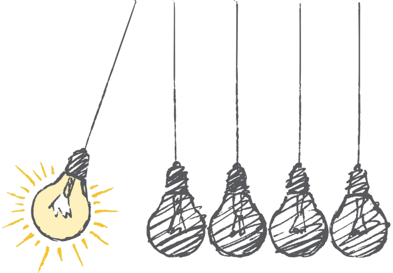
Data Analysis



DIME Analytics RA Onboarding

February 27, 2020

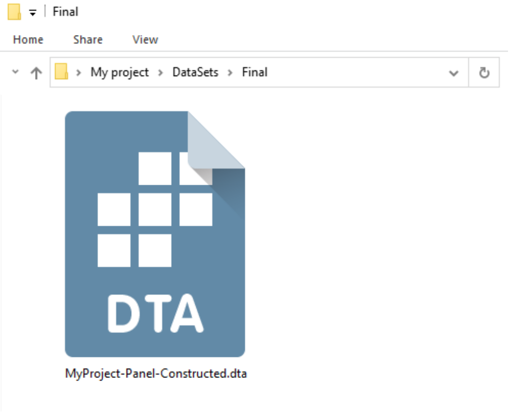
Development Impact Evaluation (DIME)
The World Bank

Introduction

Inputs

Constructed dataset

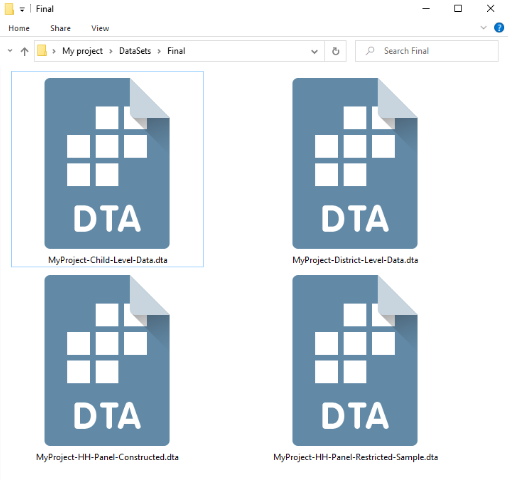
- Include only variables needed for analysis
- Accompanying codebook with description and definition of variables
- Custom-made to answer your analysis questions
  - Sample
  - Unit of observation


Final
Home
Share
View
My project › DataSets › Final
DTA
MyProject-Panel-Constructed.dta

Inputs

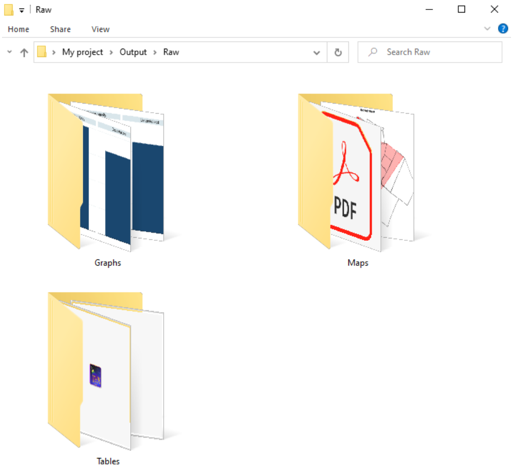
Constructed datasets

- Include only variables needed for analysis
- Accompanying codebook with description and definition of variables
- Custom-made to answer your analysis questions
  - Sample
  - Unit of observation


Final
Home
Share
View
My project › DataSets › Final
Search Final
DTA
MyProject-Child-Level-Data.dta
DTA
MyProject-District-Level-Data.dta
DTA
MyProject-HH-Panel-Constructed.dta
DTA
MyProject-HH-Panel-Restricted-Sample.dta

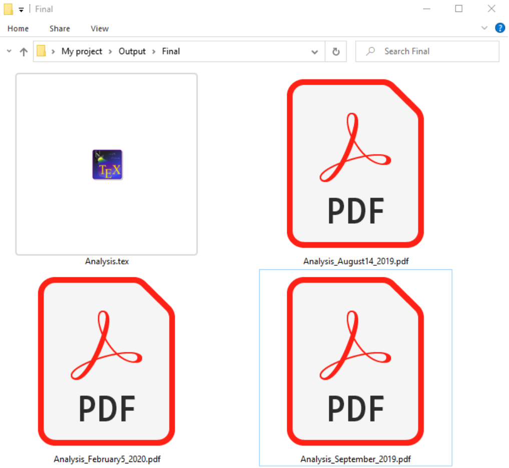
Outputs

- Results are exported to files that can be used as inputs for papers and reports
- Self-standing tables and graphs
- Accessible formats



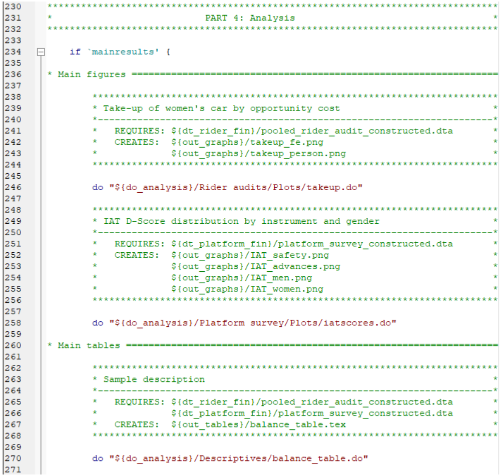
Outputs

- Final outputs such as papers, brief and even reports created to discuss results should be updated automatically when the raw outputs are updated
- LaTeX is an extremely useful tool for doing this
- If you don't know how to use it, check our LaTeX training


Final
Home
Share
View
My project › Output › Final
Search Final
Analysis.tex
Analysis_August14_2019.pdf
Analysis_February5_2020.pdf
Analysis_September_2019.pdf

Documentation

- Another important analysis output is a map of how outputs were created
- The master script is the best way to do this: it should track what are the inputs and outputs of each script that it runs
- A README file is also a good way to do this, particularly when using languages and software



```
********************************************************************************
*                           PART 4: Analysis                                   *
********************************************************************************

    if `mainresults' {

* Main figures =================================================================

        ************************************************************************
        * Take-up of women's car by opportunity cost                           *
        *----------------------------------------------------------------------*
        *   REQUIRES: ${dt_rider_fin}/pooled_rider_audit_constructed.dta       *
        *   CREATES:  ${out_graphs}/takeup_fe.png                              *
        *             ${out_graphs}/takeup_person.png                          *
        ************************************************************************

        do "${do_analysis}/Rider audits/Plots/takeup.do"

        ************************************************************************
        * IAT D-Score distribution by instrument and gender                    *
        *----------------------------------------------------------------------*
        *   REQUIRES: ${dt_platform_fin}/platform_survey_constructed.dta       *
        *   CREATES:  ${out_graphs}/IAT_safety.png                             *
        *             ${out_graphs}/IAT_advances.png                           *
        *             ${out_graphs}/IAT_men.png                                *
        *             ${out_graphs}/IAT_women.png                              *
        ************************************************************************

        do "${do_analysis}/Platform survey/Plots/iatscores.do"

* Main tables ==================================================================

        ************************************************************************
        * Sample description                                                   *
        *----------------------------------------------------------------------*
        *   REQUIRES: ${dt_rider_fin}/pooled_rider_audit_constructed.dta       *
        *             ${dt_platform_fin}/platform_survey_constructed.dta       *
        *   CREATES:  ${out_tables}/balance_table.tex                          *
        ************************************************************************

        do "${do_analysis}/Descriptives/balance_table.do"

```

The analysis process

The analysis process

- Data analysis can be divided into two stages
- During **exploratory data analysis**, the research team will typically look for patterns in the data, in a more descriptive fashion
- The process then progresses into **final analysis** when the team starts to decide what are the main result, that will be part of the research output
- For projects that have pre-analysis plans, the main specifications will be pre-defined, so the exploratory phase has less implications for final outputs

Data work during analysis

- The way you deal with code and outputs for exploratory and final analysis is different
- During exploratory data analysis, you will be tempted to write lots of analysis into one big script, or even directly into the console
- This subtly encourages risky practices such as not clearing the workspace and not reloading the relevant data
- To avoid mistakes, it’s important to take the time to organize the code that you want to use again in a clean manner

Dynamic documents during exploratory analysis

- One way to avoid falling into bad practices during exploratory data analysis is to create dynamic documents
- They allow you to write code, make notes about your observations, and visualize results in one single document
- Stata options include `markstat`, which uses a syntax similar to markdown, and `texdoc`, that combines LaTeX and Stata code
- In R, `RMarkdown` is widely adopted
- The main constraint of this type of dynamic documents is the limited formatting options offered, and the difficulty of handling code and text at the same time

Dynamic documents for final analysis

- Given the limitations of creating dynamic documents in statistical software, team tend to prefer moving to text editor or document preparation systems to write final research outputs
- When setting up this workflow, it's important to think of the integration between code outputs and text
- Code is typically still evolving as papers and reports are written, and it's important to keep code outputs up to date in the final documents
- LaTeX is the most popular way to do this
- It allows you to write references to the files containing analysis results, so that they are updated every time the LaTeX document is compiled

An automated workflow for outputs

Exporting outputs

- It's okay to not export each and every table and graph created during exploratory analysis
- Final outputs should be exported so they are ready to be included to a paper or report
- No manual edits, including formatting, should be necessary after exporting final outputs
- Don't create a workflow that involves copying and pasting across different software

Automating outputs

- Manual edits are difficult to replicate, and you will inevitably need to make changes to the outputs
- The amount of work needed in a copy-paste workflow increases rapidly with the number of outputs, and so do the chances of having the wrong version a result in your paper or report.
- Automating the creation of outputs will save you time by the end of the process
- Polishing final outputs can be time-consuming
- Don’t spend too much time on formatting until your team has agreed on final outputs

In case this is not clear

Don’t ever set up a workflow that requires copying and pasting results

Automating outputs

- Copying results from Excel to Word is error-prone and inefficient
- Copying results from a software console is risk-prone even more inefficient, and completely unnecessary
- There are numerous commands to export outputs from both R and Stata to a myriad of formats
- Our preferred Stata command to export tables are `esttab`, `outreg2`, and `outwrite`
- Our preferred R package to export tables is `stargazer`
- There are many more out there!

Automating outputs

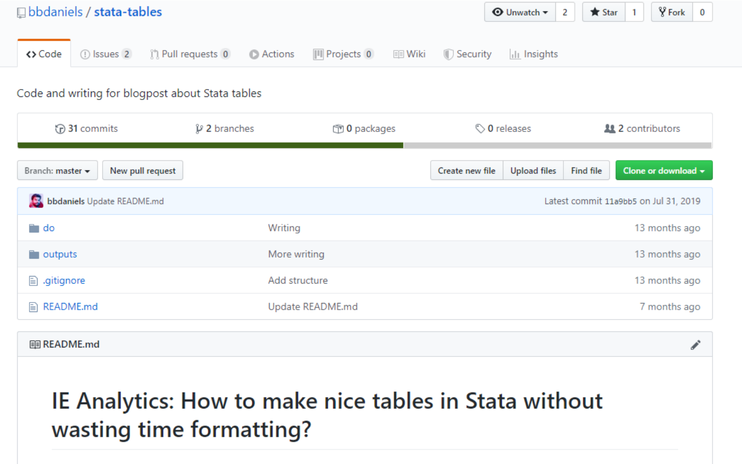
- Copying results from Excel to Word is error-prone and inefficient
- Copying results from a software console is risk-prone even more inefficient, and completely unnecessary
- There are numerous commands to export outputs from both R and Stata to a myriad of formats
- Our preferred Stata command to export tables are `estout`, `outreg2`, and `outwrite`
- Our preferred R package to export tables is `stargazer`
- There are many more out there!

Automating outputs in Stata

- `estout` can solve most of your problems
- It can export both summary statistics and regression tables easily
- It also supports a lot of customization, and exports both to Excel and LaTeX

Automating outputs in Stata

You can find a lot of example do files in `https://github.com/bbdaniels/stata-tables`



Code and writing for blogpost about Stata tables

bbdaniels Update README.md — Latest commit 11a9bb5 on Jul 31, 2019

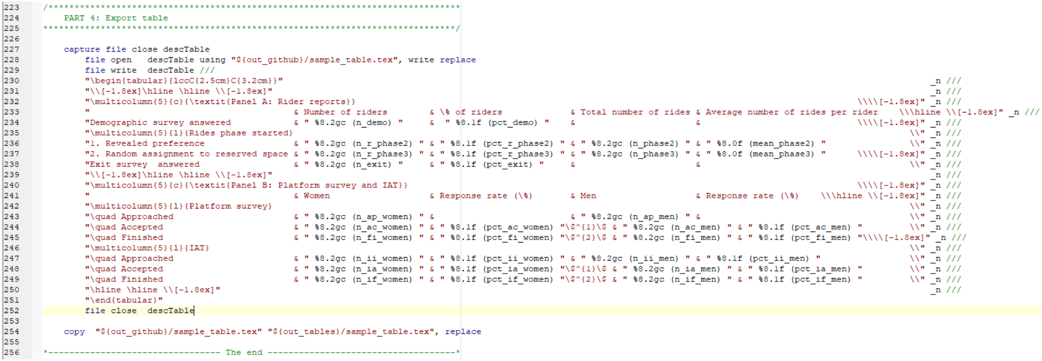
do	Writing	13 months ago
outputs	More writing	13 months ago
.gitignore	Add structure	13 months ago
README.md	Update README.md	7 months ago

README.md

IE Analytics: How to make nice tables in Stata without wasting time formatting?

Automating outputs in Stata

If you need to create a table with a very particular format, consider writing it manually using `file write`:



```
/*******************************************************************************
    PART 4: Export table
*******************************************************************************/

    capture file close descTable
        file open   descTable using "${out_github}/sample_table.tex", write replace
        file write  descTable ///
        "\begin{tabular}{lccC{2.5cm}C{3.2cm}}"                                                                                                                          _n ///
        "\\[-1.8ex]\hline \hline \\[-1.8ex]"                                                                                                                            _n ///
        "\multicolumn{5}{c}{\textit{Panel A: Rider reports}}                                                                                               \\\\[-1.8ex]" _n ///
        "                                      & Number of riders        & \% of riders              & Total number of rides & Average number of rides per rider    \\\hline \\[-1.8ex]" _n ///
        "Demographic survey answered           & " %8.2gc (n_demo) "     &  " %8.1f (pct_demo) "    &                       &                              \\\\[-1.8ex]" _n ///
        "\multicolumn{5}{l}{Rides phase started}                                                                                                                    \\" _n ///
        "1. Revealed preference                & " %8.2gc (n_r_phase2) " & " %8.1f (pct_r_phase2) " & " %8.2gc (n_phase2) " & " %8.0f (mean_phase2) "               \\" _n ///
        "2. Random assignment to reserved space & " %8.2gc (n_r_phase3) " & " %8.1f (pct_r_phase3) " & " %8.2gc (n_phase3) " & " %8.0f (mean_phase3) "      \\\\[-1.8ex]" _n ///
        "Exit survey  answered                 & " %8.2gc (n_exit) "     & " %8.1f (pct_exit) "     &                       &                                       \\" _n ///
        "\\[-1.8ex]\hline \hline \\[-1.8ex]"                                                                                                                            _n ///
        "\multicolumn{5}{c}{\textit{Panel B: Platform survey and IAT}}                                                                                     \\\\[-1.8ex]" _n ///
        "                                      & Women                   & Response rate (\%)       & Men                   & Response rate (\%)    \\\hline \\[-1.8ex]" _n ///
        "\multicolumn{5}{l}{Platform survey}                                                                                                                        \\" _n ///
        "\quad Approached                      & " %8.2gc (n_ap_women) " &                          & " %8.2gc (n_ap_men) " &                                       \\" _n ///
        "\quad Accepted                        & " %8.2gc (n_ac_women) " & " %8.1f (pct_ac_women) "\$^{1}\$ & " %8.2gc (n_ac_men) " & " %8.1f (pct_ac_men) "        \\" _n ///
        "\quad Finished                        & " %8.2gc (n_fi_women) " & " %8.1f (pct_fi_women) "\$^{2}\$ & " %8.2gc (n_fi_men) " & " %8.1f (pct_fi_men) "\\\\[-1.8ex]" _n ///
        "\multicolumn{5}{l}{IAT}                                                                                                                                    \\" _n ///
        "\quad Approached                      & " %8.2gc (n_ii_women) " & " %8.1f (pct_ii_women) " & " %8.2gc (n_ii_men) " & " %8.1f (pct_ii_men) "                \\" _n ///
        "\quad Accepted                        & " %8.2gc (n_ia_women) " & " %8.1f (pct_ia_women) "\$^{1}\$ & " %8.2gc (n_ia_men) " & " %8.1f (pct_ia_men) "        \\" _n ///
        "\quad Finished                        & " %8.2gc (n_if_women) " & " %8.1f (pct_if_women) "\$^{2}\$ & " %8.2gc (n_if_men) " & " %8.1f (pct_if_men) "        \\" _n ///
        "\hline \hline \\[-1.8ex]"                                                                                                                                      _n ///
        "\end{tabular}"                                                                                                                                                 _n ///
        file close  descTable

    copy  "${out_github}/sample_table.tex" "${out_tables}/sample_table.tex", replace

*--------------------------------- The end ------------------------------------*
```

Automating outputs in Stata

- You may also edit the data set directly and export the data to Excel with `export excel`, to csv with `export delimited` or to LaTeX with `dataout`
- If you feel fancy, you can create matrices and export them using `mat2txt` or `outwrite`
- Finally, you can export one and two-way tabulations using `tabout`

Automating outputs in R

- For R users, the `stargazer` package is the easiest way to export formatted regression and summary statistics tables to LaTeX (and html)
- Creating custom tables is also much easier in R, since you can combine objects to data frames and matrices, and use `stargazer` or `write.csv` to export them
- You can find sample codes and examples in our DIME R training repository at `https://github.com/worldbank/dime-r-training`

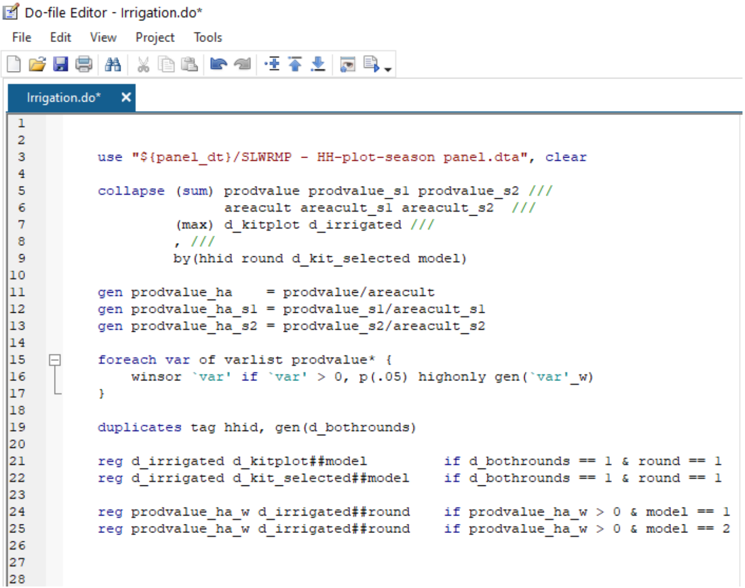
Writing analysis scripts

Script organization

A well-organized analysis script

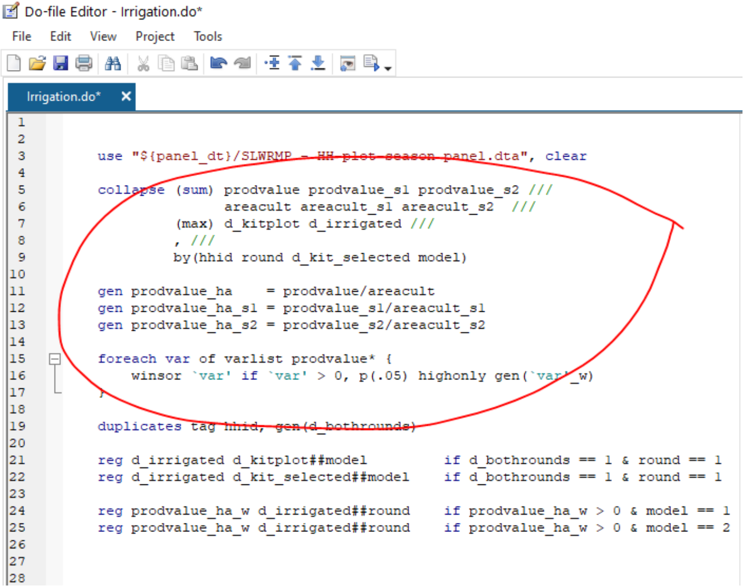
- Starts with a completely fresh workspace
- Loads the constructed dataset
- Makes research decisions explicitly (sampling, clustering, inclusion of controls)
- Has simple code that allows the user to focus on the econometrics
- Exports the results obtained
- Runs completely independently of all other code, except for the master script
- Can be linked to its output by name

Example



```
use "${panel_dt}/SLWRMP - HH-plot-season panel.dta", clear

collapse (sum) prodvalue prodvalue_s1 prodvalue_s2 ///
               areacult areacult_s1 areacult_s2  ///
         (max) d_kitplot d_irrigated ///
         , ///
         by(hhid round d_kit_selected model)

gen prodvalue_ha    = prodvalue/areacult
gen prodvalue_ha_s1 = prodvalue_s1/areacult_s1
gen prodvalue_ha_s2 = prodvalue_s2/areacult_s2

foreach var of varlist prodvalue* {
    winsor `var' if `var' > 0, p(.05) highonly gen(`var'_w)
}

duplicates tag hhid, gen(d_bothrounds)

reg d_irrigated d_kitplot##model        if d_bothrounds == 1 & round == 1
reg d_irrigated d_kit_selected##model   if d_bothrounds == 1 & round == 1

reg prodvalue_ha_w d_irrigated##round   if prodvalue_ha_w > 0 & model == 1
reg prodvalue_ha_w d_irrigated##round   if prodvalue_ha_w > 0 & model == 2
```

Example



```
use "${panel_dt}/SLWRMP - HH-plot-season-panel.dta", clear

collapse (sum) prodvalue prodvalue_s1 prodvalue_s2 ///
               areacult areacult_s1 areacult_s2  ///
         (max) d_kitplot d_irrigated ///
         , ///
         by(hhid round d_kit_selected model)

gen prodvalue_ha    = prodvalue/areacult
gen prodvalue_ha_s1 = prodvalue_s1/areacult_s1
gen prodvalue_ha_s2 = prodvalue_s2/areacult_s2

foreach var of varlist prodvalue* {
    winsor `var' if `var' > 0, p(.05) highonly gen(`var'_w)
}

duplicates tag hhid, gen(d_bothrounds)

reg d_irrigated d_kitplot##model        if d_bothrounds == 1 & round == 1
reg d_irrigated d_kit_selected##model   if d_bothrounds == 1 & round == 1

reg prodvalue_ha_w d_irrigated##round   if prodvalue_ha_w > 0 & model == 1
reg prodvalue_ha_w d_irrigated##round   if prodvalue_ha_w > 0 & model == 2
```

Script organization

- Analysis code should be clean and simple – you may even create one script for each output
- If you have multiple analysis datasets, each of them should have a descriptive name about its sample and unit of observation, so it’s clear which dataset should be used for each piece of analysis
- In both cases, naming should be intuitive so you can trace inputs and outputs of each script

Script organization

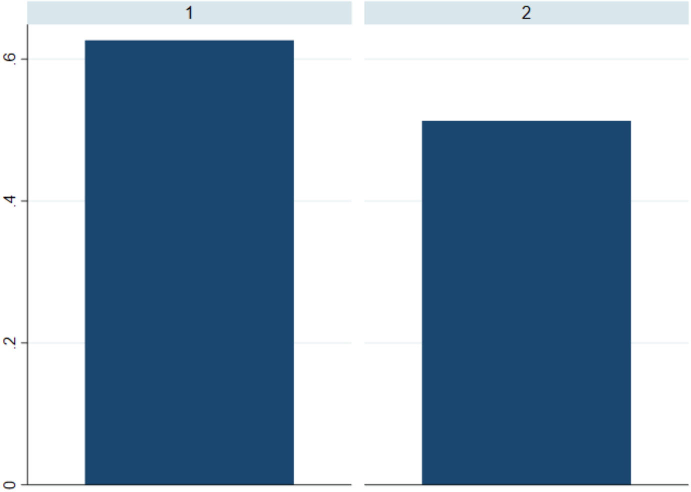
- When your team makes decisions about model specification, can create globals or objects in the master script to use across scripts
- This will ensure specifications are consistent throughout the analysis
- It will also make your code more dynamic, so it is easy to update specifications and results without changing every script
- Use pre-existing commands whenever possible: avoid cluttering your code with complicated commands to create and append intermediate matrices

Final outputs

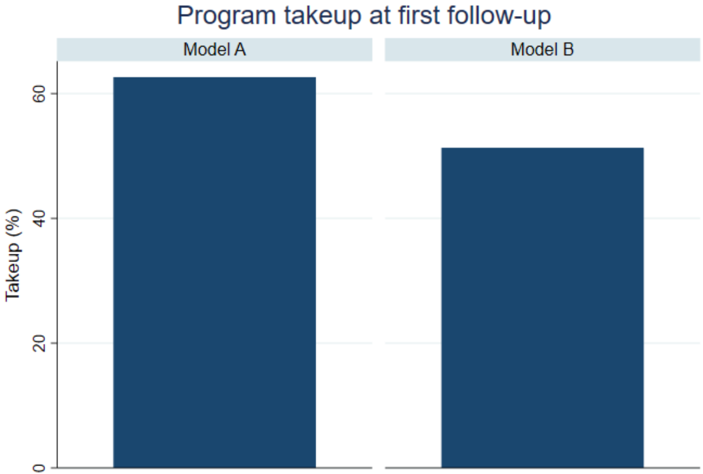
Look at your output!

- Are the looks decent?
- Can someone else understand it?
- Check the number of observations
- Ask yourself if the results make sense
- Check the number of observations again
- Try to interpret the result
- Check the scales

Example

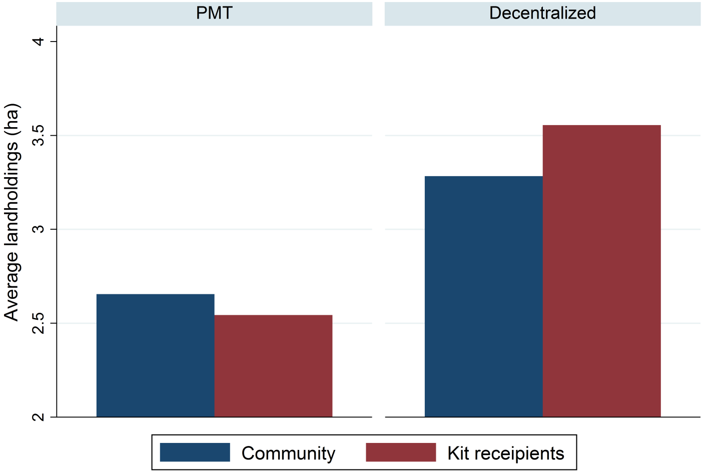


Example



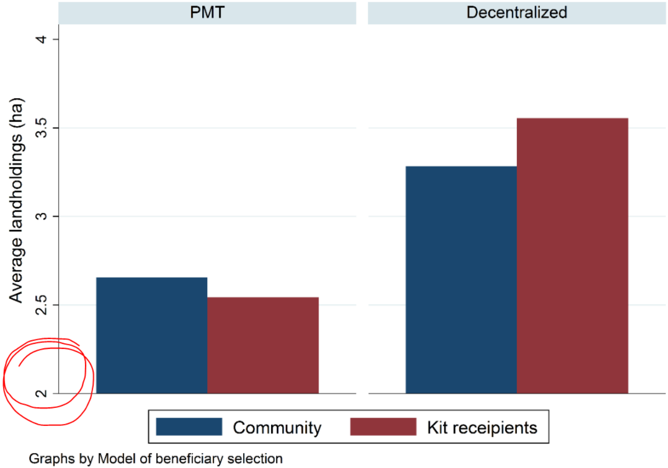
Sample includes all participants selected to receive treatment that were surveyed at follow-up.

Example



Graphs by Model of beneficiary selection

Example


PMT
Decentralized
Average landholdings (ha)
4
3.5
3
2.5
2
Community
Kit receipients
Graphs by Model of beneficiary selection

Beautifying

- Don’t worry about making every exploratory table or graph the best-looking output your team has ever seen
- Getting outputs into publication-ready format is time consuming
- Focus on getting the content right, and only get into the nitty-gritty formatting once your team has agreed on a final version of the output (it won’t be the final version)
- The goal is to reduce the number of times you will need to make very precise adjustments to the aesthetics of the output

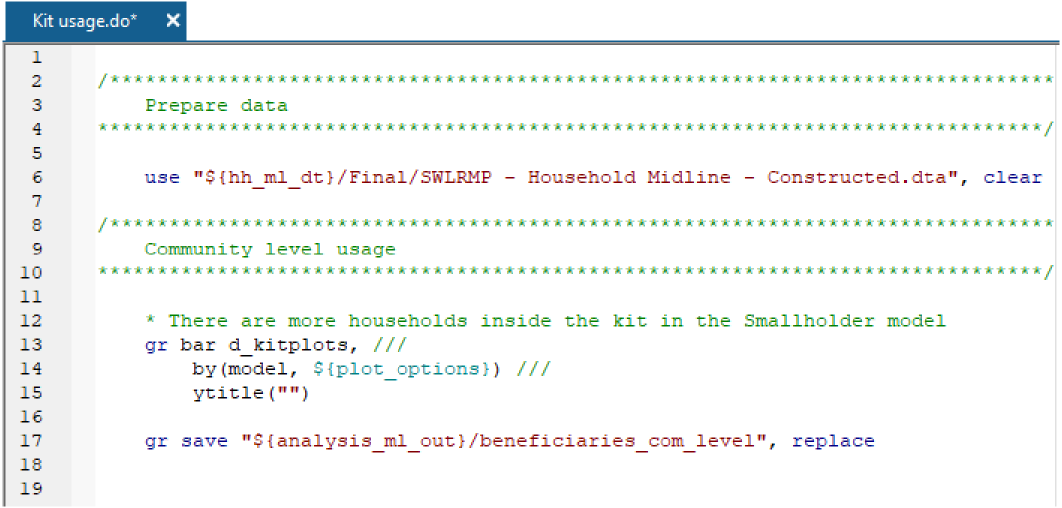
Saving graphs

- saving graph in gph

Final outputs

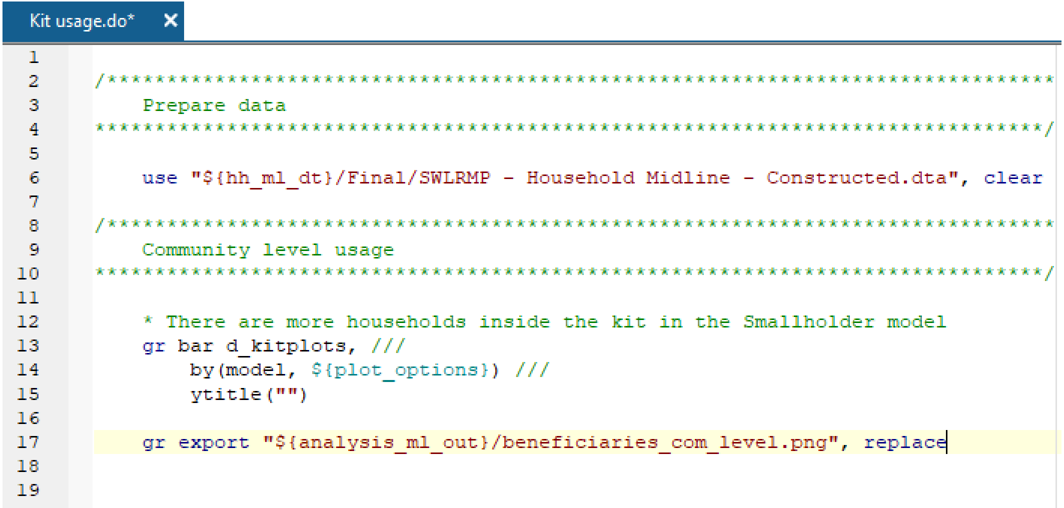
- Should be self-standing: it should be easy to read and understand them with only the information they contain
  - Remember to add labels to variables and axes
  - Include in the notes all relevant information, such as sample used, model specification, units and variable definitions
- Should be saved in accessible formats (pdf, png, jpeg, xls), preferably ones that are lightweight can be version-controlled (tex, csv, eps)

Example



```

/*******************************************************************************
    Prepare data
*******************************************************************************/

    use "${hh_ml_dt}/Final/SWLRMP - Household Midline - Constructed.dta", clear

/*******************************************************************************
    Community level usage
*******************************************************************************/

    * There are more households inside the kit in the Smallholder model
    gr bar d_kitplots, ///
        by(model, ${plot_options}) ///
        ytitle("")

    gr save "${analysis_ml_out}/beneficiaries_com_level", replace


```

Example



```

/*******************************************************************************
    Prepare data
*******************************************************************************/

    use "${hh_ml_dt}/Final/SWLRMP - Household Midline - Constructed.dta", clear

/*******************************************************************************
    Community level usage
*******************************************************************************/

    * There are more households inside the kit in the Smallholder model
    gr bar d_kitplots, ///
        by(model, ${plot_options}) ///
        ytitle("")

    gr export "${analysis_ml_out}/beneficiaries_com_level.png", replace


```

What next?

What next?

- If you follow the steps outlined in this chapter, most of the data work involved in the last step of the research process – publication – will already be done.
- Your analysis code will be organized in a reproducible way, so all you will need to do release a replication package is a last round of code review.
- This will allow you to focus on what matters: writing up your results into a compelling story.

Appendix

Useful resources

- R Graphics Cookbook
- R Graph Gallery
- Stata Visual Library
- Checklist: Reviewing graphs
- Checklist: Reviewing tables