Final Project

73-374 Econometrics II

11/29/2020

Due: Friday, December 18, Midnight EST

The goal of this project is to apply your knowledge of econometrics to understanding of modern empirical economic research by reading, replicating the main results of, and critically evaluating a recent research paper examining an important policy issue using the methods discussed in this course. The result will be a report, produced by a group of 2 to 3 students, consisting of written analysis approximately 5-7 pages in length (single spaced) accompanied by several tables of statistical output and the accompanying **R** code which produces that output, partially replicating and critically evaluating the econometric methods in a recent empirical research paper in economics.

A list of suggested papers which your group could choose to replicate, including links and data sources, follows.

(1) Ambrus, Attila, Erica Field, and Robert Gonzalez. 2020. "Loss in the Time of Cholera: Long-Run Impact of a Disease Epidemic on the Urban Landscape." *American Economic Review*, 110 (2): 475-525. https://doi.org/10.1257/aer.20190759

(2) Bazzi, Samuel, Gabriel Koehler-Derrick, and Benjamin Marx. May 2020. "The Institutional Foundations of Religious Politics: Evidence from Indonesia." *The Quarterly Journal of Economics*, 135 (2): 845–911. https://doi.org/10.1093/qje/qjz038

These papers each use a novel data set and techniques we have learned about in class to answer important economic questions, like about the long run effects of infectious disease or of land redistribution. The paper by Ambrus, Field, and Gonzalez (henceforth AFG) reexamines data from one of the earliest known studies applying observational data for causal inference, John Snow's study of the 1854 London cholera outbreak, applying modern regression discontinuity methods to measure the effect of exposure to cholera on neighborhood house prices in the short and long run. The paper by Bazzi, Koehler-Derrick, and Marx (henceforth BKDM) uses data on a proposed land redistribution scheme in Indonesia which targeted landholdings of different sizes in different districts based on a cutoff rule in population density to study the effect of such redistribution schemes on donation to religious institutions (which were exempt from redistribution), and, over the long run, the effect that these donations had on the role of religion in politics.

Links to the papers, data, and replication files used for each paper will be made available on Canvas.

Groups:

You may choose any group of 2-3 classmates. Please use the Canvas tool (under the people tab) to sign up for groups. A tab labeled "Group Suggestions" has created randomly assigned groups as a suggestion to you if you do not know who to ask. These are not binding and you must move your name into a group under the heading "Final Project Groups" to submit the assignment. A post will also be created on Piazza to help facilitate searching for group members. As an additional suggestion, if you do not know which paper to choose, odd numbered groups are suggested to replicate AFG, even ones to replicate BKDM.

If you have not found a group by Monday, December 7th, please email me and I will attempt to facilitate organizing a group for you. The grade for the paper will be shared by all group members. This project contains many subcomponents which could be divided among group members, but *all* members should review the final product to ensure coherence and accuracy.

Format:

The report should be structured as a coherent essay, interspersed with appropriate equations and figures. Tables and graphs may be placed either interspersed with the text or at the end, clearly labeled so that it is apparent what information is being presented. The file can be created in word processing software or, preferably, in R Markdown, knitted to display all results. Code should be included, with comments to make clear what procedures are being performed and what the intended output is. This should be in the form of R code at the end of the file, after the essay and tables. No extra points will be provided for aesthetically pleasing formatting, but several R packages exist which make producing neatly formatted tables of results simple: the R textbook recommended in the syllabus provides examples of the use of the library **stargazer** for this purpose. External sources, if used, should be properly cited in any standard citation style.

Alternative Paper Choice options

If so desired, your group may choose a paper to replicate other than the ones listed. Selections from outside the list provided, subject to the following conditions, are encouraged, but subject to approval: please email me with your potential selection by Wednesday December 9th and I will review the choice and provide guidelines on what to include in the report.

You may choose a paper published the American Economic Review, Quarterly Journal of Economics, or American Economic Journal: Applied Economics, from 2015 to present such that

(i) The data set is available. Click the link for Data Set on the article page for AER or AEJ papers, or Supplementary Data for QJE papers: many papers, for a variety of reasons, do not provide the full data set used, so check the file to make sure it actually contains a complete data set.

(ii) The paper uses for its primary analysis some set of methods described in this class beyond ordinary least squares regression. For example, instrumental variables, panel data, nonlinear or nonparametric regression, or regression discontinuity. A nontrivial fraction of papers use methods beyond the scope of this class. You may choose such a paper if these are confined to supplementary analyses, but the primary analysis should be based on material we have covered. In such cases, your report need not cover or replicate these supplementary analyses.

I will post on Canvas a list of suggested papers which satisfy these requirements; you are not restricted to choosing one from the list, if you can find another, but for these papers I have at least verified that data is available.

As the degree to which the results of a research paper may be feasibly replicated, and the difficulty of doing so, may vary, assessment of the quality of this replication project will be made on a scale which reflects the challenges of the analysis chosen. Analyses demonstrating solid comprehension of a variety of methods explored in 73-374 will be assessed favorably, with the understanding that some analyses may be more challenging to reproduce and assess than others. In particular, finding that some results are not capable of being replicated will **not** result in lower scores, and partial or incomplete results for more complicated analyses which nevertheless demonstrate understanding of the concepts involved are to be expected. Do not hesitate to ask for advice if some components of the analysis cannot be deciphered.

Questions:

After reading through this assignment to determine the goals of the project, read the paper you chose, looking carefully at the question asked, the data and econometric methods the authors chose to answer the questions, and considering whether the methods are appropriate and if the results are consistent with the authors chosen interpretations. It is *not* necessary to understand all parts of the paper, such as the models providing theoretical motivation e.g., the model of neighborhood sorting in Appendix A of AFG, or detailed descriptions of historical or institutional context. These provide context and support the econometric modeling assumptions made by suggesting which variables may be related and how and potentially suggest functional forms or other variables that should be included, but a solid understanding of the empirical results can be achieved without these supplemental discussions. [1] While these components of the research are

[1] It may be helpful to follow the advice in the online lectures in Econometrics from Duke University available at https://youtu.be/ZlAKPtPZqyI on how to skim an empirical research paper to find and evaluate the important results, rather than

valuable, the assessment for this project should be of the econometric component of the research.

For the purpose of all questions in this project, “replicate” will mean - to perform the same analysis using the data provided, and report the results. - If the numbers are the same (up to rounding) as those in the table, note that you were able to replicate. - If you are not able to obtain the same results, and the results differ in some nontrivial way, report this, and if possible, attempt a brief explanation.

For example, the mathematical definition of some reported objects may not be clear, the calculations may differ, or there may be some error in the original analysis. For the papers provided, the replication files contain code to reproduce the analysis using Stata, an alternative statistical programming language. Your replication should be done using R and the packages described in class. These may differ from the Stata results due to minor changes in definitions. This may especially be the case for standard error calculations: when using the **sandwich** package to produce robust or clustered standard errors, multiple options are available for scaling (e.g., the estimated variance may be a sum multiplied by $\frac{1}{n}$ or $\frac{1}{n-1}$). You may report such differences rather than attempting to exactly duplicate the tables, but should also report the sources of any other discrepancies.

Your report should answer the following questions and perform the following analyses:

1. (a) Each paper is interested in measuring the causal effect of a variable or set of variables X on an outcome or set of outcomes Y. For example, for AFG, X is exposure to cholera due to an infected water source, Y is a measure of house prices. For BKDM, X is a measure of risk of expropriation, Y is a measure of Islamic institutions.

- Briefly describe the data set used to answer the question.
- What are the units of observation? (This may be different in different components of the analysis. If so, describe the units in each.)
- What measures do they use for the outcome variable and the predictor variable?
- If controls or instrumental variables are used, what measures do they use for these?
- Provide an opinion on whether the measures used provide adequate measures of the concepts of interest.

(b) Check that the summary statistics for the variables reported in the papers are correct.

- Replicate the tables of summary statistics (e.g. Table 1 in AFG, Table A1 in the appendix of BKDM) using the data sets provided, in R.

2. (a) Describe the main empirical model used to estimate the causal relationship between X and Y.

- What is the structure of the relationship between these variables that makes the model they use valid?
- Provide a structural equation and list the assumptions needed for the estimator they used to consistently estimate its parameters. For example, if the paper used panel data methods such as difference in differences, this should be an unobserved effects model.
- Describe the properties of the residuals, and sources of unobserved heterogeneity that the authors mention as reasons why the specification is used.
- If the paper used regression (with or without control, and possibly using panel data transformations such as fixed effects or differences) to estimate a causal effect, describe in terms of the structural equation the assumptions made regarding the relationship between observed and unobserved variables, and describe the author’s justification, if provided, for the conditional exogeneity condition.
- Draw a causal graph relating the variables used or described in the analysis under which the primary coefficient of interest in the regression equation estimated in the paper is interpretable as a causal effect, and which describes a set of relationships between variables which is plausible given the statements made in the paper regarding these relationships.

Hint: There may be many such plausible causal graphs. Draw *one* which seems to best agree with the statements in the paper. Each variable or set of variables (including fixed effects) should be described by a node. A variable at different times may be drawn as multiple nodes, one for each time, or, if the relationship between variables is the same at all times, one can draw the graph for a single time period, in the same way

reading straight through from start to finish.

one draws a graph which describes a relationship between variables which is the same for all units. When devising the graph, it may be helpful to consider whether the implications of the graph are supported by the data. Any time two variables do *not* have a link between them, this implies conditional independence of certain variables, which will imply a 0 coefficient in a multivariate regression. For example, if $A \perp B|C$, Then a (correctly specified) regression of A on B and C will have 0 coefficient on B. You can find the conditional independencies implied by your DAG by using command `impliedConditionalIndependencies` in library `dagitty`.

- If the paper used instrumental variables methods, describe potential sources of endogeneity of the regressors in this equation listed by the author and describe what assumptions are needed on the instrument for the procedure to provide a consistent estimate of the causal effect.

(b) Replicate the main empirical analysis in the papers. E.g., for AFG this is Table 3, for BKDM this is Table I.

- If there is a figure accompanying this main result (for example, a scatter plot corresponding to a regression line), replicate this figure along with the table.
- Briefly summarize what the authors find about the effect of X on Y

(c) If the method used has testable implications, perform a statistical evaluation of their validity and interpret the results. Depending on the empirical method used in the paper, one or more of the following analyses may be appropriate.

- It is likely that some or all of these checks have been performed in the paper itself: if so, replicate the result. Note that some results may be located in the supplementary appendices of the paper.
- Otherwise, perform and interpret the test yourself.
- If no such test is possible, report this and (briefly) explain why.

(i) For an experiment (or natural-experiment) where observable covariates not used in the main regression specification are available, perform a test of covariate balance; do observable covariates appear to be independent of treatment assignment? If so, a regression of the treatment on covariates should have 0 coefficients for all predictors except a constant.

(ii) For 2SLS estimates, perform a statistical evaluation of their validity.

- Report the first stage and reduced form, and perform tests of the IV conditions.
- Test the first stage relevance condition with an F test.
- Compute the OLS estimates (if not done already) and compare to the 2SLS estimates. Are they larger or smaller? Is this what would be predicted by the story that the authors present for the source of endogeneity? (You can use the omitted variables formula to predict the sign).
- Perform a Hausman test to see if endogeneity is an issue. If possible, also test the exclusion restrictions by an overidentification test. (*Hint*: all these tests can be performed by including **diagnostics=TRUE** as an option in **ivreg**.) Interpret the test results.

(iii) For regression discontinuity estimates, perform tests of

- Covariate balance above and below the discontinuity (if applicable): this can be run by ensuring that a regression discontinuity estimate with the covariate as the outcome variable has no significant difference above or below the discontinuity.
- Continuity of the density of the running variable. (E.g. via **dens_test** in library **rddtools** or **DCdensity** in library **rdd**, which implement the McCrary test of continuous density.)
- For RD estimates combined with another identification strategy, also do the checks associated with that strategy. For example, for *fuzzy* regression discontinuity estimates, perform the 2SLS robustness checks with the discontinuity as instrument. For *difference in discontinuity* estimates, which compare outcomes on either side of a discontinuity before and after a treatment is implemented on one side only, perform the difference-in-difference checks.

(iv) For an event study or difference-in-differences design with more than $T = 2$ time periods, or a regression discontinuity, perform a placebo check.

- Estimate the main regression specification with the change that the time period (or location of the running variable) after which “treatment” starts is changed to another time period, for each possible time period, and plot the estimated “treatment effects” against time (or the running variable). The time period corresponding to the true specification should stand out. Does this appear to be the case? If not, there may be time effects not accounted for by the estimator. (AFG, who have a running variable in space, refer to this as a “false boundaries” test.)

(v) For a difference-in-differences design with more than $T = 2$ time periods, compare, graphically, whether trends in the observed outcome in the treatment and control units appear to be parallel in the periods *before* treatment occurs. If differential trends are already included and more the $T = 3$ periods are included, estimate the differential trends on the pre-treatment data and subtract, then compare if the pre-treatment outcome paths are comparable after subtracted the estimated trends. (See the Angrist and Pischke textbook for a discussion of this analysis.) Although this check does not and cannot test for violation of the assumption of parallel trends in the *potential outcomes*, which may be true or false regardless of the outcome of this check, failure of parallel trends in *observed outcomes* pre-treatment is suggestive of differences between treatment and control units which may not be accounted for by using difference in differences.

(vi) For random effects or pooled OLS estimate on panel data, compare to a fixed effects estimate using a Hausman test to test for correlated persistent unobserved effects.

(vii) The above list of specification tests is not exhaustive; if a method other than the ones listed is described, or another test is possible, you may perform this test and report and interpret the results.

3. (a) Consider possible reasons why the main analysis might not provide a valid estimate of the causal effect.

- Are there important excluded variables or other sources of endogeneity beyond those accounted for by the main analysis?
- Are there reasons to believe any instruments used might not satisfy the exclusion restrictions?
- Are there functional forms which may be misspecified? (Especially consider trends and interaction terms, or cutoff values which may have been chosen arbitrarily.)
- For methods which rely on a bandwidth or similar tuning parameter (including local regression or HAC standard errors), how is this chosen, and might other choices lead to a better trade-off of bias and variance of the estimates?

The authors will provide alternative specifications to address these issues (often an extremely large set of them) as well as discussion of other sources of error.

Briefly list the alternative specifications which are tried, and remark if any of them lead to qualitatively different conclusions from the main analysis.

- Choose (at least) one of these additional specifications and replicate the results.

List also potential sources of error which are *mentioned* but not accounted for. Does any one strike you as likely to invalidate the main analysis, and if so why?

(b) Authors, even scrupulous ones, have a tendency to include only those robustness checks and potential concerns which do not contradict the main result.

Can you think of any other specifications which might potentially be worth considering as an empirical model? - Estimate at least one additional specification which is feasible to estimate using available data. - Describe whether the results of the main analysis change, and if so, why.

Examples of possible specifications to check might include nonlinear transformations or adding additional controls or interaction terms, using fixed effects instead of first differences or vice versa, changing the specification of a time trend, replacing a difference in difference estimate with a pure difference in mean or event study or vice versa, or using a nonparametric regression estimate in place of a parametric one.

(c) Are there any sources of endogeneity or misspecification that they don't mention which might change the conclusions?

- If so, what are they? Explain. Make reference to the causal graph you constructed before, considering particular assumptions embedded in that graph that might fail (like the absence of a certain arrow, or of an unobserved node with particular properties.)

4. Although many papers concentrate primarily on a single cause and outcome, many report multiple outcomes or attempt to estimate sources of heterogeneity in the treatment effect.

- Briefly list other outcomes beyond the primary outcome variable for which an effect was estimated, and variables for which an interaction effect with the the treatment was estimated.
- If such estimates were performed separately from the main specification, choose *one* and replicate it, and interpret the results.

Are there any additional concerns regarding this specification that affect the credibility of the estimate beyond those for the primary specification?

- Provide a brief explanation (no need to perform a full set of robustness checks, but describe any pertinent conditions may not be valid for these estimates).

5. Conclude with an overall assessment of whether the paper has provided a credible estimate of the causal effect.

- If so, describe (in equations and words) the model of the data which the results appear to support.
- If not, explain what assumptions of the model appear to (be likely to) be violated and why.

Assessment:

An ideal report will provide a complete, correct, and comprehensible response to the questions above, demonstrating an understanding of the properties of the procedures used and a comprehension of how they apply to the particular issue addressed by the study. The replication results need not be precisely the same as in the original studies (indeed, one use of replication is to verify that the procedures used in the study are accurate), but the results and code should demonstrate proper application and interpretation of the econometric methods used.

Notes Regarding Analyses in Particular Papers

The paper you choose, whether among the ones listed or not, is likely to include some methods not discussed in class, especially in the supplementary analyses. Notes on handling this for the listed papers are provided. If you chose another paper, please do ask for advice; I may be able to provide resources including references or software. You are not required for the project to replicate results based on these additional methods. However, I discuss them here so that you can have a sense of what role they are playing.

AFG: Due to the spatial nature of the data and the concern about correlation across space, which is analogous to correlation across time, many of the analyses use "Spatial Heteroskedasticity and Autocorrelation Consistent" (HAC) standard errors, referred to in the paper as Conley (1999) Standard Errors after the reference which introduced them. These are entirely analogous to time series HAC standard errors discussed in Chapter 12 of the Wooldridge textbook and which will be discussed in class, except that rather than just the distance in time, the correlation is over space and the bandwidth is chosen in two dimensions, "latitude" and "longitude". While there is an R package for these, it is not available on CRAN, so you should obtain it by downloading the folder from https://github.com/darinchristensen/conley-se and placing it in your project folder for the report. You can then load the command using *source()* and apply the command *ConleySEs()* with syntax as in their [help file]https://darinchristensen.com/post/conley-correction/ to get the standard errors.

Some other analyses use clustered standard errors by units which are not panel units. This can be achieved by using, in place of *lm()*, the command *felm()* in R package *lfe*, which provides an alternative to package **plm** for estimating panel data models with fixed effects. In particular, command **felm** estimates fixed effects regression and permits clustering standard errors with respect to arbitrary units, using a formula identical to that used in Stata. See the details of the help file for **felm** for syntax and references.

The particular method used to choose a bandwidth for regression discontinuity in this paper, due to Calonico, Cattaneo, and Titiunik (2014), is exactly the one implemented in R library **rdrobust** (as the authors provided both R and Stata commands for this choice), so this should be used to replicate the results. However, alternative bandwidth selection methods, like those implemented in other regression discontinuity libraries like **rdd** or **rddtools** can be tried if there is difficulty with this, and perhaps should be tried as part of the robustness checks. Some supplemental analyses which use local polynomial regression choose the order of the polynomial by a criterion known as "“Akaike’s criterion”, or AIC. Replicating these analyses is optional, but if you want to do it, instead of using *lm* to run the regression, you can use **glm** with *family = gaussian(link = “identity”)* so that the glm command runs OLS. The AIC value is then reported in the summary. The model with the *smallest* AIC should be chosen as best fitting the data.

BKDM The primary specification in this paper is a *difference-in-discontinuity* design, which is slightly different from those we have covered in class, though related in a straightforward way. Effectively, this method takes the difference over types (here high vs low expropriable land prevalence) in two regression discontinuity estimates instead of the difference in types of two difference in means estimates like a standard difference in differences. Here, the identifying assumption is parallel trends in the discontinuity in potential outcomes at the cutoff. This allows for the presence of discontinuity in potential outcomes at the cutoff (which would invalidate standard regression discontinuity) so long as the change over types in the potential outcome given no treatment at the cutoff is the same to the left and to the right. Alternately, if one believes that in the treatment type potential outcomes functions are continuous, this estimator corresponds to an interaction of the discontinuity with another variable, and so the coefficients represent heterogeneous treatment effects. Mechanically, this estimator can be calculated by running standard regression discontinuity estimators before and after and then subtracting, or by adding a type dummy and an interaction of the discontinuity with a type dummy to a regression, as with standard DiD.

For standard error calculations, most results in this paper use clustering, but some robustness checks use Conley spatial HAC standard errors: see above note on **felm**. Some also use the *bootstrap* method. The bootstrap (and particularly the “wild cluster bootstrap”) that they reference is a method for constructing standard error estimates for panel data under the same conditions in which clustered errors are valid. Some authors prefer this method because the approximation may be more precise than the explicit formula when sample sizes are small or moderate. The bootstrap procedure is implemented in R package **lfe**, which provides an alternative to package **plm** for estimating panel data models with fixed effects. In particular, command **felm** estimates fixed effects regression and permits clustering standard errors in this way. See the details of the help file for **felm** for syntax and references.

Regression discontinuity estimates here are done just by OLS in the main tables (though local linear regression with a kernel is used in the figures) with ad hoc bandwidth choices: you may also want to see the above note on regression discontinuity methods in R for alternatives.