Analysis Code Primer

Guidance and examples for conducting Geospatial Impact Evaluations (GIEs) using AidData's Stata analysis code repository.

Topics include:

- Difference-in-difference (DID) analysis
  - Single treatment time
  - Multiple treatment times
- Pre-trends analysis
- Event study analysis

Acronyms:

- DID - Difference-in-Difference
- GIE - Geospatial Impact Evaluations
- NDVI - Normalized Difference Vegetation Index
- SE - Standard Error

Section Overview

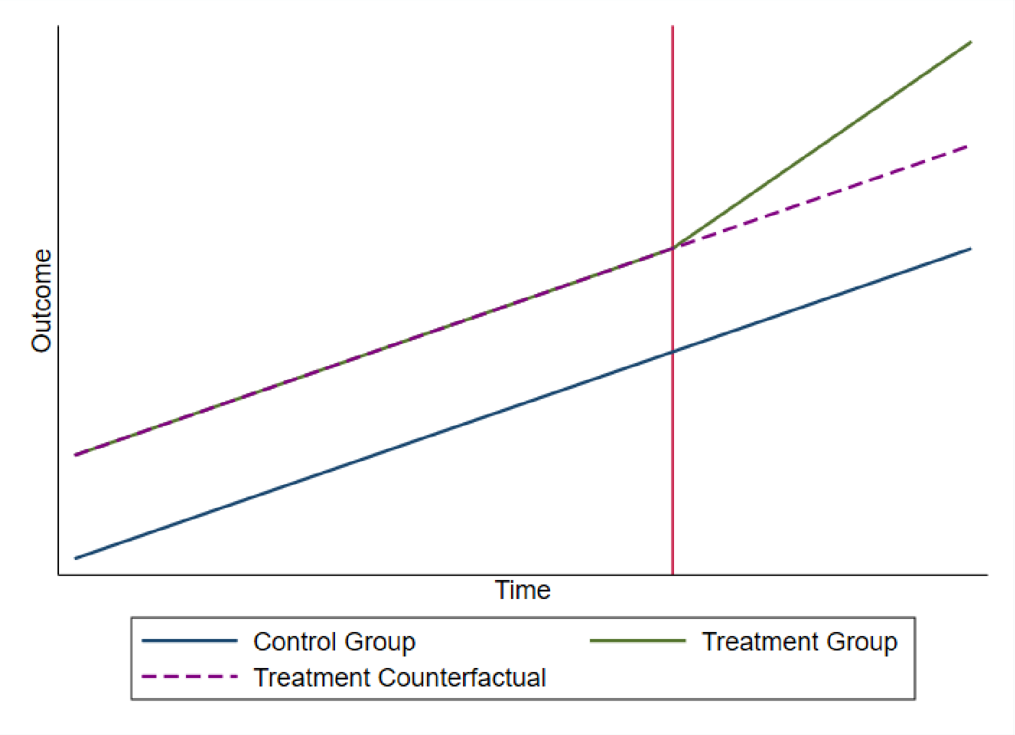
Difference-in-Difference (DID) Single Treatment Time

Difference-in-Difference (DID) is a quasi-experimental method of analysis that compares the changes in an outcome variable between the treatment and control group in order to identify the treatment effect of an intervention. DID can be used to estimate the effect of a treatment that is implemented at one time only (i.e. there is a single treatment time) or a treatment that has a staggered implementation (i.e. there are multiple treatment times). Here we discuss the use of DID to estimate the treatment effect of an intervention with a single treatment time.

Empirical Specification

DID allows for estimation of the treatment effect by utilizing the change in the outcome variable for the control group to create a counterfactual for the treatment group. That is, with DID, we assume that, in the absence of treatment, the change in the outcome variable for the treatment group would have followed the same trend as for the control group. As such, any difference between this counterfactual and the actual trajectory of the outcome variable for the treatment group can be attributed to the treatment.

Figure xx: Difference-in-Difference with a Single Treatment Time



In this way, DID utilizes two differences. The first difference is the difference in the post- and pre-treatment outcome for the treatment group. This difference controls for unobservables that are shared by members of the treatment group in both the post- and pre-treatment period. The second difference is the difference in the post- and pre-treatment outcome for the control group, controlling for unobservables that are shared by members of the control group in both the post-and pre-treatment period. DID is then calculated as the difference between these two differences. This leaves us with the treatment effect.

The empirical specification used to estimate the treatment effect for an intervention with a single treatment time using DID is the following:

$$Outcome_{itg} = \alpha + \beta_1 PostTreatment_t + \beta_2 TreatmentGroup_i + \beta_3 PostTreatment_t * TreatmentGroup_i + Controls_{itg} + \delta_g + \delta_t + \epsilon_{itg}$$

where $Outcome_{itg}$ is the outcome of interest for individual *i* at time *t* in geographical area *g.* $PostTreatment_t$ is a binary indicator that is 1 if time *t* is after the time of treatment and 0 otherwise. As such, β_1 is the average difference in the outcome variable between the pre- and post-treatment periods, whether individual *i* is in the treatment or control group. $TreatmentGroup_i$ is a binary indicator that is 1 if individual *i* is in the treatment group and 0 if it is in the control group. As such, β_2 is the average difference in the outcome variable between the treatment and control group, whether the individual is observed before or after treatment. $PostTreatment_t * TreatmentGroup_i$ is the interaction between these groups, and as such it is a binary indicator that is 1 if individual *i* is part of the treatment group and observed at time *t* after treatment occurs, and 0 otherwise. As such, β_3 is the coefficient of interest and represents the average treatment effect. $Controls_{itg}$ is an optional vector of control variables. δ_g is an optional geographic fixed effect that controls for time-invariant unobservables within geographic unit *g*. δ_t is an optional time fixed effect that controls for geographic-invariant unobservables within temporal unit *t.*

Identifying Assumptions

DID relies on the parallel trends assumption (sometimes called the equal tends assumption). This assumption is that the control and treatment group would have parallel time trends in the outcome variable in the absence of treatment. Note that the treatment and control group can have different levels of the outcome variable, as long as the overall time trend in the outcome is parallel.

This assumption is essential to producing an unbiased estimate of the treatment effect using DID. We will discuss methods to test for whether this assumption holds later in the document. However, it is important to consider the circumstances of the intervention prior to using a DID approach to understand whether this assumption holds.

Some common violations of the parallel trend assumption that would indicate that it is inappropriate to use DID for the intervention at hand include the following:

- Treatment is given to areas that are on a better trajectory with respect to the outcome variable (e.g. primary schools are targeted in areas where school enrollment had been increasing more rapidly prior to treatment)
- Treatment is given to areas that were on a worse trajectory with respect to the outcome variable (e.g. nutrition intervention targeted to areas where improvements in child health have stalled)

Data Requirements

The analysis code for the DID with a single treatment time can be used to estimate treatment effects of interventions that occurred at one time (i.e. not staggered). The data used for this analysis can be either cross-sectional or panel. The following are required of the data prior to using the analysis code:

- Data from multiple time periods (the time between the post-treatment time period and the time of treatment should be large enough for effects of the intervention to be detectable)
- Data from observations that received treatment and those who did not
- A binary variable that is 0 for individuals in the control group and 1 for individuals in the treatment group
- A binary variable that is 0 for observations observed before the time of treatment and 1 for those observed after
- A variable designated for clustering standard errors (SEs)

Analysis Code

This subsection describes how to utilize the DID single treatment time do-file to estimate the treatment effect of an intervention. It includes instructions to complete and run the do-file tailored to your dataset and intervention, as well as an overview of the output produced by the do-file.

Code Use

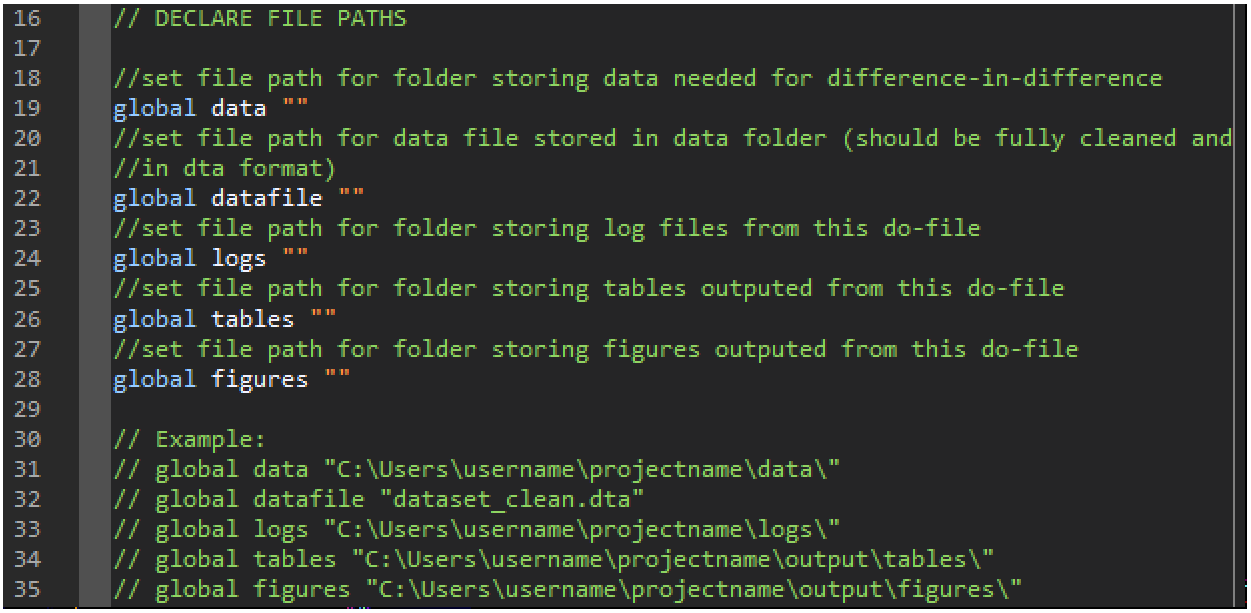
This subsection includes instructions on completing the do-file, with screenshots of the code. The code requires you to assign global macros with information relevant to your data and intervention at the start of the do-file. The do-file will then call these global macros to perform the analysis.

Declaring File Paths (Lines 16-35)

The first part of the do-file requires you to specify the file paths relevant to your data in global macros. All fields in this part of the do-file are required to make the do-file run.

- ☐ Line 19: specify the file path for the folder where your data file is stored in global *data* (Example: *global data "C:\Users\username\projectname\data\"*)
- ☐ Line 22: specify the name of your data file, which is located in your data folder. This file must be in dta format.
  (Example: *global datafile "dataset_clean.dta"*)

- □ Line 24: specify the file path for the folder where your log files from this do-file will be saved. This log file will record the commands and output of the do-file for you to refer back to at a later time
(Example: *global logs “C:\Users\username\projectname\logs\”*)
- □ Line 26: specify the file path for the folder where any tables produced in this do-file will be saved.
(Example: *global logs “C:\Users\username\projectname\tables\”*)
- □ Line 28: specify the file path for the folder where any figures, including graphs, produced in this do-file will be saved.
(Example: *global logs “C:\Users\username\projectname\figures\”*)



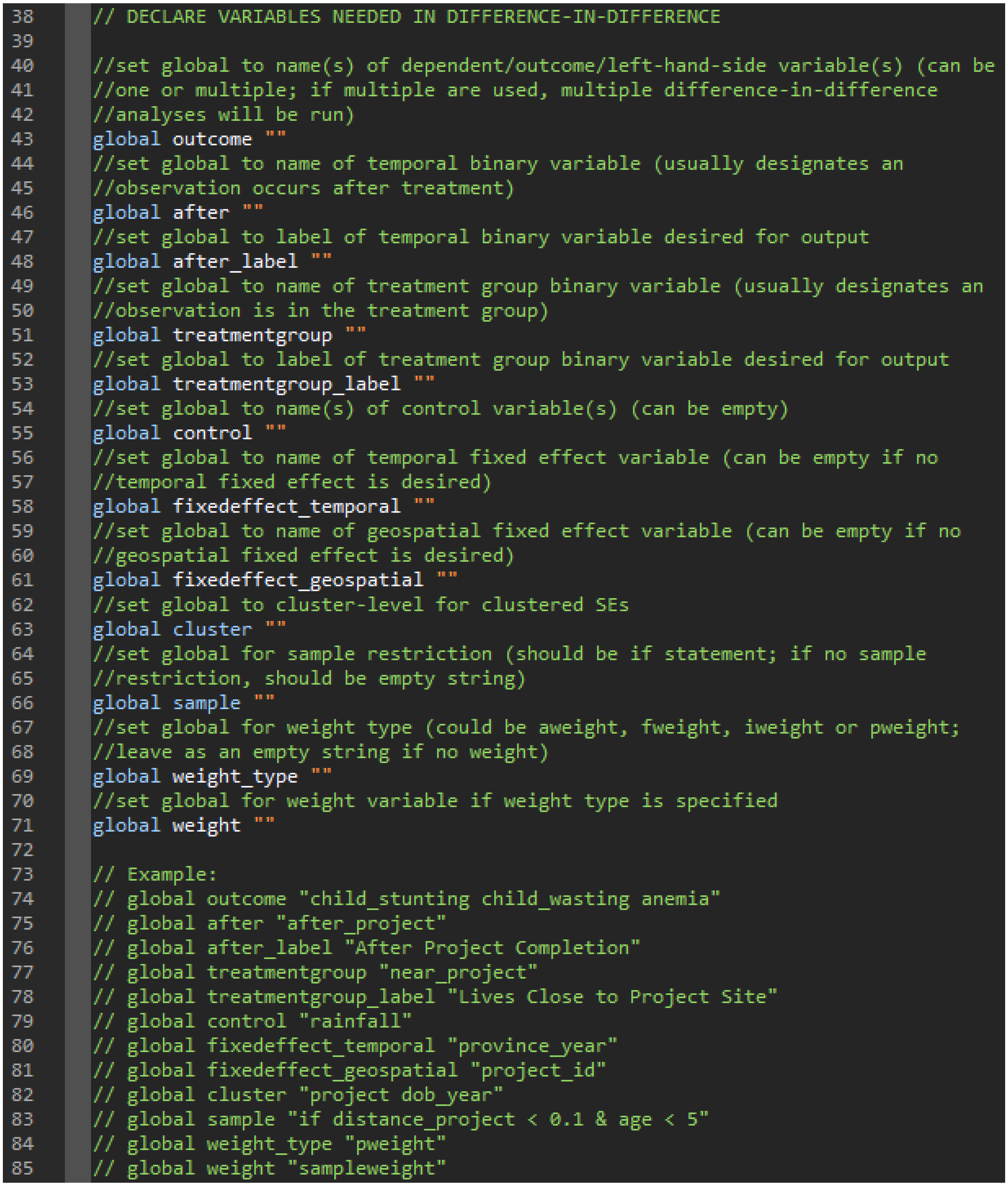
```
16  // DECLARE FILE PATHS
17
18  //set file path for folder storing data needed for difference-in-difference
19  global data ""
20  //set file path for data file stored in data folder (should be fully cleaned and
21  //in dta format)
22  global datafile ""
23  //set file path for folder storing log files from this do-file
24  global logs ""
25  //set file path for folder storing tables outputed from this do-file
26  global tables ""
27  //set file path for folder storing figures outputed from this do-file
28  global figures ""
29
30  // Example:
31  // global data "C:\Users\username\projectname\data\"
32  // global datafile "dataset_clean.dta"
33  // global logs "C:\Users\username\projectname\logs\"
34  // global tables "C:\Users\username\projectname\output\tables\"
35  // global figures "C:\Users\username\projectname\output\figures\"
```

Declaring Variables Needed in Analysis (Lines 38-85)

The second part of the do-file requires you to specify the variables relevant to your data in global macros. This section will also ask you for variable labels for key variables, in order to ensure any output in the form of tables or graphs is well-formatted. Some parts are optional, depending on your empirical specification; the optional parts will be marked accordingly. For the examples in this section, we imagine investigating the effect of an irrigation project on child stunting and child wasting. In this specification, we imagine that the treatment group is those who live near project sites and the control group is those who live further away from project sites. The post-treatment time is the time after project completion.

- □ Line 43: specify the name of the outcome variable(s) for the analysis. If there is only one outcome, only one variable name should be listed. If there are multiple, all variable names should be listed and separated by a space. The do-file will perform the analysis for each specified outcome separately. For concerns about multiple testing, indices should be constructed prior to using this do-file and the index variable name should be

specified in the outcome list. All outcome variables should be labeled for tables and graphs prior to using this do-file.
(Example: *global outcome “child_stunting child_wasting”)*



```
// DECLARE VARIABLES NEEDED IN DIFFERENCE-IN-DIFFERENCE

//set global to name(s) of dependent/outcome/left-hand-side variable(s) (can be
//one or multiple; if multiple are used, multiple difference-in-difference
//analyses will be run)
global outcome ""
//set global to name of temporal binary variable (usually designates an
//observation occurs after treatment)
global after ""
//set global to label of temporal binary variable desired for output
global after_label ""
//set global to name of treatment group binary variable (usually designates an
//observation is in the treatment group)
global treatmentgroup ""
//set global to label of treatment group binary variable desired for output
global treatmentgroup_label ""
//set global to name(s) of control variable(s) (can be empty)
global control ""
//set global to name of temporal fixed effect variable (can be empty if no
//temporal fixed effect is desired)
global fixedeffect_temporal ""
//set global to name of geospatial fixed effect variable (can be empty if no
//geospatial fixed effect is desired)
global fixedeffect_geospatial ""
//set global to cluster-level for clustered SEs
global cluster ""
//set global for sample restriction (should be if statement; if no sample
//restriction, should be empty string)
global sample ""
//set global for weight type (could be aweight, fweight, iweight or pweight;
//leave as an empty string if no weight)
global weight_type ""
//set global for weight variable if weight type is specified
global weight ""

// Example:
// global outcome "child_stunting child_wasting anemia"
// global after "after_project"
// global after_label "After Project Completion"
// global treatmentgroup "near_project"
// global treatmentgroup_label "Lives Close to Project Site"
// global control "rainfall"
// global fixedeffect_temporal "province_year"
// global fixedeffect_geospatial "project_id"
// global cluster "project dob_year"
// global sample "if distance_project < 0.1 & age < 5"
// global weight_type "pweight"
// global weight "sampleweight"
```

□ Line 46: specify the name of the binary indicator variable that is 1 if an observation is observed post-treatment and 0 if it is observed pre-treatment.
(Example: *global after “after_project”)*

□ Line 48: specify the label for the binary indicator variable specified in line 46. This label will be used in the outputted tables and graphs.
(Example: *global after_label “After Project Completion”*)

□ Line 51: specify the name of the binary indicator variable that is 1 if an observation is in the treatment group and 0 if it is in the control group.
(Example: *global treatmentgroup "near_project"*)

□ Line 53: specify the label for the binary indicator variable specified in line 51. This label will be used in the outputted tables and graphs.
(Example: *global treatmentgroup_label "Lives Close to Project Site"*)

□ Line 55 (optional): specify the name of the control variable(s) for the analysis. If there are no control variables, leave this field as empty quotation marks. If there is only one control variable, only one variable name should be listed, If there are multiple, all variable names should be listed and separated by a space. All control variables should be labeled for tables and graphs prior to using this do-file.
(Example: *global control "rainfall"*)

□ Line 58 (optional): specify the name of the temporal fixed effect variable for the analysis. This fixed effect will control for geographic-invariant unobservables within each temporal unit. If there is no temporal fixed effect variable, leave this field as empty quotation marks.
(Example: *global fixedeffect_temporal "province_year"*)

□ Line 61 (optional): specify the name of the geospatial fixed effect variable for the analysis. This fixed effect will control for time-invariant unobservables within each geospatial unit. If there is no geospatial fixed effect variable, leave this field as empty quotation marks.
(Example: *global fixedeffect_geospatial "project_id"*)

□ Line 63: specify the name of the variable(s) for which you want to cluster SEs by. If there is only one variable you want to cluster SEs by, only one variable name should be listed. If there are multiple, all variable names should be listed and separated by a space.
(Example: *cluster "project dob_year"*)

□ Line 66 (optional): specify the sample restriction for the analysis using an *if* statement. If there is no sample restriction, leave this field as empty quotation marks. If there are multiple sample restrictions, all restrictions should be included in one *if* statement using logic operators (i.e. & for and; | for or).
(Example: *global sample "if distance_project < 0.1 & age < 5"*)

□ Line 69 (optional): specify the type of weight that should be used for the analysis, if you wish to weight observations. You can specify *aweight, fweight, iweight,* or *pweight*. If you do not wish to use weights, leave this field as empty quotation marks.
(Example: *global weight_type "pweight"*)

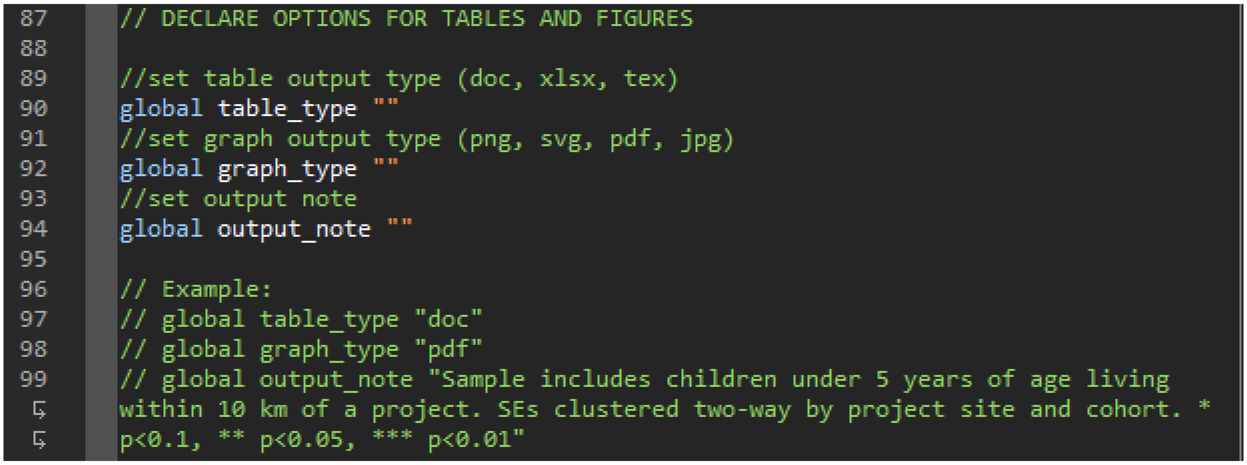
□ Line 71 (optional): specify the name of the variable that should be used for weighting according to Line 69. If you do not wish to use weights, leave this field as empty

quotation marks. If Line 69 is not specified, Line 71 should not be specified, and vice versa.
(Example: *global weight "sampleweight"*)

Declaring Options for Tables and Figures (Lines 87-99)

The third part of the do-file requires you to specify options for outputting the analysis results in the form of a table and a graph.

- □ Line 90: specify the output type/file extension desired for the regression table to be outputted. The regression table will be saved in this format in the table folder specified in Line 26. You can specify one of the following types: *doc, xlsx, tex.*
(Example: *global table_type "doc"*)
- □ Line 92: specify the output type/file extension desired for the coefficient plot generated from the regression to be outputted. The coefficient plot will be saved in this format in the figure folder specified in Line 28. You can specify one of the following types: *png, svg, pdf, jpg*.
(Example: *global graph_type "pdf"*)
- □ Line 94: specify a note to add to the outputted regression table. This note usually includes sample restrictions and information on clustering of SEs.
(Example: *global output_note "Sample includes children under 5 years of age living within 10 km of a project. SEs clustered two-way by project site and cohort. * p<0.1, ** p<0.05, *** p<0.01"*)

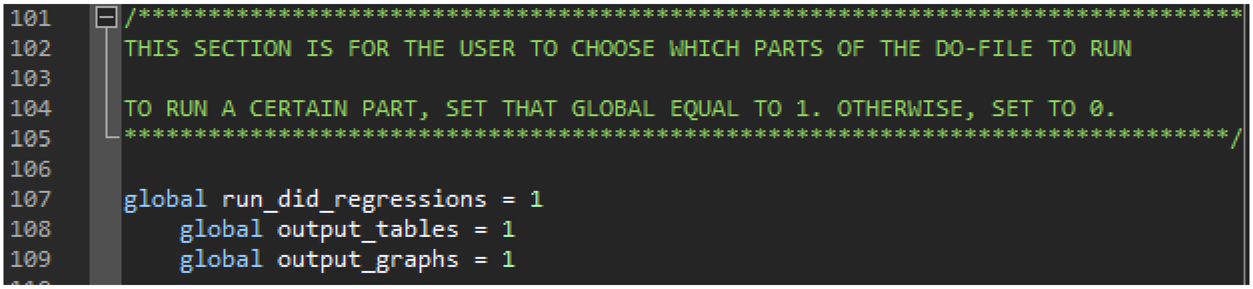


```
87  // DECLARE OPTIONS FOR TABLES AND FIGURES
88
89  //set table output type (doc, xlsx, tex)
90  global table_type ""
91  //set graph output type (png, svg, pdf, jpg)
92  global graph_type ""
93  //set output note
94  global output_note ""
95
96  // Example:
97  // global table_type "doc"
98  // global graph_type "pdf"
99  // global output_note "Sample includes children under 5 years of age living
 ↳  within 10 km of a project. SEs clustered two-way by project site and cohort. *
 ↳  p<0.1, ** p<0.05, *** p<0.01"
```

Specifying Parts of Do-File to Run (Lines 101-109)

The fourth part of the do-file requires you to specify which parts of the do file you would like to run. The purpose of this section is to allow you to select which parts of the do-file, particularly the output, that is generated each time you run it. This may help with debugging or when you are making changes to the specification. As a default, the file is set to run all parts.

- □ Line 107: specify if you would like to run the DID regressions. If you would like to run it, set it equal to 1; otherwise, set it to 0. Note that setting this to 0 will cause the majority of the do-file to not run.
- □ Line 108: specify if you would like to output the regression table from the DID analysis. If you would like to output the table, set it equal to 1; otherwise, set it to 0.
- □ Line 109: specify if you would like to output the coefficient plot from the DID analysis. If you would like to output the graph, set it equal to 1; otherwise, set it to 0.



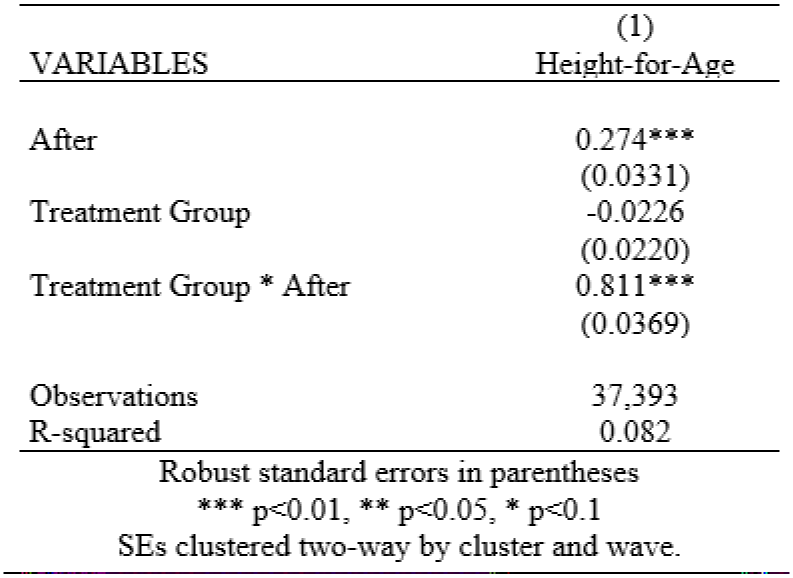
```
101  /*******************************************************************************
102  THIS SECTION IS FOR THE USER TO CHOOSE WHICH PARTS OF THE DO-FILE TO RUN
103  
104  TO RUN A CERTAIN PART, SET THAT GLOBAL EQUAL TO 1. OTHERWISE, SET TO 0.
105  *******************************************************************************/
106  
107  global run_did_regressions = 1
108      global output_tables = 1
109      global output_graphs = 1
```

Code Output

The code outputs one type of regression table and one type of coefficient plot showing the result of the DID analysis (the number of each type outputted depends on the number of outcomes specified).

The first output from the do-file is a regression table pre-specified outcome variable(s) that shows the DID results. The column title will be the outcome variable's label and the row labels will be the label for the post-treatment binary variable, the treatment group binary variable, and the interaction of these two variables (the coefficient on this variable is the estimated treatment effect), as well as labels for the control variables. Both the coefficient and SE are reported in the table, with the latter in parentheses. An example of this table is shown in Table XX.

Table XX: Example of DID Single Treatment Times Regression Table

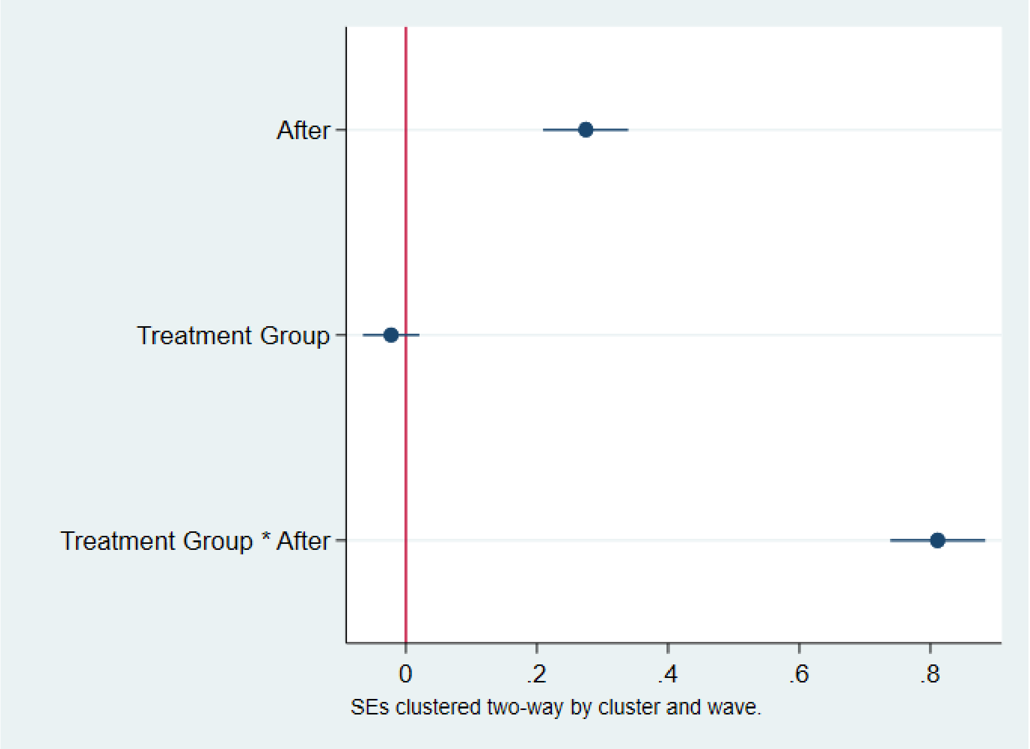


VARIABLES	(1) Height-for-Age
After	0.274***
	(0.0331)
Treatment Group	-0.0226
	(0.0220)
Treatment Group * After	0.811***
	(0.0369)
Observations	37,393
R-squared	0.082

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
SEs clustered two-way by cluster and wave.

The second output from the do-file is a coefficient plot pre-specified outcome variable(s) that shows the DID results. The x-axis shows the value of the coefficient estimates, and the y-axis lists the variable labels (including post-treatment, treatment group, and the interaction of these variables, along with control variables). Both the coefficient estimate (point) and confidence interval (line) are shown in the graph. An example of this table is shown in Table XX.

Figure XX: Example of DID SingleTreatment Times Coefficient Plot



Pre-Trends Analysis

Since Difference-in-Difference (DID) only produces an unbiased estimate of the treatment effect of an intervention if the parallel trends assumption holds, it is vital to investigate this assumption. Of course, researchers cannot prove the parallel trends assumption holds, as they cannot know how the outcome would have progressed post-treatment in the treatment group in the absence of treatment (i.e. the true counterfactual). However, researchers can investigate the time trends in the outcome variable pre-treatment to determine whether the pre-trends of the treatment and control group were parallel. If this is the case, it is assumed that the parallel trends assumption would likely have continued to hold in the absence of treatment. Here we discuss pre-trends analysis to use in conjunction with DID for an intervention with a single treatment time, in order to investigate the parallel trends assumption.

Empirical Specification

This do-file conducts a pre-trend analysis using two techniques. The first technique is to consider the raw pre-trends in the outcome variable(s). In this specification, the outcome variable is averaged across each time period for both the treatment and control groups separately. These averages are then plotted across time in a line graph to allow for visual inspection of the raw pre-trends.

The second technique is to consider the pre-trends when accounting for control variables and fixed effects. This allows for visual inspection of the pre-trends once accounting for confounding variables. The empirical specification of this analysis is as follows:

$$Outcome_{itg} = \alpha + \beta\delta_t + Controls_{itg} + \delta_g + \epsilon_{itg}$$

where $Outcome_{itg}$ is the outcome of interest for individual *i* at time *t* in geographical area *g.* δ_t are indicator variables for each period of time, with the base referenced to the time of treatment. As such, β is a vector of coefficients showing the average of $Outcome_{itg}$ in each time period. $Controls_{itg}$ is an optional vector of control variables. δ_g is an optional geographic fixed effect that controls for time-invariant unobservables within geographic unit *g*.

Data Requirements

The analysis code for the pre-trends analysis can be used to visualize the time trends of outcome variables prior to a treatment that occurred at one time (i.e. not staggered). The data used for this analysis can be either cross-sectional or panel. The following are required of the data prior to using the analysis code:

- Data from multiple pre-treatment time periods (at least two pre-treatment time periods must be present to visualize any pre-treatment time trend)
- Data from observations that received treatment and those who did not
- A binary variable that is 0 for individuals in the control group and 1 for individuals in the treatment group

- A binary variable that is 0 for observations observed before the time of treatment and 1 for those observed after
- A variable designated for clustering standard errors (SEs)

Analysis Code

This subsection describes how to utilize the pre-trends analysis do-file to look at the pre-trends in outcome variables prior to treatment. This file is useful to understand whether DID can be used, as it allows you to consider the parallel trends assumption. This section includes instructions to complete and run the do-file tailored to your dataset and intervention, as well as an overview of the output produced by the do-file.

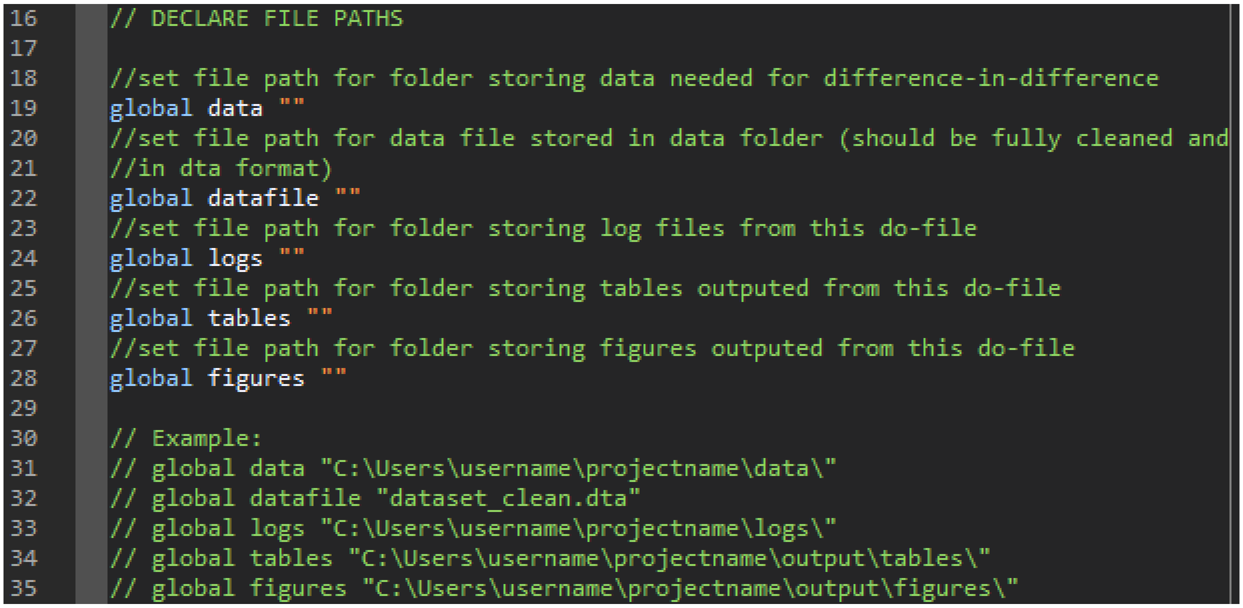
Code Use

This subsection includes instructions on completing the do-file, with screenshots of the code. The code requires you to assign global macros with information relevant to your data and intervention at the start of the do-file. The do-file will then call these global macros to perform the analysis. Note that this file is to be used with the DID single treatment time do-file.

Declaring File Paths (Lines 16-35)

The first part of the do-file requires you to specify the file paths relevant to your data in global macros. All fields in this part of the do-file are required to make the do-file run.

- □ Line 19: specify the file path for the folder where your data file is stored in global *data* (Example: *global data "C:\Users\username\projectname\data\"*)
- □ Line 22: specify the name of your data file, which is located in your data folder. This file must be in dta format.
  (Example: *global datafile "dataset_clean.dta"*)
- □ Line 24: specify the file path for the folder where your log files from this do-file will be saved. This log file will record the commands and output of the do-file for you to refer back to at a later time
  (Example: *global logs "C:\Users\username\projectname\logs\"*)
- □ Line 26: specify the file path for the folder where any tables produced in this do-file will be saved.
  (Example: *global logs "C:\Users\username\projectname\tables\"*)
- □ Line 28: specify the file path for the folder where any figures, including graphs, produced in this do-file will be saved.
  (Example: *global logs "C:\Users\username\projectname\figures\"*)

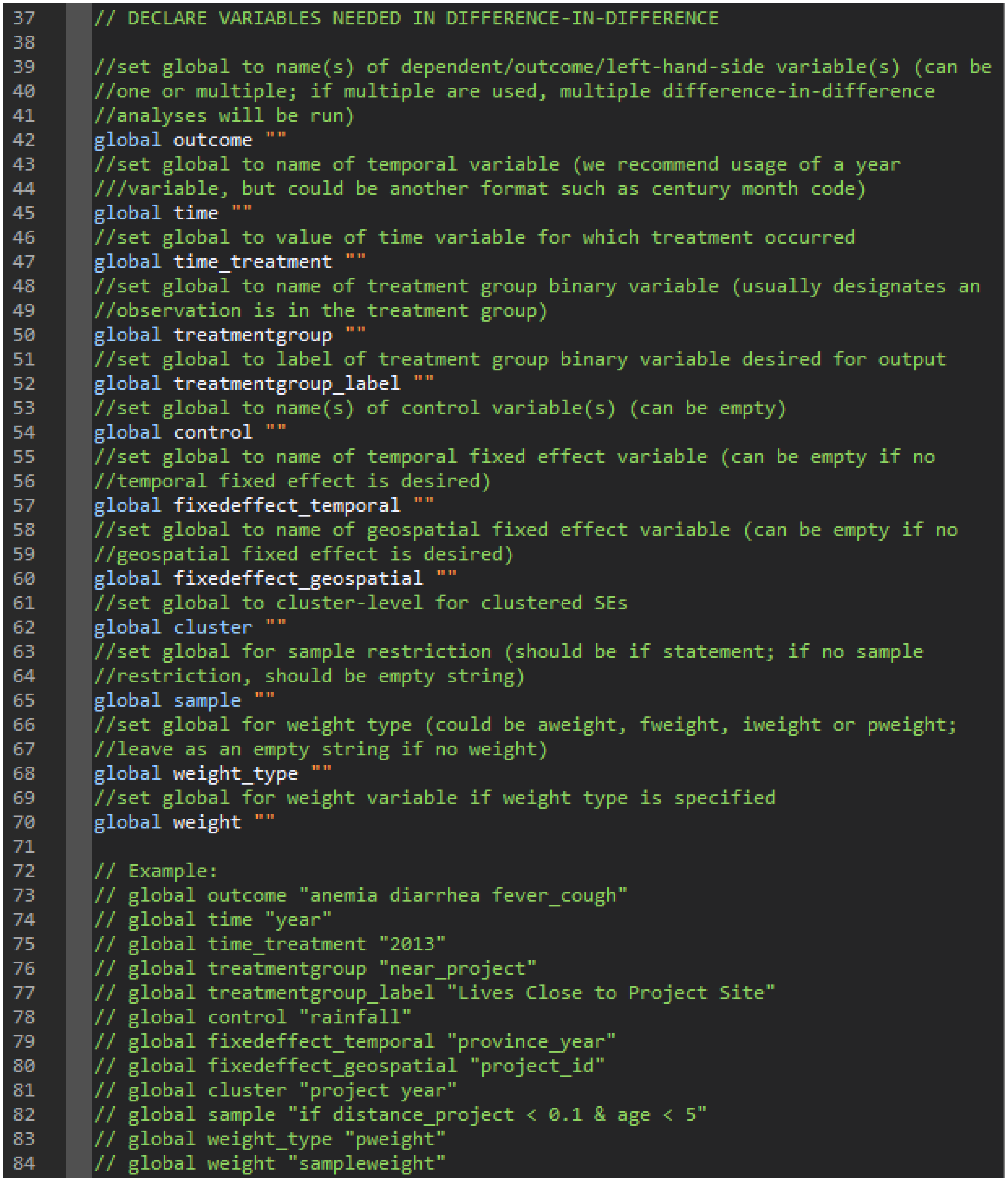


```
16  // DECLARE FILE PATHS
17
18  //set file path for folder storing data needed for difference-in-difference
19  global data ""
20  //set file path for data file stored in data folder (should be fully cleaned and
21  //in dta format)
22  global datafile ""
23  //set file path for folder storing log files from this do-file
24  global logs ""
25  //set file path for folder storing tables outputed from this do-file
26  global tables ""
27  //set file path for folder storing figures outputed from this do-file
28  global figures ""
29
30  // Example:
31  // global data "C:\Users\username\projectname\data\"
32  // global datafile "dataset_clean.dta"
33  // global logs "C:\Users\username\projectname\logs\"
34  // global tables "C:\Users\username\projectname\output\tables\"
35  // global figures "C:\Users\username\projectname\output\figures\"
```

Declaring Variables Needed in Analysis (Lines 37-83)

The second part of the do-file requires you to specify the variables relevant to your data in global macros. This section will also ask you for variable labels for key variables, in order to ensure any output in the form of graphs is well-formatted. Some parts are optional, depending on your empirical specification; the optional parts will be marked accordingly. For the examples in this section, we imagine investigating the effect of an irrigation project on child illness, including anemia, diarrhea, and fever and cough. In this specification, we imagine that the treatment group is those who live near project sites and the control group is those who live further away from project sites. The post-treatment time is the time after project completion, which occurred in 2013 in this example.

- ☐ Line 42: specify the name of the outcome variable(s) for the analysis. If there is only one outcome, only one variable name should be listed. If there are multiple, all variable names should be listed and separated by a space. The do-file will perform the analysis for each specified outcome separately. For concerns about multiple testing, indices should be constructed prior to using this do-file and the index variable name should be specified in the outcome list. All outcome variables should be labeled for tables and graphs prior to using this do-file.
  (Example: *global outcome "anemia diarrhea fever_cough")*
- ☐ Line 45: specify the name of the temporal variable used in the analysis. This variable must be in integer or double format, not string. We recommend usage of a year variable, but the variable could be in an alternative format, such as century month code.
  (Example: *global time "year")*



```
37  // DECLARE VARIABLES NEEDED IN DIFFERENCE-IN-DIFFERENCE
38
39  //set global to name(s) of dependent/outcome/left-hand-side variable(s) (can be
40  //one or multiple; if multiple are used, multiple difference-in-difference
41  //analyses will be run)
42  global outcome ""
43  //set global to name of temporal variable (we recommend usage of a year
44  ///variable, but could be another format such as century month code)
45  global time ""
46  //set global to value of time variable for which treatment occurred
47  global time_treatment ""
48  //set global to name of treatment group binary variable (usually designates an
49  //observation is in the treatment group)
50  global treatmentgroup ""
51  //set global to label of treatment group binary variable desired for output
52  global treatmentgroup_label ""
53  //set global to name(s) of control variable(s) (can be empty)
54  global control ""
55  //set global to name of temporal fixed effect variable (can be empty if no
56  //temporal fixed effect is desired)
57  global fixedeffect_temporal ""
58  //set global to name of geospatial fixed effect variable (can be empty if no
59  //geospatial fixed effect is desired)
60  global fixedeffect_geospatial ""
61  //set global to cluster-level for clustered SEs
62  global cluster ""
63  //set global for sample restriction (should be if statement; if no sample
64  //restriction, should be empty string)
65  global sample ""
66  //set global for weight type (could be aweight, fweight, iweight or pweight;
67  //leave as an empty string if no weight)
68  global weight_type ""
69  //set global for weight variable if weight type is specified
70  global weight ""
71
72  // Example:
73  // global outcome "anemia diarrhea fever_cough"
74  // global time "year"
75  // global time_treatment "2013"
76  // global treatmentgroup "near_project"
77  // global treatmentgroup_label "Lives Close to Project Site"
78  // global control "rainfall"
79  // global fixedeffect_temporal "province_year"
80  // global fixedeffect_geospatial "project_id"
81  // global cluster "project year"
82  // global sample "if distance_project < 0.1 & age < 5"
83  // global weight_type "pweight"
84  // global weight "sampleweight"
```

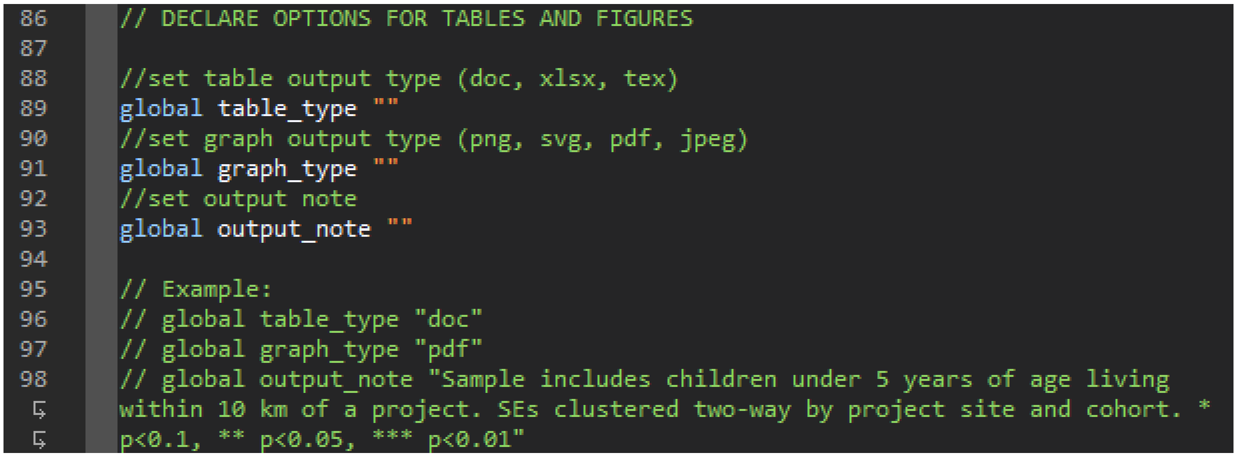
☐ Line 47: specify the value of the time variable specified in Line 45 that corresponds to the time of treatment. Remember, this file is for use with an intervention that has only one time of treatment, so there should only be one value in this field. (Example: *global time_treatment "2013")*

- ☐ Line 50: specify the name of the binary indicator variable that is 1 if an observation is in the treatment group and 0 if it is in the control group.
(Example: *global treatmentgroup "near_project"*)
- ☐ Line 52: specify the label for the binary indicator variable specified in line 50. This label will be used in the outputted tables and graphs.
(Example: *global treatmentgroup_label "Lives Close to Project Site"*)
- ☐ Line 54 (optional): specify the name of the control variable(s) for the analysis. If there are no control variables, leave this field as empty quotation marks. If there is only one control variable, only one variable name should be listed, If there are multiple, all variable names should be listed and separated by a space. All control variables should be labeled for tables and graphs prior to using this do-file.
(Example: *global control "rainfall"*)
- ☐ Line 57 (optional): specify the name of the temporal fixed effect variable for the analysis. This fixed effect will control for geographic-invariant unobservables within each temporal unit. If there is no temporal fixed effect variable, leave this field as empty quotation marks.
(Example: *global fixedeffect_temporal "province_year"*)
- ☐ Line 60 (optional): specify the name of the geospatial fixed effect variable for the analysis. This fixed effect will control for time-invariant unobservables within each geospatial unit. If there is no geospatial fixed effect variable, leave this field as empty quotation marks.
(Example: *global fixedeffect_geospatial "project_id"*)
- ☐ Line 62: specify the name of the variable(s) for which you want to cluster SEs by. If there is only one variable you want to cluster SEs by, only one variable name should be listed. If there are multiple, all variable names should be listed and separated by a space.
(Example: *cluster "project year"*)
- ☐ Line 65 (optional): specify the sample restriction for the analysis using an *if* statement. If there is no sample restriction, leave this field as empty quotation marks. If there are multiple sample restrictions, all restrictions should be included in one *if* statement using logic operators (i.e. & for and; | for or).
(Example: *global sample "if distance_project < 0.1 & age < 5"*)
- ☐ Line 68 (optional): specify the type of weight that should be used for the analysis, if you wish to weight observations. You can specify *aweight, fweight, iweight,* or *pweight*. If you do not wish to use weights, leave this field as empty quotation marks.
(Example: *global weight_type "pweight"*)
- ☐ Line 70 (optional): specify the name of the variable that should be used for weighting according to Line 68. If you do not wish to use weights, leave this field as empty quotation marks. If Line 68 is not specified, Line 70 should not be specified, and vice versa.
(Example: *global weight "sampleweight"*)

Declaring Options for Tables and Figures (lines 86-98)

The third part of the do-file requires you to specify options for outputting the analysis results in the form of a table and a graph.

- □ Line 89: specify the output type/file extension desired for the regression table to be outputted. The regression table will be saved in this format in the table folder specified in Line 26. You can specify one of the following types: *doc, xlsx, tex.*
  (Example: *global table_type "doc"*)
- □ Line 91: specify the output type/file extension desired for the coefficient plot generated from the regression to be outputted. The coefficient plot will be saved in this format in the figure folder specified in Line 28. You can specify one of the following types: *png, svg, pdf, jpg*.
  (Example: *global graph_type "pdf"*)
- □ Line 93: specify a note to add to the outputted regression table. This note usually includes sample restrictions and information on clustering of SEs. It can also include other information relevant to interpreting the analysis results.
  (Example: *global output_note "Sample includes children under 5 years of age living within 10 km of a project. SEs clustered two-way by project site and cohort. * p<0.1, ** p<0.05, *** p<0.01"*)

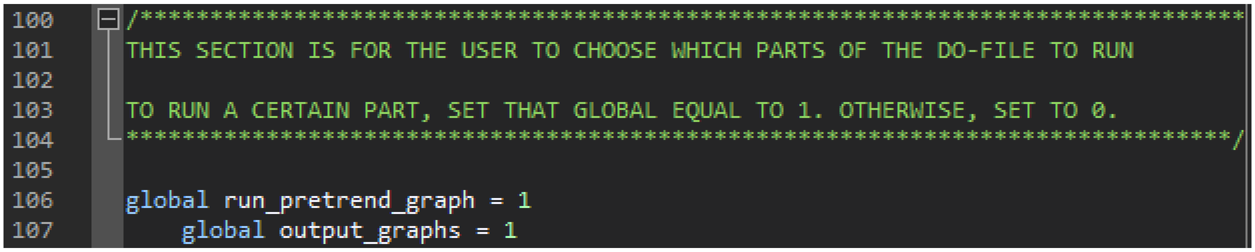


```
86  // DECLARE OPTIONS FOR TABLES AND FIGURES
87
88  //set table output type (doc, xlsx, tex)
89  global table_type ""
90  //set graph output type (png, svg, pdf, jpeg)
91  global graph_type ""
92  //set output note
93  global output_note ""
94
95  // Example:
96  // global table_type "doc"
97  // global graph_type "pdf"
98  // global output_note "Sample includes children under 5 years of age living
 ↳  within 10 km of a project. SEs clustered two-way by project site and cohort. *
 ↳  p<0.1, ** p<0.05, *** p<0.01"
```

Specifying Parts of Do-File to Run (Lines 100-107)

The fourth part of the do-file requires you to specify which parts of the do file you would like to run. The purpose of this section is to allow you to select which parts of the do-file, particularly the output, that is generated each time you run it. This may help with debugging or when you are making changes to the specification. As a default, the file is set to run all parts.

- □ Line 106: specify if you would like to run the pre-trend analysis. If you would like to run it, set it equal to 1; otherwise, set it to 0. Note that setting this to 0 will cause the majority of the do-file to not run.
- □ Line 107: specify if you would like to output the pre-trends graph from the pre-trend analysis. If you would like to output the graphs, set it equal to 1; otherwise, set it to 0.



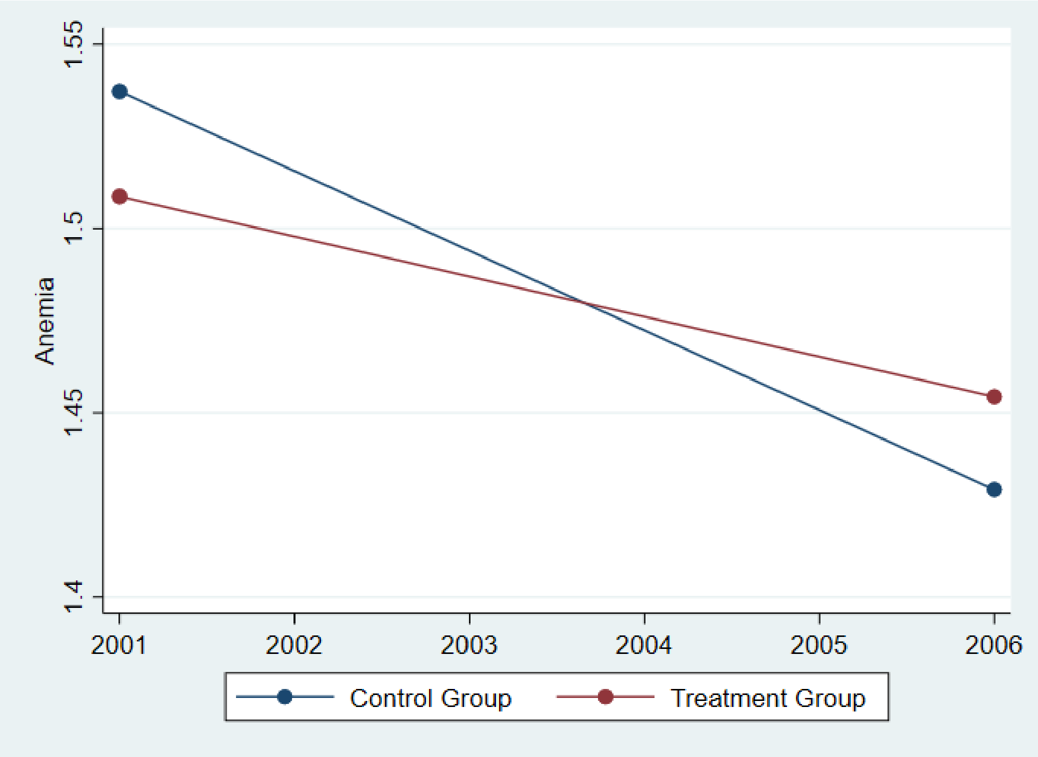
```
/*****************************************************************************
THIS SECTION IS FOR THE USER TO CHOOSE WHICH PARTS OF THE DO-FILE TO RUN

TO RUN A CERTAIN PART, SET THAT GLOBAL EQUAL TO 1. OTHERWISE, SET TO 0.
*****************************************************************************/

global run_pretrend_graph = 1
    global output_graphs = 1
```

Code Output

The code outputs two types of graphs showing the result of the pre-trends analysis (the number of each type outputted depends on the number of outcomes specified). The first type of graph shows the raw pre-treatment time trends in the outcome variable(s) for both the treatment and control group. The second type of graph shows the pre-treatment time trends in the outcome variable(s) for both the treatment and control group after controlling for specified control variables and fixed effects.

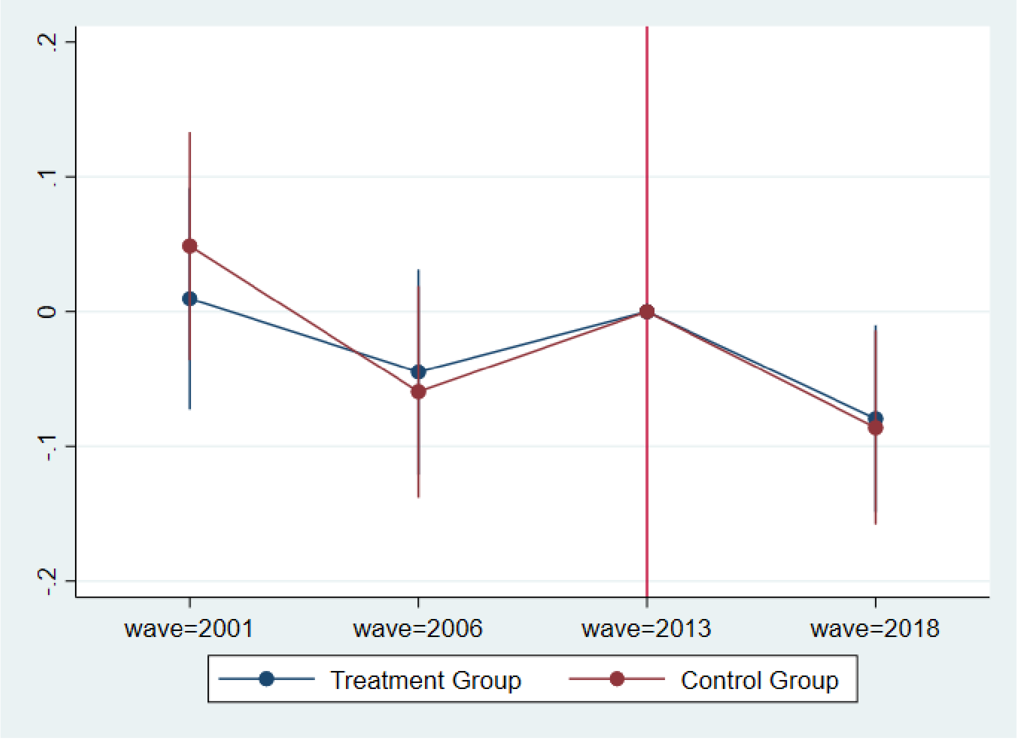
The first output from the do-file is a graph of raw pre-treatment trends in each of the pre-specified outcome variable(s). The x-axis represents time and the y-axis represents the mean of the relevant outcome variable. An example of this table is shown in Table XX.

Table XX: Example of Raw Pre-Trends Graph



The second output from the do-file is the pre-trends in the pre-specified outcome variable(s) after controlling for controls and fixed effects. The x-axis represents time and the y-axis represents the estimated mean of the relevant outcome variable after controlling for the specified controls and fixed effects. Both the coefficient estimate (point) and confidence interval (line) are shown in the graph. An example of this table is shown in Table XX.

Table XX: Example of Pre-Trends Graph



Difference-in-Difference (DID) Multiple Treatment Times

Difference-in-Difference (DID) is a quasi-experimental method of analysis that compares the changes in an outcome variable between the treatment and control group in order to identify the treatment effect of an intervention. DID can be used to estimate the effect of a treatment that is implemented at one time only (i.e. there is a single treatment time) or a treatment that has a staggered implementation (i.e. there are multiple treatment times). Here we discuss the use of DID to estimate the treatment effect of an intervention with multiple treatment times.

Empirical Specification

DID analysis of an intervention with multiple treatment times follows the same core concepts of a DID analysis of an intervention with a single treatment time. That is, DID allows for the estimation of the treatment effect by utilizing the change in the outcome variable for the control group to create a counterfactual for the treatment group. However, when the intervention has multiple treatment times or a staggered rollout, the treatment time must be statistically aligned to ensure a comparison of pre- and post-treatment outcomes is not conflated with general time trends.

As such, DID for an intervention with staggered rollouts incorporates two-way fixed effects to statistically align staggered treatment times. Similarly to DID for interventions with single treatment times, DID utilizes two differences (the difference in the post- and pre-treatment outcome for the treatment group and the difference in the post- and pre-treatment outcome for the control group). However, for a DID for an intervention with staggered rollouts, two-way fixed effects are used to de-mean the averages before calculating these differences and then calculating the treatment effect. These fixed effects include a temporal fixed effect and a geospatial fixed effect; the geospatial fixed effect should be at the same level as the unit (e.g. if the unit of observation is at the grid-cell*year level, the geospatial fixed effect should be at the grid-cell level).

In this way, a DID for an intervention with staggered rollouts sets the treatment group as observations that have already received treatment, and the control group as observations that have not yet received treatment. In this way, individuals/units serve as both the treatment and control group, depending on when they are observed relative to treatment. This specification therefore requires a panel dataset of individuals/units.

The empirical specification used to estimate the treatment effect for an intervention with multiple treatment times using DID is the following:

$$Outcome_{it} = \alpha + \beta Treated_{it} + Controls_{it} + \delta_i + \delta_t + \epsilon_{itg}$$

where $Outcome_{it}$ is the outcome of interest for individual/unit *i* at time $t.$ $Treated_{it}$ is a binary indicator that is 1 if individual *i* is part of the treatment group and is observed at time *t* after treatment occurs, and 0 otherwise. As such, is the coefficient of interest and represents the average treatment effect. $Controls_{it}$ is an optional vector of control variables. δ_i is a required geospatial/unit fixed effect that controls for time-invariant unobservables within individual/unit *i*. δ_t is a required time fixed effect that controls for geographic-invariant unobservables within temporal unit *t.*

Identifying Assumptions

Like a DID of an intervention with a single treatment time, DID of an intervention with multiple treatment times relies on the parallel trends assumption, which assumes that the control and treatment group would have parallel time trends in the outcome variable in the absence of treatment. This assumption is essential to producing an unbiased estimate of the treatment effect using DID. We will discuss methods to test for whether this assumption holds later in the document.

Since observations serve as members of both the treatment and control group, depending on when they are observed relative to treatment, there is an additional assumption underlying this parallel trends assumption. That assumption is that the time of treatment was assigned either randomly or as-if random. Importantly, this means that time of treatment cannot be prioritized based on the outcome variable or another characteristic correlated with the outcome variable.

In addition, the effects of treatment must not be heterogeneous across time (i.e. individuals/units benefitted the same way from treatment, no matter whether they were early or late recipients) in order for the estimated treatment effect to reflect the actual treatment effect on all treated individuals. Similarly, the treatment must be homogenous across time.

Some common violations of these assumptions that would indicate that it is inappropriate to use DID for the intervention at hand include the following:

- Treatment is prioritized among those individuals/units with worse outcomes, even though all eventually receive it (e.g. nutrition intervention targeted first to areas with the worst child health outcomes at baseline and later to those with better child health outcomes at baseline)
- Treatment is prioritized among those individuals/units with better outcomes, even though all eventually receive it (e.g. primary school construction is targeted first to areas with higher primary school completion at baseline and later to those with lower primary school completion at baseline)
- Effects of treatment are smaller among those individuals/units who receive it later (e.g. a nutritional information intervention has information spillovers, such that the new information in the intervention is less novel to communities receiving it later)
- Effects of treatment are larger among those individuals/units who receive it later (e.g. a road building intervention improves market access more for individuals/units who receive it later, because it can harness the other roads built during earlier treatment times)
- The treatment changes/evolves over time (e.g. implementers learn over time and change the implementation of the intervention in response)

Data Requirements

The analysis code for the DID with multiple treatment times can be used to estimate treatment effects of interventions that occurred at more than one time (i.e. staggered). The data used for this analysis must be panel data. The following are required of the data prior to using the analysis code:

- Data from multiple time periods for each observation (the time between the post-treatment time period and the time of treatment should be large enough for effects of the intervention to be detectable)
- A binary variable that is 1 for individuals in the treatment group that were observed after being treated and 0 otherwise
- A variable designated for clustering standard errors (SEs)

Analysis Code

This subsection describes how to utilize the DID multiple treatment times do-file to estimate the treatment effect of an intervention. It includes instructions to complete and run the do-file tailored to your dataset and intervention, as well as an overview of the output produced by the do-file.

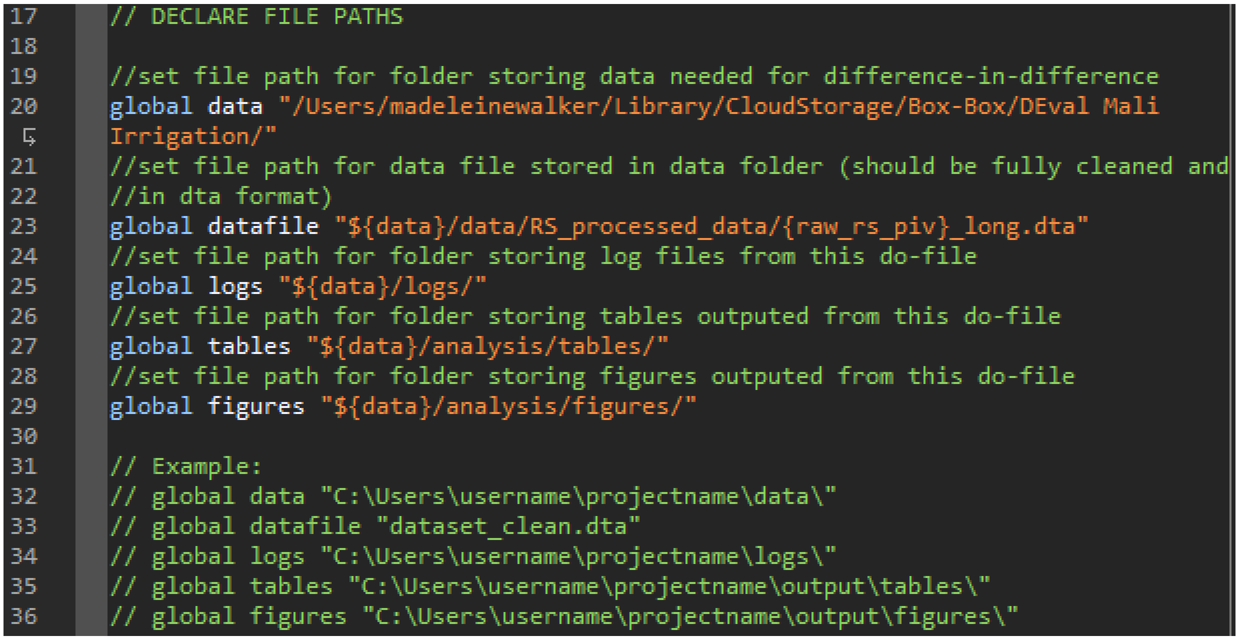
Code Use

This subsection includes instructions on completing the do-file, with screenshots of the code. The code requires you to assign global macros with information relevant to your data and intervention at the start of the do-file. The do-file will then call these global macros to perform the analysis.

Declaring File Paths (Lines 17-36)

The first part of the do-file requires you to specify the file paths relevant to your data in global macros. All fields in this part of the do-file are required to make the do-file run.

- □ Line 20: specify the file path for the folder where your data file is stored in global *data*
(Example: *global data “C:\Users\username\projectname\data\”*)
- □ Line 23: specify the name of your data file, which is located in your data folder. This file must be in dta format.
(Example: *global datafile “dataset_clean.dta”*)
- □ Line 25: specify the file path for the folder where your log files from this do-file will be saved. This log file will record the commands and output of the do-file for you to refer back to at a later time
(Example: *global logs “C:\Users\username\projectname\logs\”*)
- □ Line 27: specify the file path for the folder where any tables produced in this do-file will be saved.
(Example: *global logs “C:\Users\username\projectname\tables\”*)
- □ Line 29: specify the file path for the folder where any figures, including graphs, produced in this do-file will be saved.
(Example: *global logs “C:\Users\username\projectname\figures\”*)

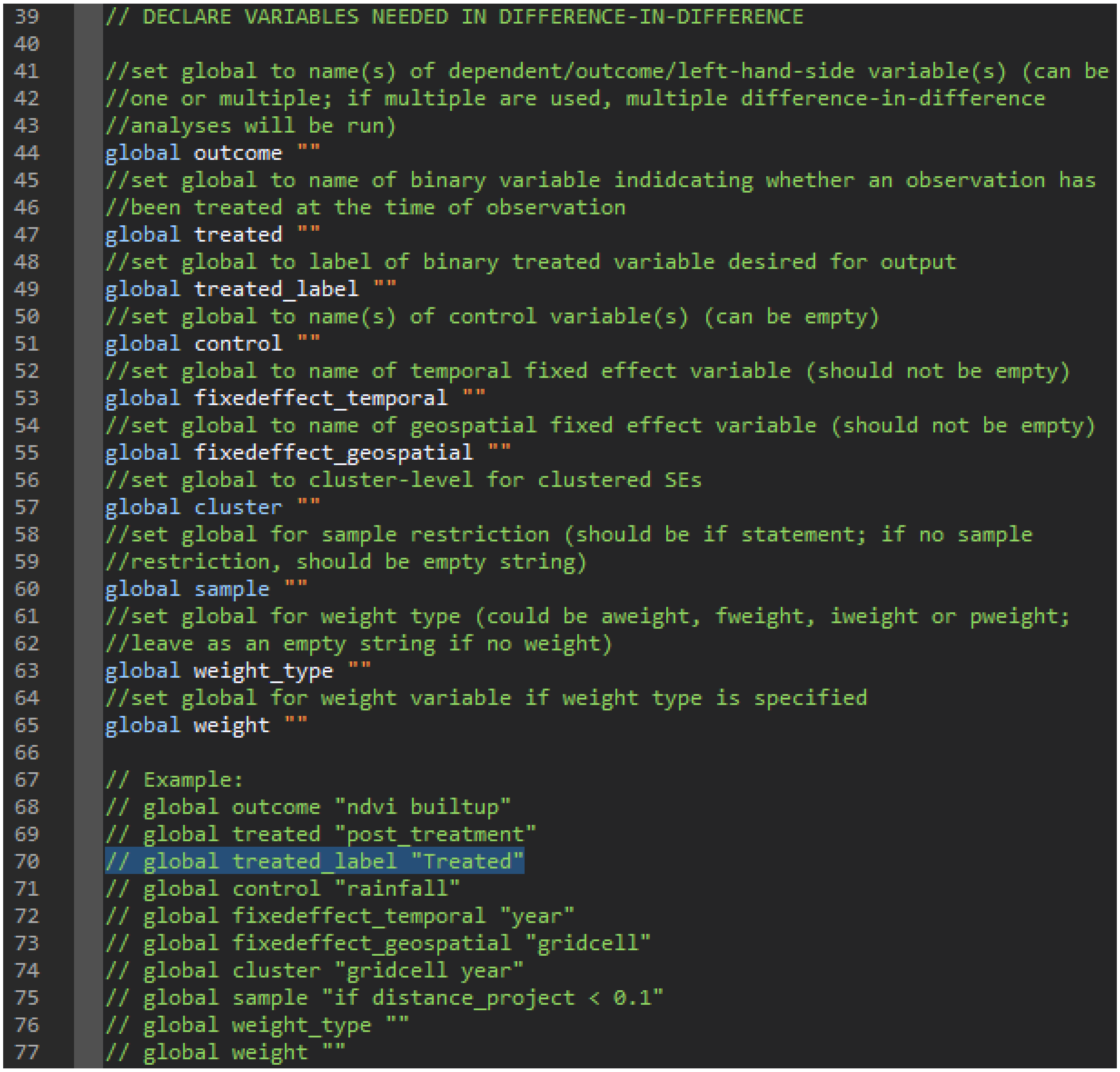


```
17  // DECLARE FILE PATHS
18
19  //set file path for folder storing data needed for difference-in-difference
20  global data "/Users/madeleinewalker/Library/CloudStorage/Box-Box/DEval Mali
 ↳  Irrigation/"
21  //set file path for data file stored in data folder (should be fully cleaned and
22  //in dta format)
23  global datafile "${data}/data/RS_processed_data/{raw_rs_piv}_long.dta"
24  //set file path for folder storing log files from this do-file
25  global logs "${data}/logs/"
26  //set file path for folder storing tables outputed from this do-file
27  global tables "${data}/analysis/tables/"
28  //set file path for folder storing figures outputed from this do-file
29  global figures "${data}/analysis/figures/"
30
31  // Example:
32  // global data "C:\Users\username\projectname\data\"
33  // global datafile "dataset_clean.dta"
34  // global logs "C:\Users\username\projectname\logs\"
35  // global tables "C:\Users\username\projectname\output\tables\"
36  // global figures "C:\Users\username\projectname\output\figures\"
```

Declaring Variables Needed in Analysis (Lines 39-77)

The second part of the do-file requires you to specify the variables relevant to your data in global macros. This section will also ask you for variable labels for key variables, in order to ensure any output in the form of tables or graphs is well-formatted. Some parts are optional, depending on your empirical specification; the optional parts will be marked accordingly. For the examples in this section, we imagine investigating the effect of an irrigation project on agriculture and development, using normalized difference vegetation index (NDVI) and percent built-up as proxies. In this specification, we imagine that the treatment group is grid-cells near project sites that have already undergone treatment and the control group is grid-cells near project sites that have not yet undergone treatment but will eventually.

- ☐ Line 44: specify the name of the outcome variable(s) for the analysis. If there is only one outcome, only one variable name should be listed. If there are multiple, all variable names should be listed and separated by a space. The do-file will perform the analysis for each specified outcome separately. For concerns about multiple testing, indices should be constructed prior to using this do-file and the index variable name should be specified in the outcome list. All outcome variables should be labeled for tables and graphs prior to using this do-file.
  (Example: *global outcome "ndvi builtup")*



```
39  // DECLARE VARIABLES NEEDED IN DIFFERENCE-IN-DIFFERENCE
40  
41  //set global to name(s) of dependent/outcome/left-hand-side variable(s) (can be
42  //one or multiple; if multiple are used, multiple difference-in-difference
43  //analyses will be run)
44  global outcome ""
45  //set global to name of binary variable indidcating whether an observation has
46  //been treated at the time of observation
47  global treated ""
48  //set global to label of binary treated variable desired for output
49  global treated_label ""
50  //set global to name(s) of control variable(s) (can be empty)
51  global control ""
52  //set global to name of temporal fixed effect variable (should not be empty)
53  global fixedeffect_temporal ""
54  //set global to name of geospatial fixed effect variable (should not be empty)
55  global fixedeffect_geospatial ""
56  //set global to cluster-level for clustered SEs
57  global cluster ""
58  //set global for sample restriction (should be if statement; if no sample
59  //restriction, should be empty string)
60  global sample ""
61  //set global for weight type (could be aweight, fweight, iweight or pweight;
62  //leave as an empty string if no weight)
63  global weight_type ""
64  //set global for weight variable if weight type is specified
65  global weight ""
66  
67  // Example:
68  // global outcome "ndvi builtup"
69  // global treated "post_treatment"
70  // global treated_label "Treated"
71  // global control "rainfall"
72  // global fixedeffect_temporal "year"
73  // global fixedeffect_geospatial "gridcell"
74  // global cluster "gridcell year"
75  // global sample "if distance_project < 0.1"
76  // global weight_type ""
77  // global weight ""
```

- □ Line 47: specify the name of the binary indicator variable that is 1 if an observation is in the treatment group and observed after it has been treated and 0 otherwise.
  (Example: *global treated "post_treatment")*
- □ Line 51 (optional): specify the name of the control variable(s) for the analysis. If there are no control variables, leave this field as empty quotation marks. If there is only one control variable, only one variable name should be listed, If there are multiple, all variable names should be listed and separated by a space. All control variables should be labeled for tables and graphs prior to using this do-file.
  (Example: *global control "rainfall"*)
- □ Line 53: specify the name of the temporal fixed effect variable for the analysis. This fixed effect will control for geographic-invariant unobservables within each temporal unit. This fixed effect must be specified.
  (Example: *global fixedeffect_temporal "year"*)

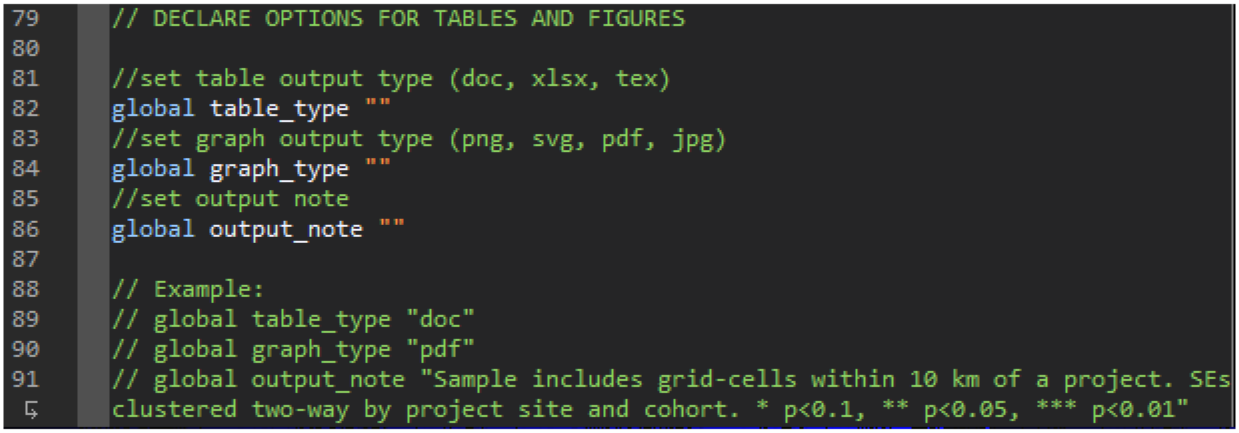
- □ Line 55: specify the name of the geospatial fixed effect variable for the analysis, which should be the same variable as the individual/unit used for the analysis. This fixed effect will control for time-invariant unobservables among each individual/unit. This fixed effect must be specified.
  (Example: *global fixedeffect_geospatial "gridcell"*)
- □ Line 57: specify the name of the variable(s) for which you want to cluster SEs by. If there is only one variable you want to cluster SEs by, only one variable name should be listed. If there are multiple, all variable names should be listed and separated by a space.
  (Example: *global cluster "gridcell year"*)
- □ Line 60 (optional): specify the sample restriction for the analysis using an *if* statement. If there is no sample restriction, leave this field as empty quotation marks. If there are multiple sample restrictions, all restrictions should be included in one *if* statement using logic operators (i.e. & for and; | for or). We recommend either limiting the dataset to individuals in the treatment group who undergo treatment during the time period analyzed (i.e. dropping those who are never treated or always in the time period analyzed) or limiting the analysis to this group using this sample restriction.
  (Example: *global sample "if distance_project < 0.1"*)
- □ Line 63 (optional): specify the type of weight that should be used for the analysis, if you wish to weight observations. You can specify *aweight, fweight, iweight,* or *pweight*. If you do not wish to use weights, leave this field as empty quotation marks.
  (Example: *global weight_type ""*)
- □ Line 65 (optional): specify the name of the variable that should be used for weighting according to Line 63. If you do not wish to use weights, leave this field as empty quotation marks. If Line 63 is not specified, Line 65 should not be specified, and vice versa.
  (Example: *global weight ""*)

Declaring Options for Tables and Figures (Lines 79-91)

The third part of the do-file requires you to specify options for outputting the analysis results in the form of a table and a graph.

- □ Line 82: specify the output type/file extension desired for the regression table to be outputted. The regression table will be saved in this format in the table folder specified in Line 26. You can specify one of the following types: *doc, xlsx, tex.*
  (Example: *global table_type "doc"*)
- □ Line 84: specify the output type/file extension desired for the coefficient plot generated from the regression to be outputted. The coefficient plot will be saved in this format in the figure folder specified in Line 28. You can specify one of the following types: *png, svg, pdf, jpg*.
  (Example: *global graph_type "pdf"*)
- □ Line 86: specify a note to add to the outputted regression table. This note usually includes sample restrictions and information on clustering of SEs. It can also include other information relevant to interpreting the analysis results.

(Example: *global output_note "Sample includes grid-cells within 10 km of a project. SEs clustered two-way by project site and cohort. * p<0.1, ** p<0.05, *** p<0.01"*)

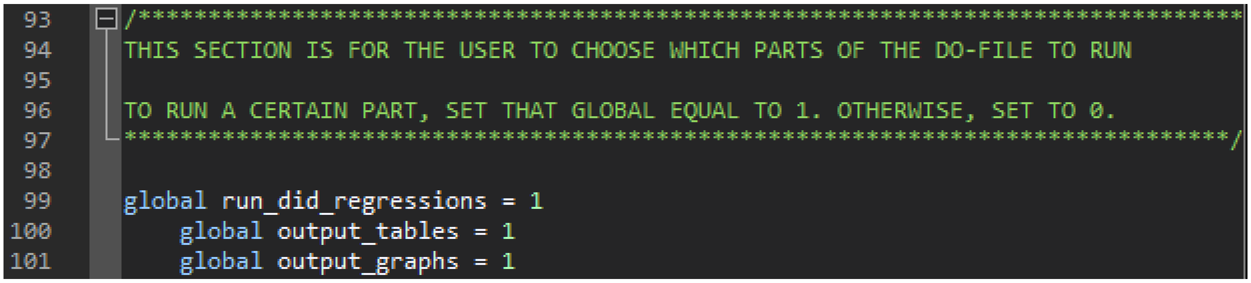


```
// DECLARE OPTIONS FOR TABLES AND FIGURES

//set table output type (doc, xlsx, tex)
global table_type ""
//set graph output type (png, svg, pdf, jpg)
global graph_type ""
//set output note
global output_note ""

// Example:
// global table_type "doc"
// global graph_type "pdf"
// global output_note "Sample includes grid-cells within 10 km of a project. SEs
clustered two-way by project site and cohort. * p<0.1, ** p<0.05, *** p<0.01"
```

Specifying Parts of Do-File to Run (Lines 93-101)

The fourth part of the do-file requires you to specify which parts of the do file you would like to run. The purpose of this section is to allow you to select which parts of the do-file, particularly the output, that is generated each time you run it. This may help with debugging or when you are making changes to the specification. As a default, the file is set to run all parts.

- □ Line 99: specify if you would like to run the DID regressions. If you would like to run it, set it equal to 1; otherwise, set it to 0. Note that setting this to 0 will cause the majority of the do-file to not run.
- □ Line 100: specify if you would like to output the regression table from the DID analysis. If you would like to output the table, set it equal to 1; otherwise, set it to 0.
- □ Line 101: specify if you would like to output the coefficient plot from the DID analysis. If you would like to output the graph, set it equal to 1; otherwise, set it to 0.



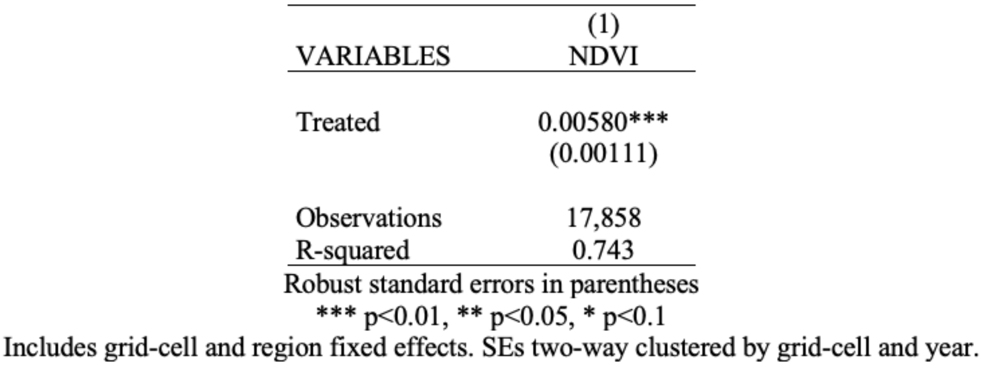
```
93   /**********************************************************************
94   THIS SECTION IS FOR THE USER TO CHOOSE WHICH PARTS OF THE DO-FILE TO RUN
95
96   TO RUN A CERTAIN PART, SET THAT GLOBAL EQUAL TO 1. OTHERWISE, SET TO 0.
97   **********************************************************************/
98
99   global run_did_regressions = 1
100      global output_tables = 1
101      global output_graphs = 1
```

Code Output

The code outputs one type of regression table and one type of coefficient plot showing the result of the DID analysis (the number of each type outputted depends on the number of outcomes specified)..

The first output from the do-file is a regression table pre-specified outcome variable(s) that shows the DID results. The column title will be the outcome variable's label and the row labels will be the label for the treatment variable (the coefficient is the estimated treatment effect) and the labels for the control variables. Both the coefficient and SE are reported in the table, with the latter in parentheses. An example of this table is shown in Table XX.

Table XX: Example of DID Multiple Treatment Times Regression Table

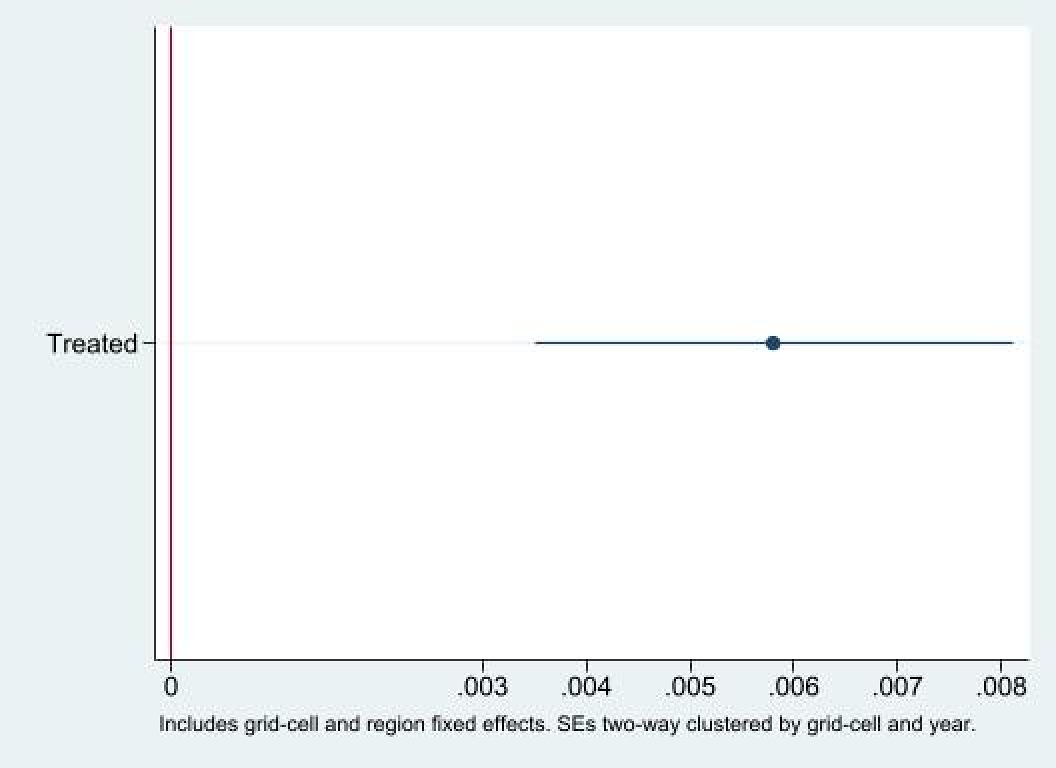


VARIABLES	(1) NDVI
Treated	0.00580*** (0.00111)
Observations	17,858
R-squared	0.743

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Includes grid-cell and region fixed effects. SEs two-way clustered by grid-cell and year.

The second output from the do-file is a coefficient plot pre-specified outcome variable(s) that shows the DID results. The x-axis shows the value of the coefficient estimates, and the y-axis lists the variable labels (including treatment variable and control variables). Both the coefficient estimate (point) and confidence interval (line) are shown in the graph. An example of this table is shown in Table XX.

Figure XX: Example of DID Multiple Treatment Times Coefficient Plot



Includes grid-cell and region fixed effects. SEs two-way clustered by grid-cell and year.

Event Study Analysis

When we consider an intervention with multiple treatment times, we are often interested in considering pre-treatment trends in the outcome variable, as well as any post-treatment trends in the treatment effect. An event study analysis is an extension of difference-in-differences that allows us to accomplish both these goals in one analysis. An event study aligns the time of treatment by looking at the relationship between the outcome variable and the time between observation and treatment. While a difference-in-differences analysis averages all effects into one treatment effect, an event study allows you to decompose these effects across pre-specified time steps. In this way, an event study analysis allows you to see the temporal nature of the treatment effect. For instance, it allows you to understand if the treatment effect grows or shrinks over time and if the treatment effect is transitory or permanent.

Empirical Specification

This do-file conducts an event study analysis for the outcome variable(s). In this specification, a categorical variable is constructed that groups observations based on time between observation and treatment according to user-specified time steps. The relationship between indicator variables for each value of this categorical variable and the outcome variable(s) is then estimated, controlling for controls and temporal and geospatial fixed effects (the latter of which is at the level of the individual/unit). The empirical specification of this analysis is as follows:

$$Outcome_{it} = \beta TimeToTreat_{it} + Controls_{it} + \delta_i + \delta_t + \epsilon_{it}$$

where $Outcome_{it}$ is the outcome of interest for individual/unit *i* at time *t.* $TimeToTreat_{it}$ is a vector of indicator variables that groups observations on the basis of the categorical variable equal to the difference between the time of observation and the time of treatment according to user-specified time steps, with the base reference of 0. As such, β is a vector of coefficients representing the average value of the outcome variable for each value of $TimeToTreat_{it}$. $Controls_{it}$ is an optional vector of control variables. δ_i is a required geospatial/unit fixed effect that controls for time-invariant unobservables within individual/unit *i*. δ_t is a required time fixed effect that controls for geographic-invariant unobservables within temporal unit *t.*

Data Requirements

The analysis code for the event study can be used to better understand the time trends of outcome variables prior to a treatment that occurred at multiple time periods (i.e. staggered). The data used for this analysis must be panel data. The following are required of the data prior to using the analysis code:

- Data from multiple time periods for each observation (the time between the post-treatment time period and the time of treatment should be large enough for effects of the intervention to be detectable)

- A binary variable that is 1 for individuals in the treatment group that were observed after being treated and 0 otherwise
- A variable designated for clustering standard errors (SEs)

Analysis Code

This subsection describes how to utilize the event study analysis do-file to look at the relationship outcome variable(s) and the time between observation and treatment. This file is useful to understand whether DID can be used, as it allows you to consider the parallel trends assumption. This section includes instructions to complete and run the do-file tailored to your dataset and intervention, as well as an overview of the output produced by the do-file.

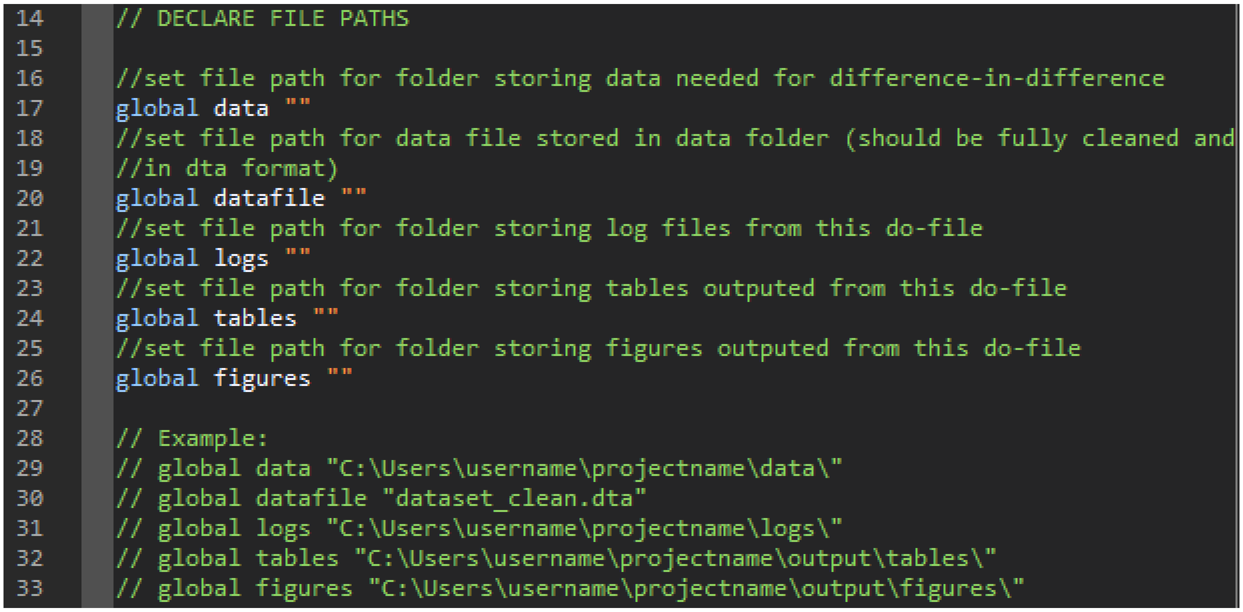
Code Use

This subsection includes instructions on completing the do-file, with screenshots of the code. The code requires you to assign global macros with information relevant to your data and intervention at the start of the do-file. The do-file will then call these global macros to perform the analysis. Note that this file is to be used with the DID multiple treatment times do-file.

Declaring File Paths (Lines 14-33)

The first part of the do-file requires you to specify the file paths relevant to your data in global macros. All fields in this part of the do-file are required to make the do-file run.

- ☐ Line 17: specify the file path for the folder where your data file is stored in global *data* (Example: *global data "C:\Users\username\projectname\data\"*)
- ☐ Line 20: specify the name of your data file, which is located in your data folder. This file must be in dta format.
  (Example: *global datafile "dataset_clean.dta"*)
- ☐ Line 22: specify the file path for the folder where your log files from this do-file will be saved. This log file will record the commands and output of the do-file for you to refer back to at a later time
  (Example: *global logs "C:\Users\username\projectname\logs\"*)
- ☐ Line 25: specify the file path for the folder where any tables produced in this do-file will be saved.
  (Example: *global logs "C:\Users\username\projectname\tables\"*)
- ☐ Line 26: specify the file path for the folder where any figures, including graphs, produced in this do-file will be saved.
  (Example: *global logs "C:\Users\username\projectname\figures\"*)

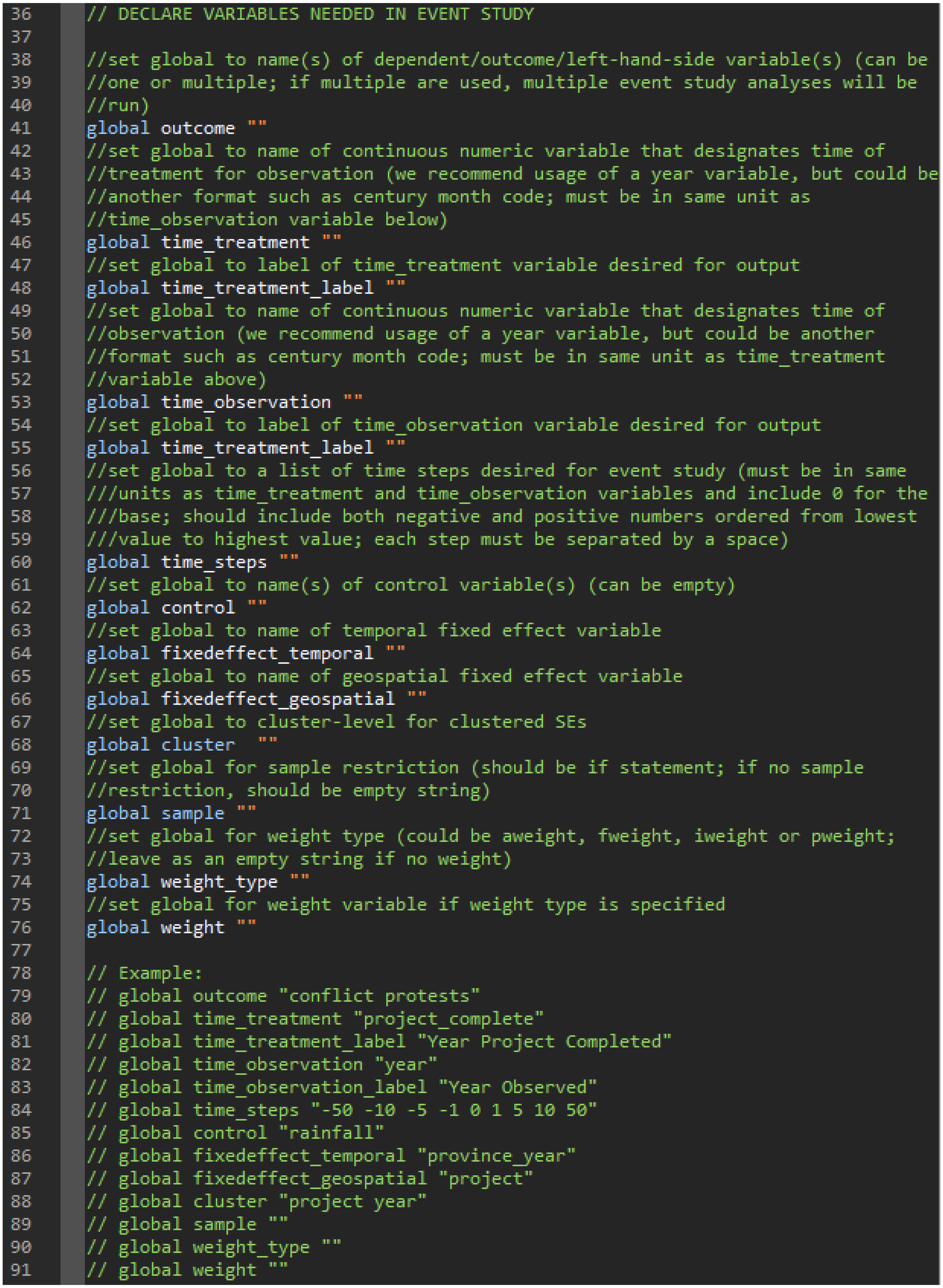


```
14  // DECLARE FILE PATHS
15
16  //set file path for folder storing data needed for difference-in-difference
17  global data ""
18  //set file path for data file stored in data folder (should be fully cleaned and
19  //in dta format)
20  global datafile ""
21  //set file path for folder storing log files from this do-file
22  global logs ""
23  //set file path for folder storing tables outputed from this do-file
24  global tables ""
25  //set file path for folder storing figures outputed from this do-file
26  global figures ""
27
28  // Example:
29  // global data "C:\Users\username\projectname\data\"
30  // global datafile "dataset_clean.dta"
31  // global logs "C:\Users\username\projectname\logs\"
32  // global tables "C:\Users\username\projectname\output\tables\"
33  // global figures "C:\Users\username\projectname\output\figures\"
```

Declaring Variables Needed in Analysis (Lines 36-91)

The second part of the do-file requires you to specify the variables relevant to your data in global macros. This section will also ask you for variable labels for key variables, in order to ensure any output in the form of tables or graphs is well-formatted. Some parts are optional, depending on your empirical specification; the optional parts will be marked accordingly. For the examples in this section, we imagine investigating the effect of an irrigation project on conflict and protest events. In this specification, we imagine that the treatment group is project sites that have already undergone treatment and the control group is project sites that have not yet undergone treatment but will eventually.

- ☐ Line 41: specify the name of the outcome variable(s) for the analysis. If there is only one outcome, only one variable name should be listed. If there are multiple, all variable names should be listed and separated by a space. The do-file will perform the analysis for each specified outcome separately. For concerns about multiple testing, indices should be constructed prior to using this do-file and the index variable name should be specified in the outcome list. All outcome variables should be labeled for tables and graphs prior to using this do-file.
  (Example: *global outcome "conflict protests")*



```
// DECLARE VARIABLES NEEDED IN EVENT STUDY

//set global to name(s) of dependent/outcome/left-hand-side variable(s) (can be
//one or multiple; if multiple are used, multiple event study analyses will be
//run)
global outcome ""
//set global to name of continuous numeric variable that designates time of
//treatment for observation (we recommend usage of a year variable, but could be
//another format such as century month code; must be in same unit as
//time_observation variable below)
global time_treatment ""
//set global to label of time_treatment variable desired for output
global time_treatment_label ""
//set global to name of continuous numeric variable that designates time of
//observation (we recommend usage of a year variable, but could be another
//format such as century month code; must be in same unit as time_treatment
//variable above)
global time_observation ""
//set global to label of time_observation variable desired for output
global time_treatment_label ""
//set global to a list of time steps desired for event study (must be in same
///units as time_treatment and time_observation variables and include 0 for the
///base; should include both negative and positive numbers ordered from lowest
///value to highest value; each step must be separated by a space)
global time_steps ""
//set global to name(s) of control variable(s) (can be empty)
global control ""
//set global to name of temporal fixed effect variable
global fixedeffect_temporal ""
//set global to name of geospatial fixed effect variable
global fixedeffect_geospatial ""
//set global to cluster-level for clustered SEs
global cluster  ""
//set global for sample restriction (should be if statement; if no sample
//restriction, should be empty string)
global sample ""
//set global for weight type (could be aweight, fweight, iweight or pweight;
//leave as an empty string if no weight)
global weight_type ""
//set global for weight variable if weight type is specified
global weight ""

// Example:
// global outcome "conflict protests"
// global time_treatment "project_complete"
// global time_treatment_label "Year Project Completed"
// global time_observation "year"
// global time_observation_label "Year Observed"
// global time_steps "-50 -10 -5 -1 0 1 5 10 50"
// global control "rainfall"
// global fixedeffect_temporal "province_year"
// global fixedeffect_geospatial "project"
// global cluster "project year"
// global sample ""
// global weight_type ""
// global weight ""
```

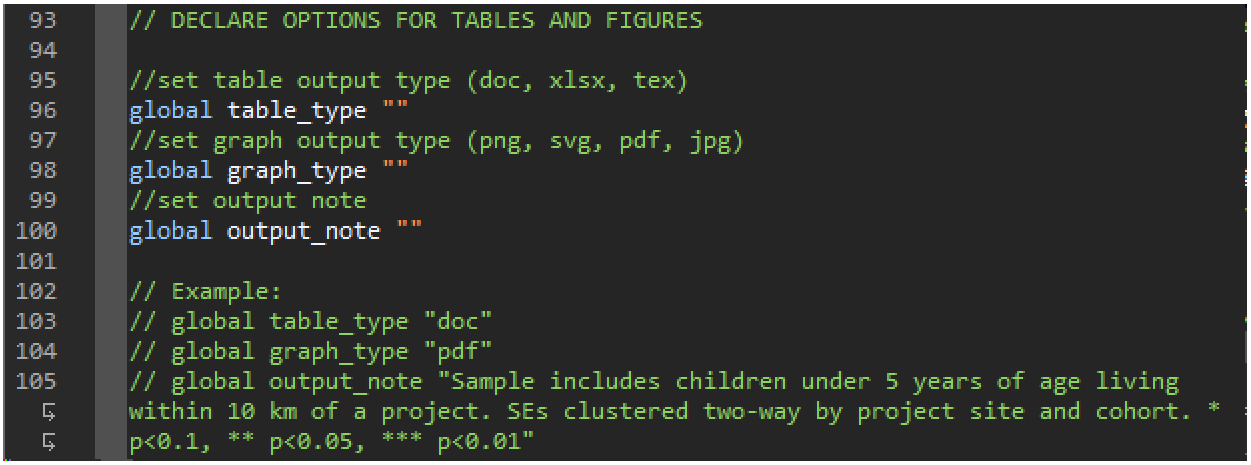
- □ Line 46: specify the name of the continuous variable representing time of treatment for the analysis. This variable must be in integer or double format, not string. We recommend usage of a year variable, but the variable could be in an alternative format, such as century month code.
  (Example: *global time_treatment "project_complete")*
- □ Line 48: specify the label of the continuous variable representing time of treatment specified in Line 46. This label will be used in the outputted tables and graphs.
  (Example: *global time_treatment_label "Year Project Completed")*
- □ Line 53: specify the name of the continuous variable representing time of observation for the analysis. This variable must be in integer or double format, not string. We recommend usage of a year variable, but the variable could be in an alternative format, such as century month code. This variable must be in the same time unit as that specified in Line 46.
  (global time_observation "year"*)*
- □ Line 55: specify the label of the continuous variable representing time of observation specified in Line 53. This label will be used in the outputted tables and graphs.
  (Example: *global time_observation_label "Year Observed")*
- □ Line 60: specify the list of time steps you would like to use to group observations based on the time between observation and treatment. This list must be in the same units as the variables specified in Lines 46 and 53. In addition, it must include negative numbers (representing observations observed before treatment), 0 (representing observations observed at the time of treatment), and positive numbers (representing observations observed after treatment). Each step must be separated by a space.
  (Example: *global time_steps "-50 -10 -5 -1 0 1 5 10 50")*
- □ Line 62 (optional): specify the name of the control variable(s) for the analysis. If there are no control variables, leave this field as empty quotation marks. If there is only one control variable, only one variable name should be listed, If there are multiple, all variable names should be listed and separated by a space. All control variables should be labeled for tables and graphs prior to using this do-file.
  (Example: *global control "rainfall"*)
- □ Line 64: specify the name of the temporal fixed effect variable for the analysis. This fixed effect will control for geographic-invariant unobservables within each temporal unit. This fixed effect must be specified.
  (Example: *global fixedeffect_temporal "province_year"*)
- □ Line 66: specify the name of the geospatial fixed effect variable for the analysis, which should be the same variable as the individual/unit used for the analysis. This fixed effect will control for time-invariant unobservables among each individual/unit. This fixed effect must be specified.
  (Example: *global fixedeffect_geospatial "project"*)
- □ Line 68: specify the name of the variable(s) for which you want to cluster SEs by. If there is only one variable you want to cluster SEs by, only one variable name should be listed. If there are multiple, all variable names should be listed and separated by a space.
  (Example: *global cluster "project year"*)

- □ Line 71 (optional): specify the sample restriction for the analysis using an *if* statement. If there is no sample restriction, leave this field as empty quotation marks. If there are multiple sample restrictions, all restrictions should be included in one *if* statement using logic operators (i.e. & for and; | for or). We recommend either limiting the dataset to individuals in the treatment group who undergo treatment during the time period analyzed (i.e. dropping those who are never treated or always in the time period analyzed) or limiting the analysis to this group using this sample restriction.
  (Example: *global sample ""*)
- □ Line 74 (optional): specify the type of weight that should be used for the analysis, if you wish to weight observations. You can specify *aweight, fweight, iweight,* or *pweight*. If you do not wish to use weights, leave this field as empty quotation marks.
  (Example: *global weight_type ""*)
- □ Line 76 (optional): specify the name of the variable that should be used for weighting according to Line 63. If you do not wish to use weights, leave this field as empty quotation marks. If Line 63 is not specified, Line 65 should not be specified, and vice versa.
  (Example: *global weight ""*)

Declaring Options for Tables and Figures (Lines 93-105)

The third part of the do-file requires you to specify options for outputting the analysis results in the form of a table and a graph.

- □ Line 96: specify the output type/file extension desired for the regression table to be outputted. The regression table will be saved in this format in the table folder specified in Line 26. You can specify one of the following types: *doc, xlsx, tex.*
  (Example: *global table_type "doc"*)



```
93   // DECLARE OPTIONS FOR TABLES AND FIGURES
94
95   //set table output type (doc, xlsx, tex)
96   global table_type ""
97   //set graph output type (png, svg, pdf, jpg)
98   global graph_type ""
99   //set output note
100  global output_note ""
101
102  // Example:
103  // global table_type "doc"
104  // global graph_type "pdf"
105  // global output_note "Sample includes children under 5 years of age living
 ↳   within 10 km of a project. SEs clustered two-way by project site and cohort. *
 ↳   p<0.1, ** p<0.05, *** p<0.01"
```

- □ Line 98: specify the output type/file extension desired for the coefficient plot generated from the regression to be outputted. The coefficient plot will be saved in this format in the

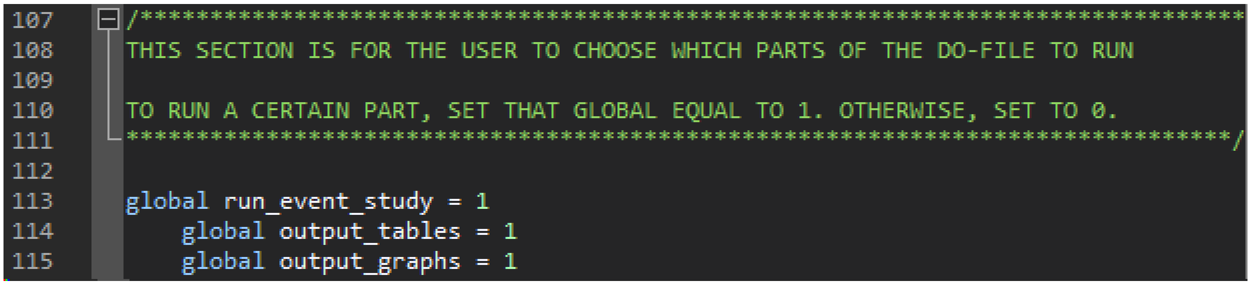
figure folder specified in Line 28. You can specify one of the following types: *png, svg, pdf, jpg*.
(Example: *global graph_type "pdf"*)

- □ Line 100: specify a note to add to the outputted regression table. This note usually includes sample restrictions and information on clustering of SEs. It can also include other information relevant to interpreting the analysis results.
(Example: *global output_note "Sample includes grid-cells within 10 km of a project. SEs clustered two-way by project site and cohort. * p<0.1, ** p<0.05, *** p<0.01"*)

Specifying Parts of Do-File to Run (Lines 107-115)

The fourth part of the do-file requires you to specify which parts of the do file you would like to run. The purpose of this section is to allow you to select which parts of the do-file, particularly the output, that is generated each time you run it. This may help with debugging or when you are making changes to the specification. As a default, the file is set to run all parts.

- □ Line 113: specify if you would like to run the event study analysis. If you would like to run it, set it equal to 1; otherwise, set it to 0. Note that setting this to 0 will cause the majority of the do-file to not run.
- □ Line 114: specify if you would like to output the regression table from the event study analysis. If you would like to output the table, set it equal to 1; otherwise, set it to 0.
- □ Line 115: specify if you would like to output the coefficient plot from the event study analysis. If you would like to output the graph, set it equal to 1; otherwise, set it to 0.



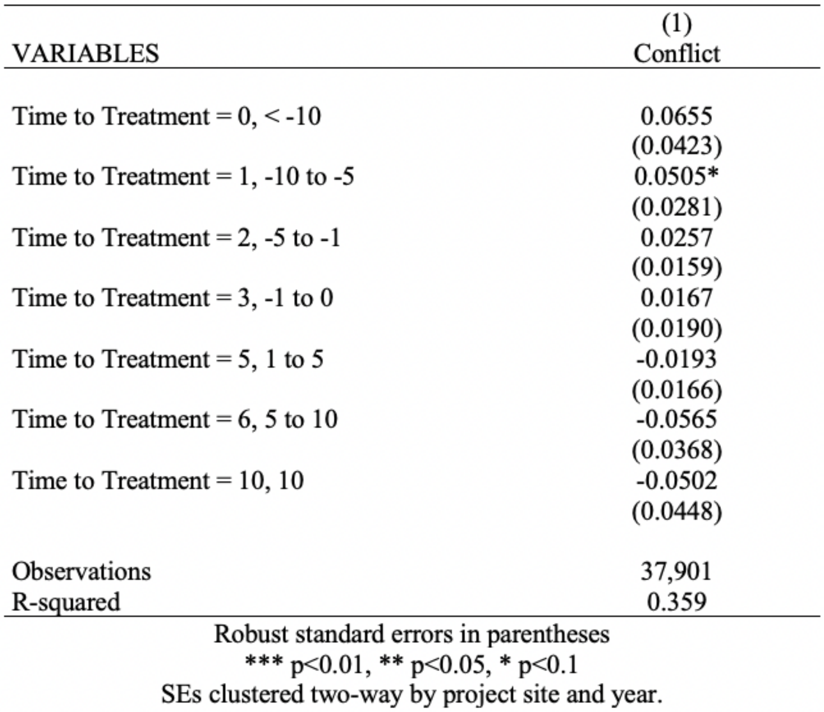
```
107  /*******************************************************************************
108  THIS SECTION IS FOR THE USER TO CHOOSE WHICH PARTS OF THE DO-FILE TO RUN
109
110  TO RUN A CERTAIN PART, SET THAT GLOBAL EQUAL TO 1. OTHERWISE, SET TO 0.
111  ******************************************************************************/
112
113  global run_event_study = 1
114      global output_tables = 1
115      global output_graphs = 1
```

Code Output

The code outputs one type of regression table and one type of coefficient plot showing the result of the event study analysis (the number of each type outputted depends on the number of outcomes specified).

The first output from the do-file is a regression table pre-specified outcome variable(s) that shows the event study results. The column title will be the outcome variable’s label and the row labels will be the time step groupings of the time between observation and treatment. Both the coefficient and SE are reported in the table, with the latter in parentheses. An example of this table is shown in Table XX.

Table XX: Example of Event Study Regression Table

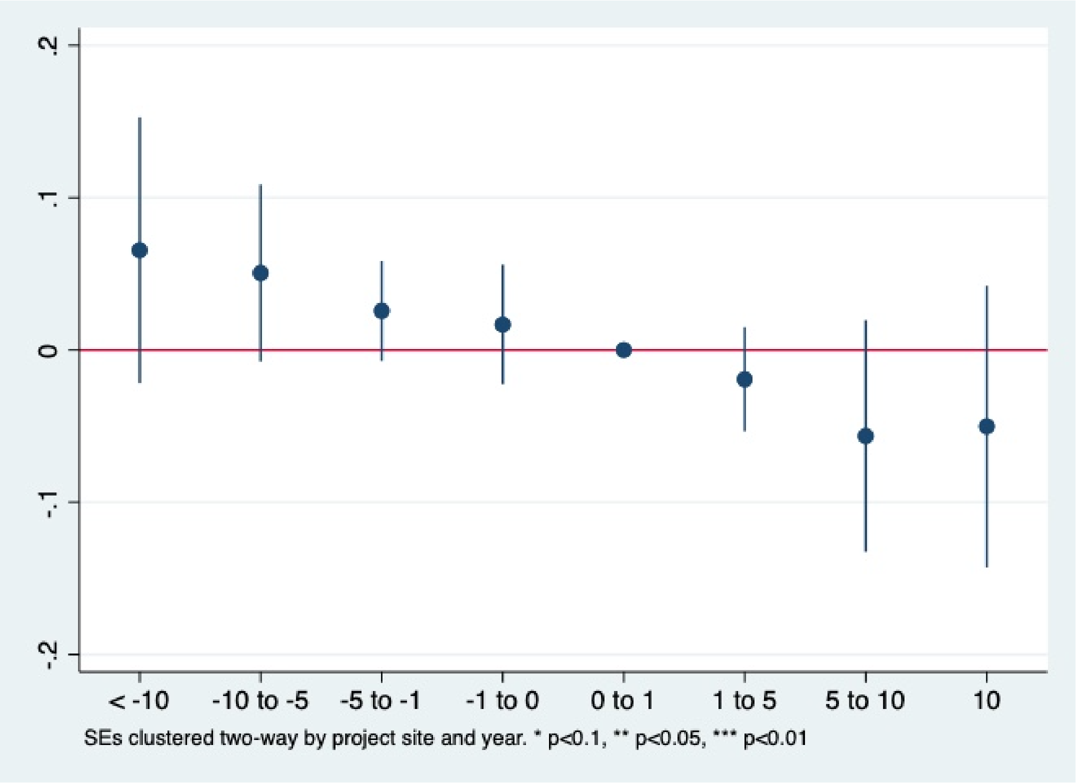


	(1)
VARIABLES	Conflict
Time to Treatment = 0, < -10	0.0655
	(0.0423)
Time to Treatment = 1, -10 to -5	0.0505*
	(0.0281)
Time to Treatment = 2, -5 to -1	0.0257
	(0.0159)
Time to Treatment = 3, -1 to 0	0.0167
	(0.0190)
Time to Treatment = 5, 1 to 5	-0.0193
	(0.0166)
Time to Treatment = 6, 5 to 10	-0.0565
	(0.0368)
Time to Treatment = 10, 10	-0.0502
	(0.0448)
Observations	37,901
R-squared	0.359

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
SEs clustered two-way by project site and year.

The second output from the do-file is a coefficient plot pre-specified outcome variable(s) that shows the event study results. The x-axis shows the time step groupings of the time between observation and treatment, and the y-axis represents the outcome variable(s). Both the coefficient estimate (point) and confidence interval (line) are shown in the graph. An example of this table is shown in Table XX.

Figure XX: Example of Event Study Coefficient Plot



SEs clustered two-way by project site and year. * p<0.1, ** p<0.05, *** p<0.01