statistical significance even with the help of weak identification robust inference.

The result I acquire is quite peculiar to say the least. It is not consistent with the extant literature on the relationship between immigration and support for conservative political parties in many other parts of the world. It is also contradictory to the prior finding by Tabellini (2020) in the early 20$^{\text{th}}$ century United States. If I were to take it at face value, I may conclude that the widespread trend in anti-immigrant sentiment is less acute in America. However, it would much more natural to search for flaws in the research design that may have yielded such an outcome. To test the possibility that the result is mostly driven by noise,

I conduct a Monte Carlo simulation study. This is in the spirit of a prior predictive check, where the statistician synthetically creates a data set according to a prior (or a structural model) and assesses its resemblance to the real-world data. I rely on Equations (7) and (8) and the underlying structure of the Bartik instrument for the simulation.

I now discuss the Monte Carlo simulation in more detail. Note that I try to create the simplest possible setting, almost as a toy example, to avoid complications that is not specific to the model. I first randomly draw vectors of initial shares at location l, $z_{l0} = (z_{l10}, z_{l20}, \dots)$, from a Dirichlet distribution, to ensure that the proportions sum to one. I add a small noise $\nu_l \overset{iid}{\sim} N(0, \sigma_1^2)$ to the initial proportions which are drawn from a normal distribution with mean zero and normalize to obtain the shares for the next period, *i.e.*,

$$z_{l,t+1} = \frac{z_{lt} + \nu}{\|z_{lt} + \nu\|}.$$

I then sample global shocks $g_{kt} \overset{iid}{\sim} U[0, 1]$ from a uniform distribution, which accounts for the fact that increase in immigration will likely be a positive number across all sending countries over the years. Using the simulated shares and global shifts, I construct the shift-share instrument as

$$B_{lt} = \sum_k z_{lk0} g_{kt}.$$

Next, I generate errors corresponding to those in Equations (7) and (8) from a bivariate normal distribution, i.e.,

$$\begin{bmatrix} \eta_{lt} \\ \epsilon_{lt} \end{bmatrix} \sim N_2(0, \Sigma),$$

where Σ is positive definite but not diagonal to model the confounding effect of U depicted in Figure 2. Now, using Equation (8), the endogenous regressor X_{lt} is created as

$$X_{lt} = \gamma B_{lt} + \eta_{lt},$$

and, finally, the response variable Y_{lt} is defined as

$$Y_{lt} = \beta X_{lt} + \epsilon_{lt}.$$

Note that I omit intercepts from Equations (7) and (8). Then, using the synthetic data, I estimate the SSIV model to obtain a point estimate $\hat{\beta}$. Finally, I apply the weak identification robust inference as in Chernozhukov and Hansen (2008) to get a 95% confidence interval within the chosen region of $\mathcal{B} = [-2, 2]$. I provide a step-by-step guide and the R code used

for the simulation in Appendix A.

To form a simple but realistic sample, I hypothesize the existence of 100 locations, *i.e.*, $l \in \{1, 2, \ldots, 100\}$, and 5 subgroups, *i.e.*, $k \in \{1, \ldots, 5\}$. This would roughly correspond to the different continents, such as Africa, East Asia-Pacific, Central Asia, Latin America, and Europe. I create two time periods, *i.e.*, $t \in \{0, 1\}$. For the specific parameterization, I choose $\alpha = \left(\frac{1}{5}, \frac{1}{5}, \ldots, \frac{1}{5}\right) \in \mathbb{R}^5$, which is *ad hoc*. To obtain a fairly small perturbation in the local shares, I choose $\sigma_1 = 0.01$ for the distribution of ν_l. To model the correlated structure of errors under endogeneity, I choose the covariance matrix of the bivariate normal as

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix},$$

where I purposefully assume relatively large variances of the error terms in an attempt to closely mirror the empirical result using actual data. To enforce weak identification, I posit $\gamma = 0.1$ for the effect of instrument B_{lt} on the regressor X_{lt}. Finally, I try two values of the true causal effect of interest, or $\beta \in \{0.1, 1\}$. The value 0.1 is the case when we have small, positive causal effect, whereas the value 1 can be considered sizable. In fact, a true effect of $\beta = 1$ would translate to 1% increase in the proportion of immigrant population in an area leading to 1% increase in the proportion of Republican voter registration. Also, to properly gauge the effect of the effect size alone, I use the same seed for the two samples which would generate identical random components. Not doing this may lead to differences purely due to the disparity in the generated values.

I now present the results of my simulation study. Table 3 summarizes the result of 200 trials for each choice of β. Panel A corresponds to the case when $\beta = 0.1$ and Panel B corresponds to the case when $\beta = 1$. "Estimate" refers to the estimated second-stage coefficient $\hat{\beta}$; "Lower Bound" refers to the lower bound of the robust 95% confidence interval; "Upper Bound" denotes the upper bound of the robust 95% confidence interval; "Significant?" refers to statistical significance, *i.e.*, whether 0 is not contained in the aforementioned confidence interval; and lastly, "Valid?" denotes whether the ground truth value of β was contained in the robust confidence interval. As I assume a large noise, I discover that the point estimates are extremely volatile. The reader can reaffirm that I have the same random parts from the fact that the standard deviations from both trials are identical. Since the intervals are rather uninformative, all of them include the ground truth in both simulations.

The most salient result is the fact that it is extremely rare to observe a statistically significant result. For the small effect simulations, a mere 1.5% of trials yielded statistical significance.

Table 3: Summary of Simulation Results

Panel A. Small Effect Simulation

Statistic	N	Mean	St. Dev.	Min	Max
Estimate	200	17.276	241.247	−75.559	3,410.241
Lower Bound	200	−1.882	0.451	−2.000	0.680
Upper Bound	200	1.872	0.380	0.110	2.000
Significant?	200	0.015	0.122	0	1
Valid?	200	1.000	0.000	1	1

Panel B. Large Effect Simulation

Statistic	N	Mean	St. Dev.	Min	Max
Estimate	200	18.176	241.247	−74.659	3,411.141
Lower Bound	200	−1.805	0.680	−2.000	1.580
Upper Bound	200	1.935	0.225	0.310	2.000
Significant?	200	0.060	0.238	0	1
Valid?	200	1.000	0.000	1	1

This is somewhat expected as it will be extremely difficult to detect a small signal amidst large noises. However, it is interesting to learn that, even with a relatively large effect ($\beta = 1$), only 6% of trials successfully rejected the null hypothesis that $\beta = 0$. In light of this, I can conjecture that it is quite plausible to observe insignificant results due to large noises in constructing the shift-share instrument.

Figure 4 presents the histogram of coefficient estimates ($\hat{\beta}$) from the 200 trials. I omit estimates with absolute values larger than 5 for visualization purposes. Again, Panel A depicts the histogram for small effect size simulations and Panel B is for the large effect size simulations. The red line denotes the ground truth. I find that, in fact, the estimates tend to be loosely clustered around the vicinity of the true value. Hence, the SSIV estimator seems to have little to no bias but inherently vulnerable to the size of the noises.

As the final part of my analysis, I try plotting the robust 95% confidence intervals for both simulations. Each horizontal line depicts a confidence interval from a single trial. Consistent with the values in Table 3, all intervals seem to contain the red line, which marks the ground truth value. Overall, the robust intervals are too large to be informative. There does not seem to be a clear positive trend, despite the true parameter being positive.

Figure 4: Histogram of Simulation Estimates

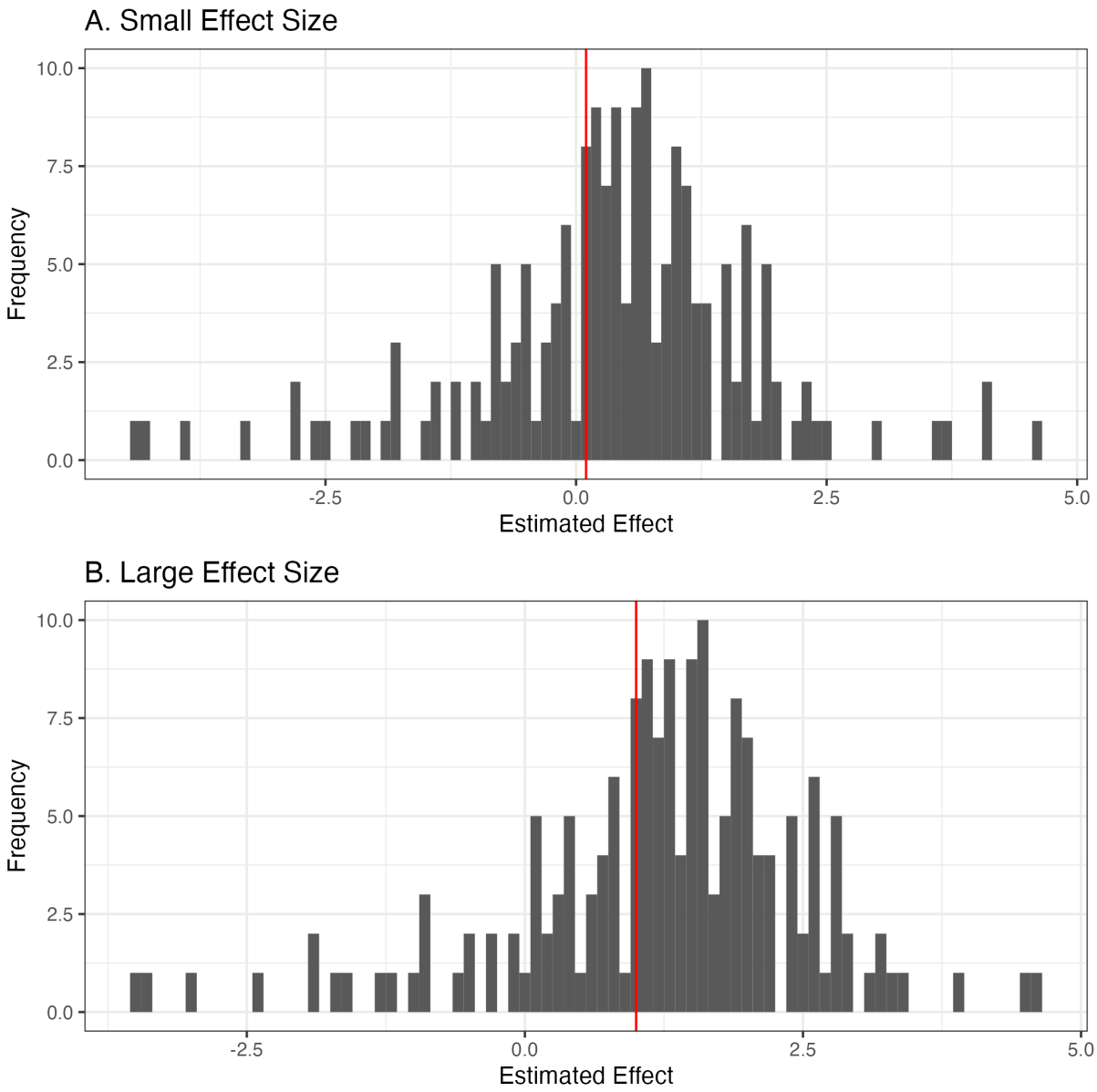
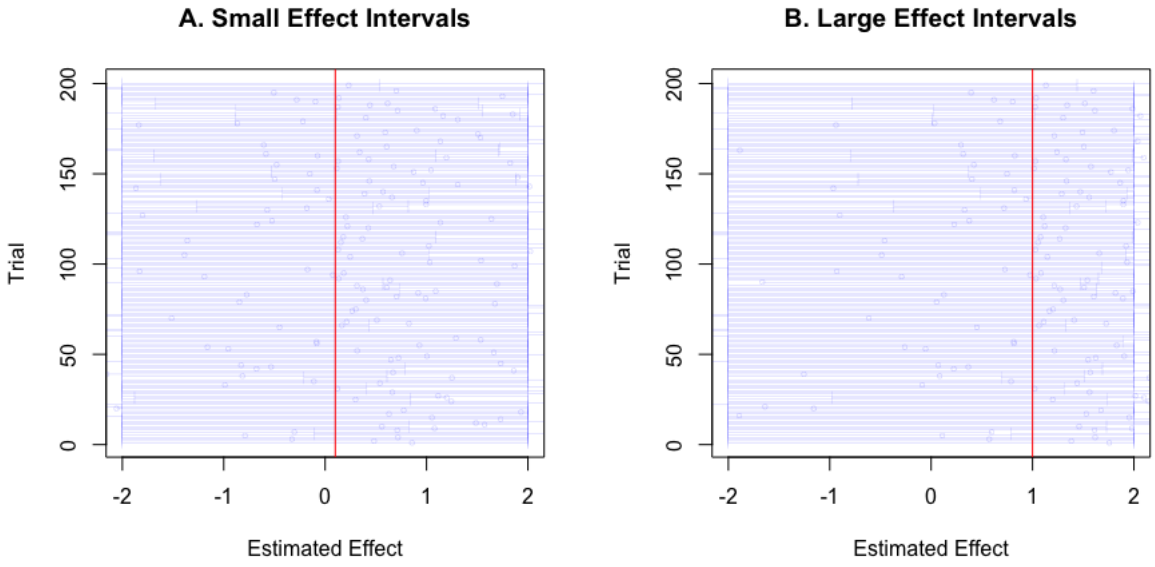


Figure 5: Simulated Robust Confidence Intervals



4 Conclusions

In this paper, I have reviewed the current literature on shift-share instrumental variables, an increasingly popular causal inference technique. SSIV depends on a *local* and *subgroup* level structure to be applicable and has the advantage that validity as an instrument can be obtained through two different channels: (1) independent local-level shares, or (2) independent global subgroup-level shifts. After studying the underlying assumptions for identification, I have tried applying this method to an interesting empirical problem of immigration and voter registration to the Republican party in the US. The estimate is counterintuitive to both the prior literature and the popular opinion. I test for the possibility that the estimator may not be robust to the presence of relatively large errors via a Monte Carlo simulation in the spirit of a prior predictive test.

While my analysis does not yield exciting results or solid evidence at the moment, it does point toward some possible paths that would be of interest to explore. First, SSIVs, much like the usual instruments, are sensitive to noise in the response and the endogenous regressor. This is of particular importance as an SSIV is likely to contain more noise from the separate estimation of shares and shifts. Hence, it is critical to gather the most granular data possible in their calculations. The current setting can be augmented with a more complete data set of immigration patterns than the American Community Survey. Second, weak identification robust inference may not be able to solve the problem for weakly identified

shift-share instruments. Such remedial properties would require more academic attention for advancement.

I plan to continue working on this topic beyond graduation, hopefully producing a full-blown research paper fit for publication in top economic or political science journals. The current plans for further developing this thesis project is to (1) apply randomization inference to test the sharp null hypothesis, (2) re-analyze the model given finer data on immigration, and (3) try variations of the prior predictive test.

A Appendix: Details of Monte Carlo Simulation

Simulation Steps

1. Randomly draw vectors of initial shares (z_{l0}) from a Dirichlet distribution.
2. Add small perturbation (ν) from a normal distribution for shares in subsequent periods.
3. Randomly draw global growth (g_{kt}) from a uniform distribution.
4. Randomly draw correlated errors for treatment (X_{lt}) and outcome (Y_{lt}) from a bivariate normal distribution.
5. Form shift-share instrument (B_{lt}), treatment (X_{lt}), and outcome (Y_{lt}), according to the structural equations.
6. Run IV regression with the simulated variables and obtain weak identification robust confidence intervals.
7. Repeat Steps 1-6 200 times.

R Code for Simulation

```
##### Required Packages #####
library(fixest)
library(ivreg)
library(lmtest)
library(sandwich)
library(extraDistr)
library(mvtnorm)

##### Analysis #####
# Weak Identification Robust Inference
weakiv.robust = function(x, y, z, vals, sig = 0.05){
  #' Weak Identification Robust Confidence Intervals for Instrumental Variables
  #'
  #' Function to obtain weak identification robust confidence interval
  #' a la Chernozhukov and Hansen (2008)
  #'
  #' @param x the vector of regressor
```

```
  #' @param y the vector of response
  #' @param z the vector of instrument
  #' @param vals the set B of values to try
  #' @param sig the significance level of confidence interval (default = 0.05)
  #' @returns a pair of numeric values for the lower bound and upper bound

  ivsig = function(b){
    #' Test Significance of Instrumental Variable
    #'
    #' Function to calculate the significance of coefficient on the instrument
    #'
    #' @param b value of beta (the coefficient on x) to try
    #' @returns 1 if significance of coefficient on the instrument is rejected,
    #' 0 otherwise

    temp_mod = lm((y - b*x) ~ z - 1)
    rej = (coeftest(temp_mod, vcov = sandwich)[4] < sig)
    return(rej)
  }

  region = data.frame() # initialize df to collect non-rejected region
  for(b in vals){
    temp = ivsig(b)
    region = rbind(region, c(b, temp))
  }
  colnames(region) = c("b", "z")

  reg = region$b[region$z == 0] # non-rejected region
  rob_conf_int = c(min(reg), max(reg))
  return(rob_conf_int)
}

sim.ssiv = function(loc, k, sig1, sig2, sig3, b1, alph, seed = 1){
  if(!is.na(seed)){
    set.seed(seed) # for replicability
  }
```

```
# Shares
z0 = rdirichlet(n = loc, alpha = rep(1/k, k)) # initial shares
nu = rnorm(n = loc, 0, sig1)
z1 = (z0 + nu) / rowSums(z0 + nu)

# Shifts
g_glob0 = runif(n = k, min = 0, max = 1) # global growth
g_glob1 = runif(n = k, min = 0, max = 1) # global growth

# Shift-share instrument
ssiv0 = rowSums(z0 * g_glob0)
ssiv1 = rowSums(z0 * g_glob1)

# Generate Errors (Correlated to mimic U)
covar0 = matrix(c(sig2, 0.5, 0.5, sig3), ncol = 2)
covar1 = matrix(c(sig2, 0.5, 0.5, sig3), ncol = 2)
err0 = rmvnorm(n = loc, mean = rep(0, 2), sigma = covar0)
err1 = rmvnorm(n = loc, mean = rep(0, 2), sigma = covar0)

# Endogenous regressor
eta0 = err0[, 1]
eta1 = err1[, 1]
x0 = alph*ssiv0 + eta0
x1 = alph*ssiv1 + eta1

# Response
eps0 = err0[, 2]
eps1 = err1[, 2]
y0 = b1*x0 + eps0
y1 = b1*x1 + eps1

sim.dat = as.data.frame(cbind(y1, x1, ssiv1))
sim.mod = feols(y1 ~ 1 | 0 | x1 ~ ssiv1, data = sim.dat)
#summary(sim.mod)
```

```
  sim.rob = weakiv.robust(x = x1, y = y1, z = ssiv1, vals = seq(-2, 2, by = 0.01))
  #sim.rob

  return(list(df = sim.dat, mod = sim.mod, rob.int = sim.rob))
}

# Small effect size
coefs = NULL
intervals = data.frame()
set.seed(1)
for(i in 1:200){
  temp = sim.ssiv(loc = 100, # number of localities
                  k = 5, # number of groups
                  sig1 = 0.01, # sd for random change in shares from t = 0 to t = 1
                  sig2 = 1, # sd for random local idiosyncratic growth
                  sig3 = 1, # sd for random variation in response
                  b1 = 0.1, # effect of x on y
                  alph = 0.1, # effect of ssiv on x (assumed to be small))
                  seed = NA
  )
  coefs = c(coefs, coef(temp$mod)[2])
  intervals = rbind(intervals, temp$rob.int)
}

# Large effect size
coef.large = NULL
inter.large = data.frame()
set.seed(1)
for(i in 1:200){
  temp = sim.ssiv(loc = 100, # number of localities
                  k = 5, # number of groups
                  sig1 = 0.01, # sd for random change in shares from t = 0 to t = 1
                  sig2 = 1, # sd for random local idiosyncratic growth
                  sig3 = 1, # sd for random variation in response
                  b1 = 1, # effect of x on y
                  alph = 0.1, # effect of ssiv on x (assumed to be small))
```

```
                  seed = NA
  )
  coef.large = c(coef.large, coef(temp$mod)[2])
  inter.large = rbind(inter.large, temp$rob.int)
}
```

References

Acemoglu, D. and P. Restrepo (2020). Robots and Jobs: Evidence from US Labor Markets. *Journal of Political Economy 128*(6), 2188–2244.

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

Andrews, D. and J. Stock (2005). Inference with weak instruments. Cowles Foundation Discussion Papers 1530, Cowles Foundation for Research in Economics, Yale University.

Angrist, J. D. and G. W. Imbens (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association 90*(430), 431–442.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association 91*(434), 444–455.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Number 8769 in Economics Books. Princeton University Press.

Autor, D., D. Dorn, and G. Hanson (2019). When Work Disappears: Manufacturing Decline and the Falling Marriage Market Value of Young Men. *American Economic Review: Insights 1*(2), 161–178.

Autor, D. H., D. Dorn, and G. H. Hanson (2013). The China Syndrome: Local Labor Market Effects of Import Competition in the United States. *American Economic Review 103*(6), 2121–2168.

Becker, S. O. and T. Fetzer (2016). Does Migration Cause Extreme Voting? CAGE Online Working Paper Series 306, Competitive Advantage in the Global Economy (CAGE).

Borusyak, K., P. Hull, and X. Jaravel (2021). Quasi-Experimental Shift-Share Research Designs. *The Review of Economic Studies 89*(1), 181–213.

Brown, J. R. and R. D. Enos (2021). The measurement of partisan sorting for 180 million voters. *Nature Human Behaviour 5*(8), 998–1008.

Burchardi, K. B., T. Chaney, and T. A. Hassan (2019). Migrants, Ancestors, and Foreign Investments. *The Review of Economic Studies 86*(4), 1448–1486.

Card, D. (2001). Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration. *Journal of Labor Economics 19*(1), 22–64.

Chernozhukov, V. and C. Hansen (2008). The reduced form: A simple approach to inference with weak instruments. *Economics Letters 100*(1), 68–71.

Dustmann, C., K. Vasiljeva, and A. Piil Damm (2019). Refugee Migration and Electoral Outcomes. *The Review of Economic Studies 86*(5), 2035–2091.

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

Greenstone, M., A. Mas, and H.-L. Nguyen (2020). Do Credit Market Shocks Affect the Real Economy? Quasi-experimental Evidence from the Great Recession and "Normal" Economic Times. *American Economic Journal: Economic Policy 12*(1), 200–225.

Halla, M., A. F. Wagner, and J. Zweimüller (2017, 03). Immigration and Voting for the Far Right. *Journal of the European Economic Association 15*(6), 1341–1385.

Imbens, G. W. (2014). Instrumental Variables: An Econometrician's Perspective. *Statistical Science 29*(3), 323–358.

Jaravel, X. (2019). The Unequal Gains from Product Innovations: Evidence from the U.S. Retail Sector*. *The Quarterly Journal of Economics 134*(2), 715–783.

Kleibergen, F. and E. Zivot (2003). Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics 114*(1), 29–72.

Nunn, N. and N. Qian (2014). US Food Aid and Civil Conflict. *American Economic Review 104*(6), 1630–1666.

Ruggles, S., S. Flood, R. Goeken, M. Schouweiler, and M. Sobek (2022). IPUMS USA: Version 12.0. [dataset], Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V12.0.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*(3), 557–586.

Tabellini, M. (2020). Gifts of the Immigrants, Woes of the Natives: Lessons from the Age of Mass Migration. *The Review of Economic Studies 87*(1), 454–486.