

Research workflow: Stata, Github, Overleaf and R

Workshop on Automation of the Research Process

Alex Hollingsworth, Ohio State University and NBER

Be replicable

Have a workflow

Invest in reducing the cost of errors, future, work.

Invest in future you.

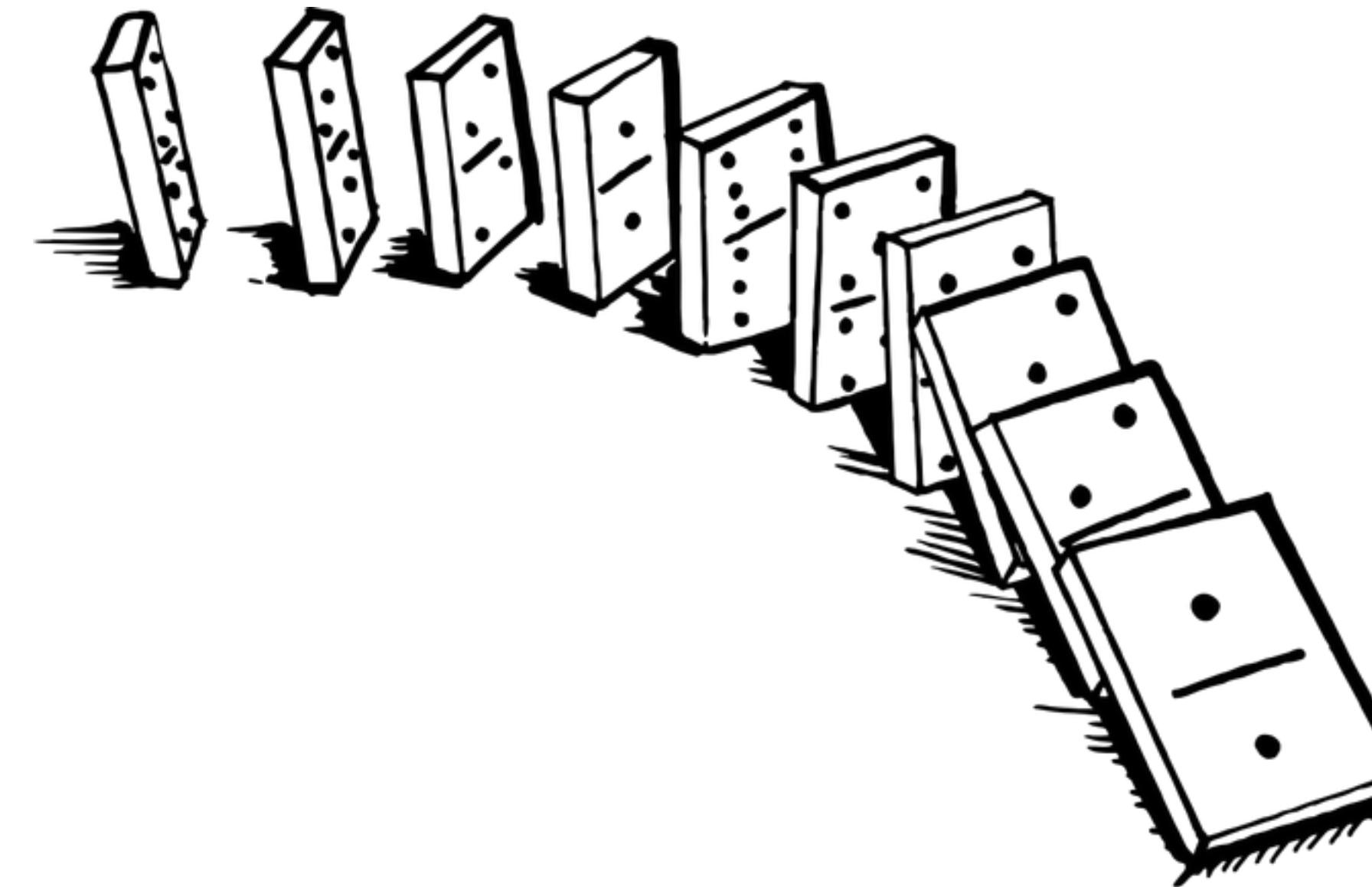
**Specific tools do not matter. Buy
in matters**

- “Imagine if you were reading a great theory paper that had elegant theorems, but then you went to go look at the proofs everything was messy and incomplete...It would make you rethink the validity of the original paper... The code underlying [an empirical] paper, is a lot like the proof [underlying a theory paper].”
- “We want to get the right answer. And having something that’s organized makes [this] more likely”

Julian Reif, Associate Professor of Finance and Economics at Gies College of Business, University of Illinois

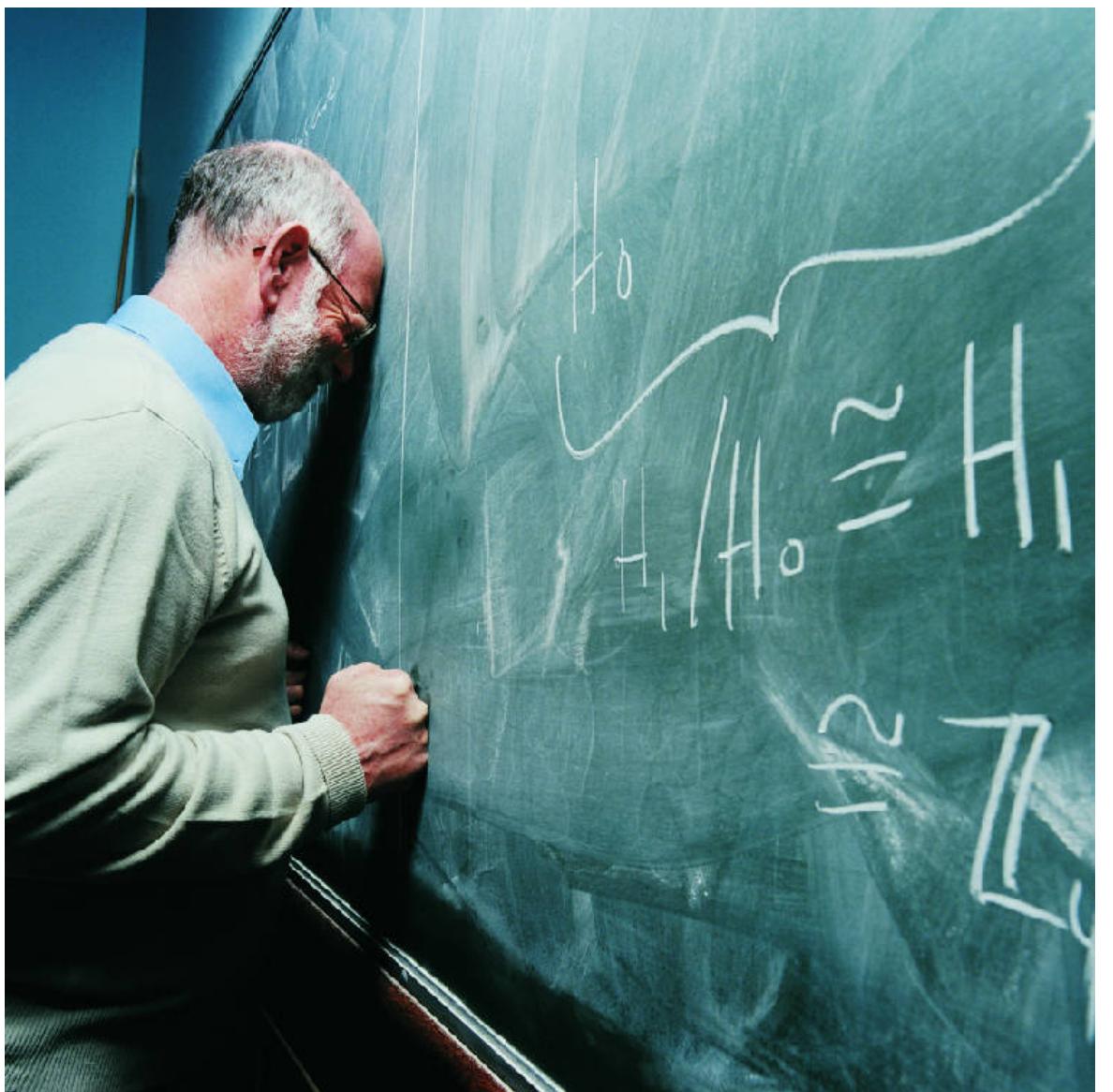
Keystone habit

- **Work in a reproducible manner that can be easily shared with others**
 - Reduces the probability of error
 - Improves the “scientific” process
 - Practice ethical research
 - Reduce temptation to specification hunt
 - p-hack
 - file-drawer results
 - Improves collaboration
 - With your current co-authors
 - Future researchers
 - Smooths team transitions
 - Lends credibility to your results
- A thoughtful, partially automated, workflow is essential to achieving this goal



Why should you care about workflow?

Future you



AMERICAN ECONOMIC ASSOCIATION

Membership About AEA Log In

Journals Annual Meeting Careers Resources More + Q

Home > Journals > AEA Data and Code Policies and Guidance > Data and Code Availability Policy

Journals

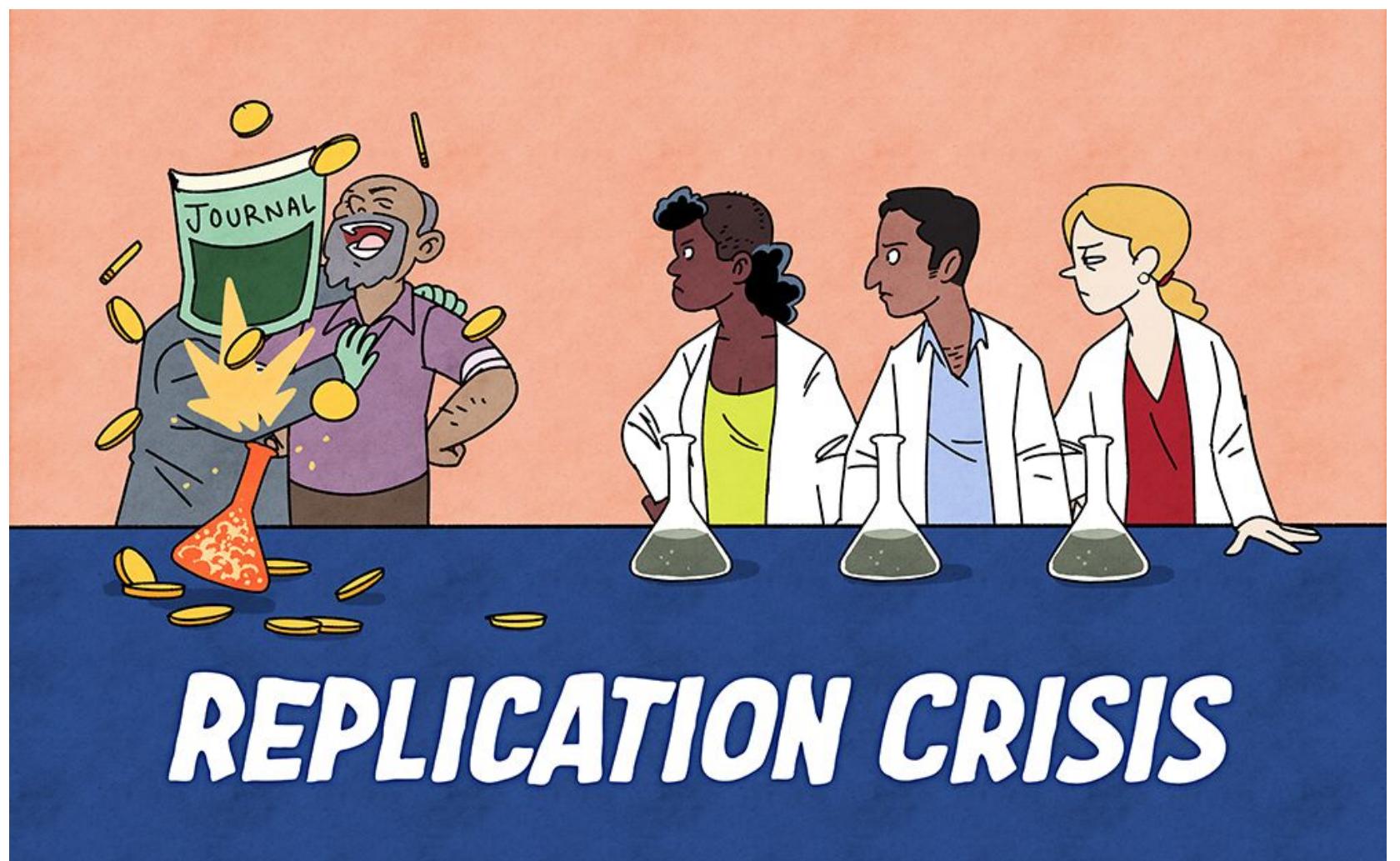
American Economic Review
AER: Insights
AEJ: Applied Economics
AEJ: Economic Policy
AEJ: Macroeconomics
AEJ: Microeconomics
Journal of Economic Literature
Journal of Economic Perspectives

Data and Code Availability Policy

- Content and Scope
 - Data and Software Citations
 - Data Availability Statement
 - Non-Public Data
 - Formats
 - Metadata
 - Version of Record
- Registration of Randomized Control Trials
- Ethics Approval
- Instructions

It is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented and access to the data and code is non-exclusive to the authors.

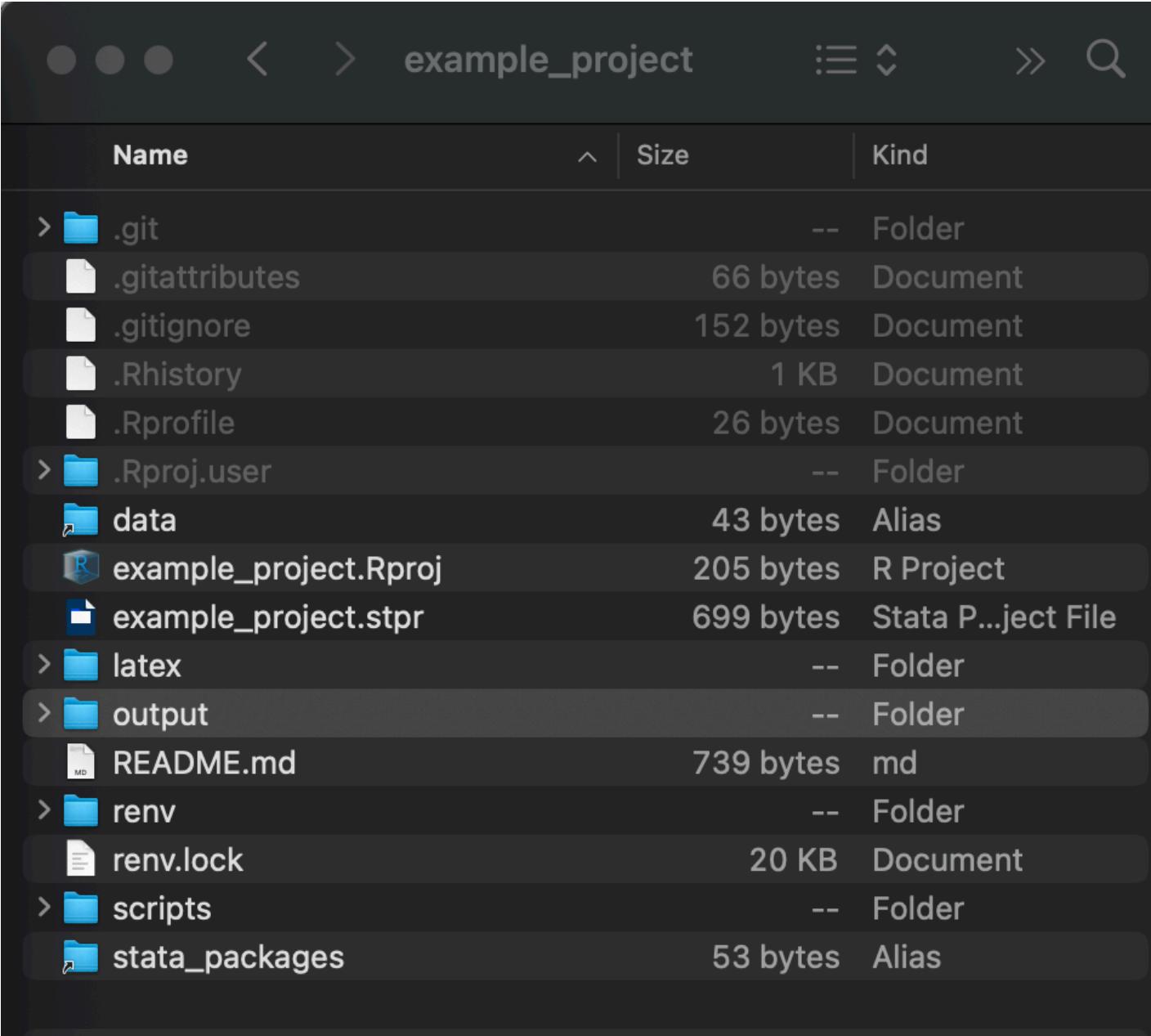
Future research



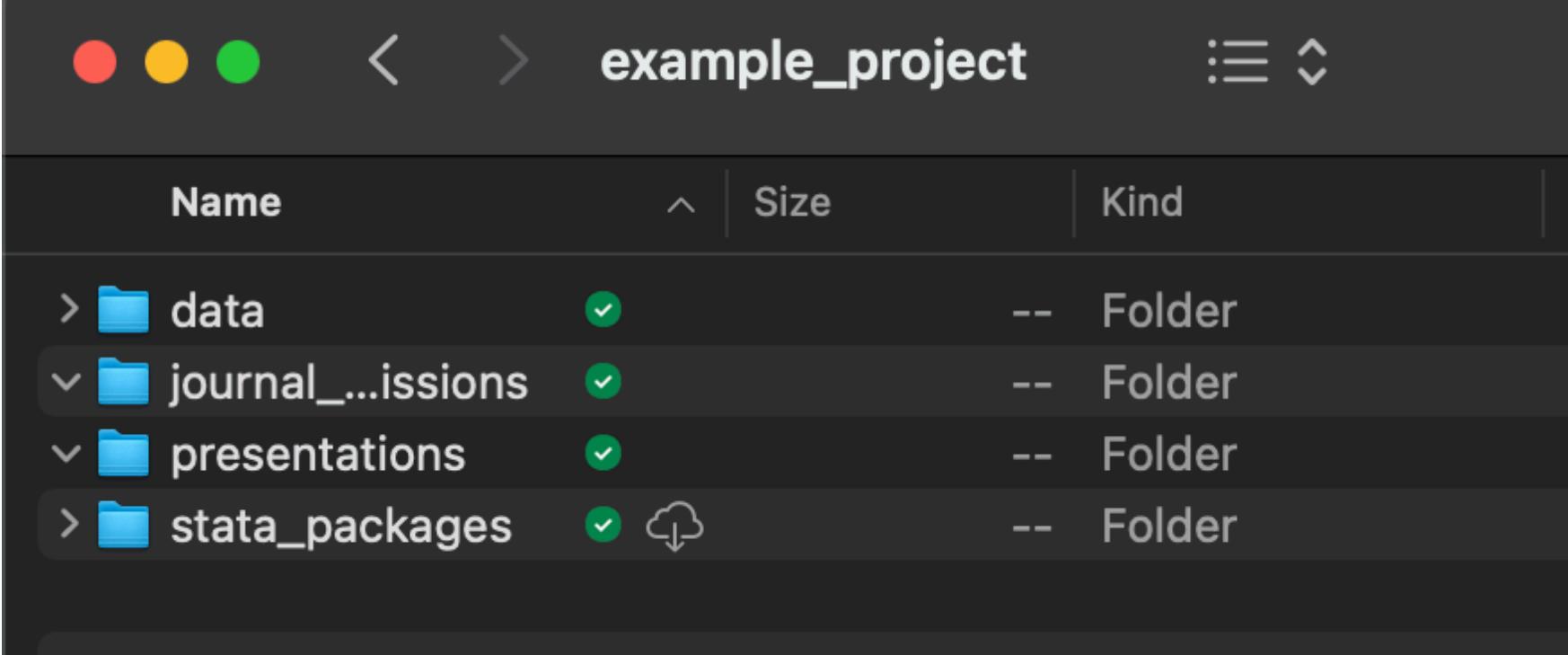
Set-up

Visual overview

GitHub for code



Dropbox for data/other things



Coding/analysis on server

```
R version 4.3.1 (2023-06-16) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin22.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> sin(2)
[1] 0.9092974
```

Local coding/writing using VSCode

A screenshot of VSCode with a LaTeX editor. The code editor shows a file named 'tables_and_figures.tex'. A preview panel on the right shows a PDF document titled 'Tables' containing a table with summary statistics. The table is as follows:

	Mean	S.D.	Min.	Max.	N
Price	6165.26	2949.50	3291.00	15906.00	74
Mileage (mpg)	21.30	5.79	12.00	41.00	74
<i>Mortality Sample</i>					
Car type	0.30	0.46	0.00	1.00	74
Headroom (in.)	2.99	0.85	1.50	5.00	74
Trunk space (cu. ft.)	13.77	4.28	2.00	32.00	74
Weight (lbs.)	3019.46	777.19	1760.00	4840.00	74
Length (in.)	187.93	22.27	142.00	233.00	74
Turn Circle (ft.)	39.65	4.40	31.00	51.00	74

Online writing using overleaf

A screenshot of the Overleaf online LaTeX editor. The project name is 'example_project'. The code editor shows the same 'tables_and_figures.tex' file as the VSCode screenshot. A preview panel on the right shows the same PDF output as the VSCode preview. The table data is identical to the one shown in the VSCode preview.

GitHub is the “glue”

- It's more complicated than it needs to be for your purposes
- Think of it as an extra save button you need to click
 - fancy save
 - Handles merging really well

The screenshot shows a GitHub interface. At the top, there are dropdown menus for 'Current Repository' (set to 'example_project'), 'Current Branch' (set to 'master'), and 'Fetch origin' (last fetched 7 minutes ago). Below this, a pull request titled 'update for new presentation' is shown, created by Alex Hollingsworth 82e24f0 ago, with +425 -323 changes. The 'History' tab is selected, showing a list of commits:

- update (Alex Hollingsworth, 7 minutes ago)
- Update .gitignore (Alex Hollingsworth, 7 minutes ago)
- update for new presentation (Alex Hollingsworth, 12 minutes ago)
- edit a typo (hollina, 9 months ago)
- Update 2.test_script.R (hollina, 9 months ago)
- update to include renv file (hollina, 9 months ago)
- add dropbox link (Alex Hollingsworth, 3 years ago)
- Update README.md (Alex Hollingsworth, 3 years ago)
- fixed issue where CRAN mirror was not set (Alex Hollingsworth, 3 years ago)
- this is the first upload of these files (Alex Hollingsworth, 3 years ago)
- initial commit

The right side of the interface displays a detailed diff for the 'update for new presentation' commit, specifically for the file 'latex/latex_extras/dynamic_tables.tex'. The diff shows 13 changed files, with 15 additions and 16 deletions. The changes include modifications to command definitions and document structure.

Line	Change Type	Text
15	+	\makeatother
16	+	%\let\estinput=\@input % define a new input command so that we can still flatten the document
9	17	
10	18	\newcommand{\estwide}[3]{
11	-	\vspace{.75ex}{
12	19	+ \vspace{.25ex}{
13	20	\textsymbols% Note the added command here
14	21	\begin{tabular}{
15	22	{\textwidth}{@\hskip\tabcolsep\extracolsep\fill}l*{#2}{#3}}
16	23	\toprule
17	24	- \estinput{#1}
18	25	+ \estinput{#1}
19	26	\bottomrule
20	27	\addlinespace{.75ex}
21	28	\end{tabular}
22	29	}
23	30	
24	31	\newcommand{\estaauto}[3]{
25	32	- \vspace{.75ex}{
26	33	+ \vspace{.25ex}{
27	34	\textsymbols% Note the added command here
28	35	\begin{tabular}{l*{#2}{#3}}
29	36	\toprule

Basic idea of GitHub

- You write/edit some code or a document
 - You click save. This saves locally.

Basic idea of GitHub

- You write/edit some code or a document
 - You click save. This saves locally.
- After a while, say you finish a do-file, need to end the work day,
 - You write a little message in your GitHub saying what you did
 - AKA commit

Basic idea of GitHub

- You write/edit some code or a document
 - You click save. This saves locally.
- After a while, say you finish a do-file, need to end the work day,
 - You write a little message in your GitHub saying what you did
- You check to see if anyone else has made changes
 - AKA fetch

Basic idea of GitHub

- You write/edit some code or a document
 - You click save. This saves locally.
- After a while, say you finish a do-file, need to end the work day,
 - You write a little message in your GitHub saying what you did
- You check to see if anyone else has made changes
- Then you “push” these changes to a server that can then be accessed on all your devices

Basic idea of GitHub

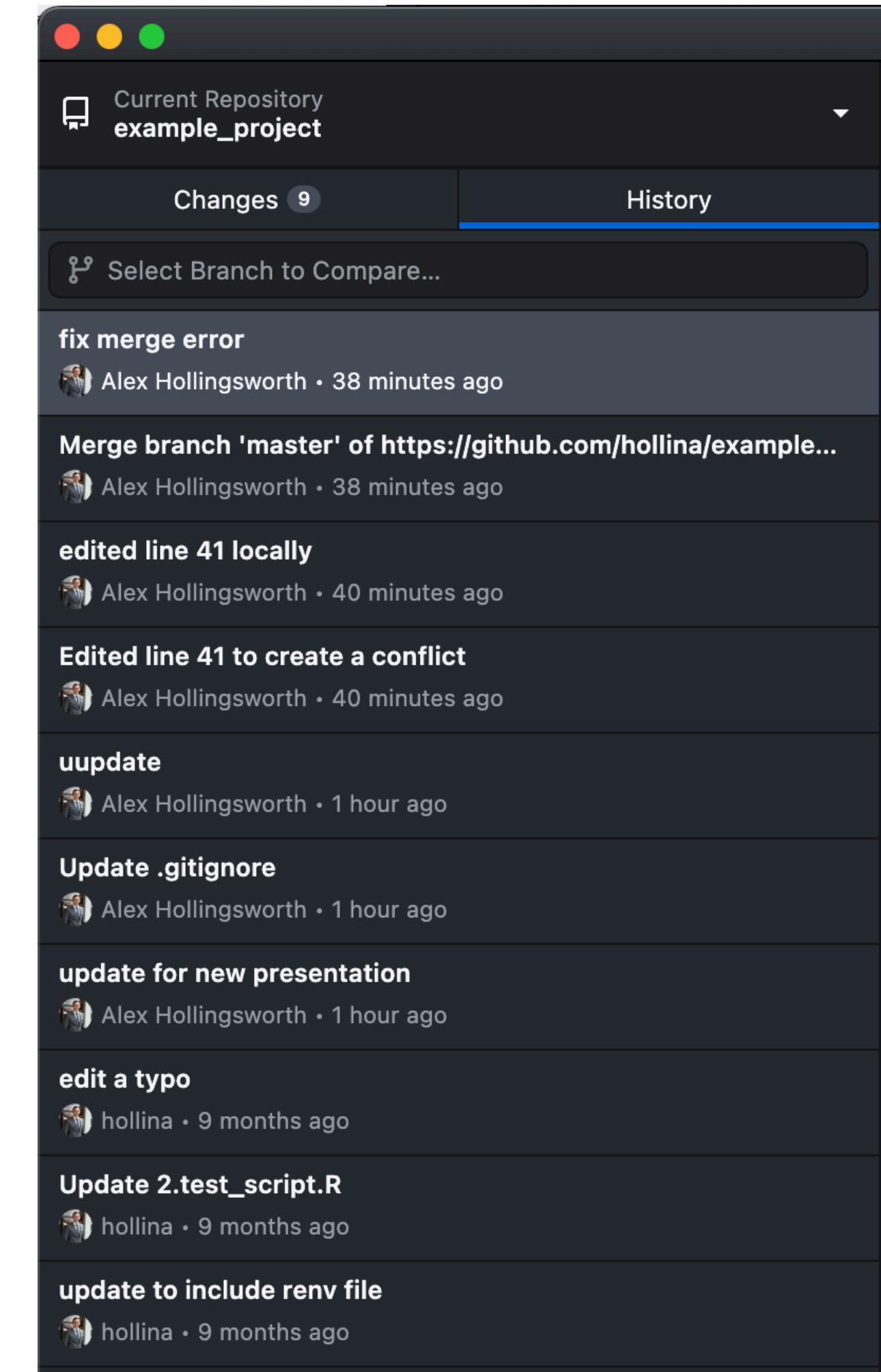
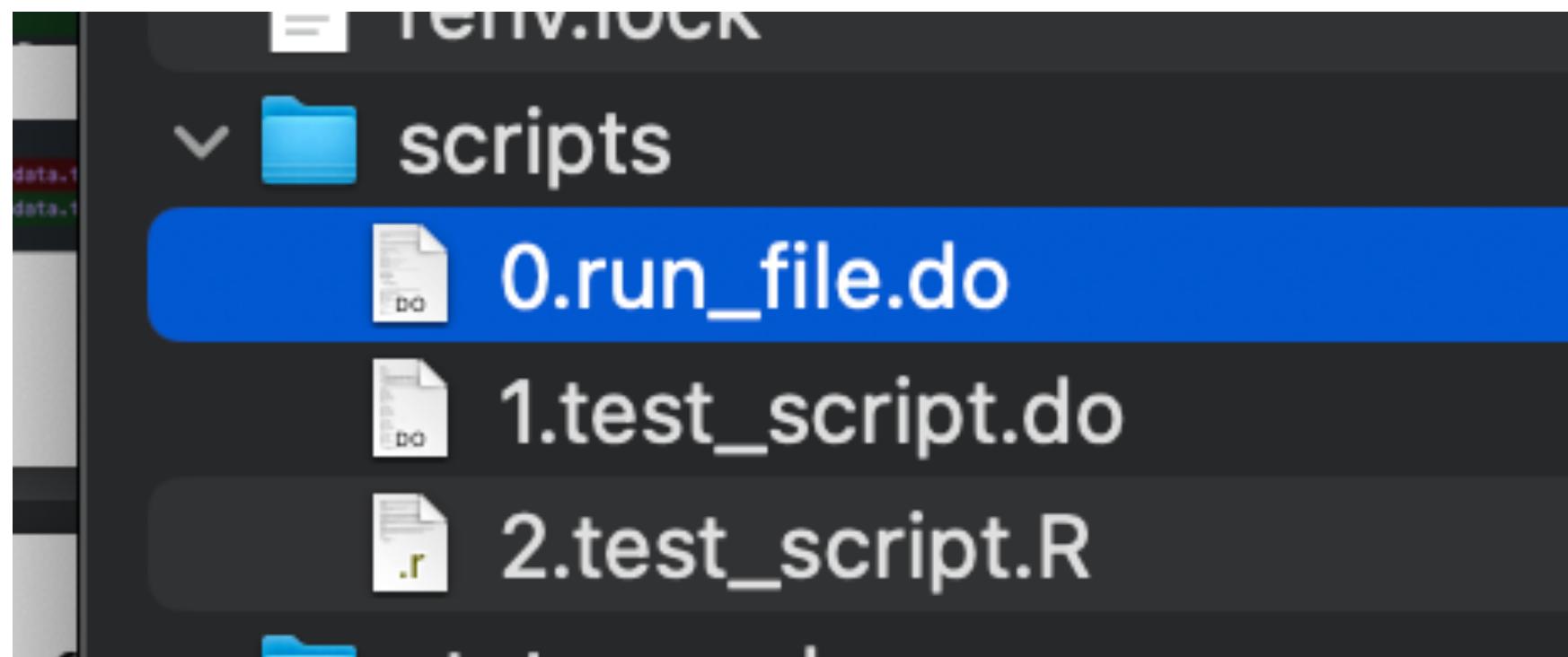
- You write/edit some code or a document
 - You click save. This saves locally.
- After a while, say you finish a do-file, need to end the work day,
 - You write a little message in your GitHub saying what you did
- You check to see if anyone else has made changes
- Then you “push” these changes to a server that can then be accessed on all your devices
- If there are any conflicts, you solve them via “merge”

Basic idea of GitHub

- There is a detailed log kept of all changes made to code
- Everything that you want coarser version control for is stored using another system like dropbox

Main benefit: version control

- Only one file! No more of this
``version_1245_final_for_real_this_time.do``



Main benefit: version control

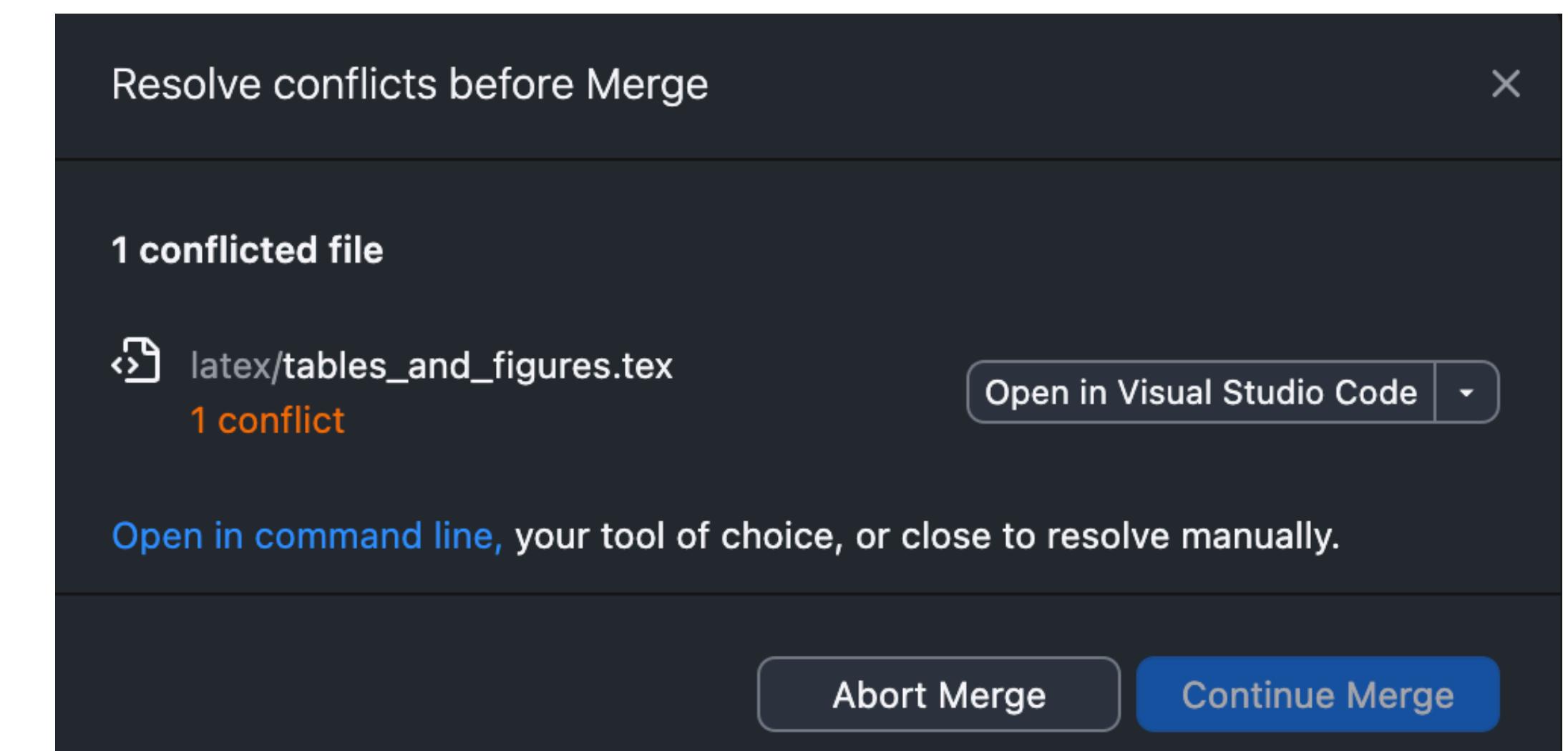
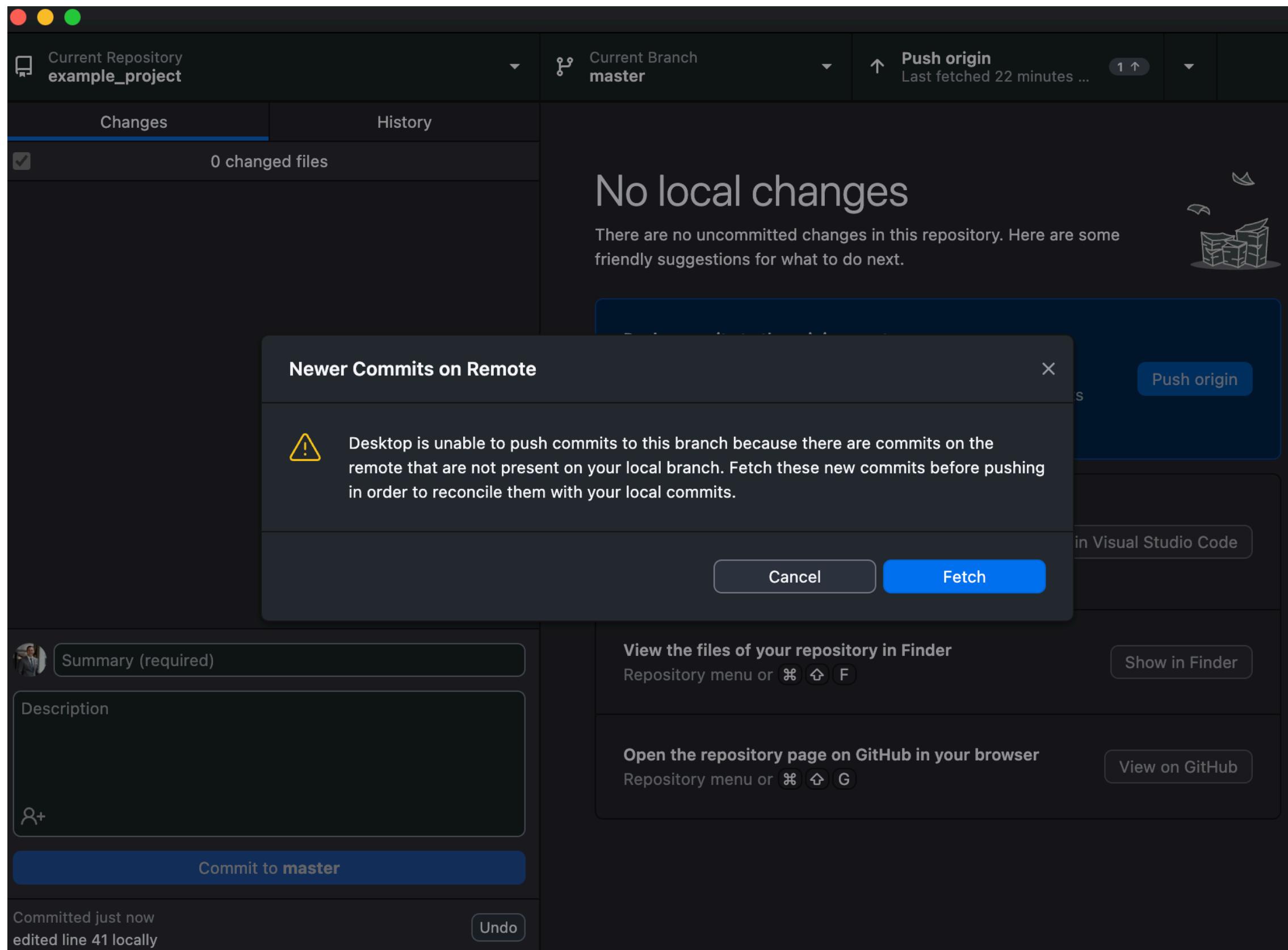
```
latex/duke_endowment.tex
  2518 | 2516 | \Fignote{* p $< $ 0.1, ** p $< $ 0.05, *** p $< $ 0.01.
  2519 | 2517 | Each coefficient comes from a separate Poisson pseudo maximum likelihood regression.
  2520 | 2518 | In columns 1 to 3, the unit of observation is a county-by-year of birth cell.
  2521 | - | In columns 4 to 6, the unit of observation is a birth county by birth year by follow-up
        | year triplet.
  2519 | + | In columns 4 to 6, the unit of observation is a birth county by birth year by follow-up
        | age triplet.
```

replication/analysis/output/appendix/table_hX_maternal_mortality_diff_specs.tex

```
@@ -1,11 +1,11 @@
  1 | 1 | & \multicolumn{3}{c}{\shortstack{Number of Deaths}} & \multicolumn{3}{c}{\shortstack{Number of Births}}
  2 | 2 | \multicolumn{1}{c}{(1)} & \multicolumn{1}{c}{(2)} & \multicolumn{1}{c}{(3)}
  3 | - | \midrule \emph{A. Maternal mortality} &&&& \\ \addlinespace\hline
        | & -2.60 & -11.87 & -12.83
  4 | - | \sim & (8.39) & (7.33)
  5 | 3 | + \midrule \emph{A. Maternal mortality} &&&& \\ \addlinespace\hline
        | & -3.60 & -16.84* & -18.37**
  6 | 4 | + \sim & (8.39) & (7.33)
  7 | 5 | \addlinespace\hline Observations&\multicolumn{3}{l}{51 [1, 100]} & \multicolumn{3}{l}{51 [1, 100]} & \multicolumn{3}{l}{51 [1, 100]}
```

Second benefit: merge conflicts

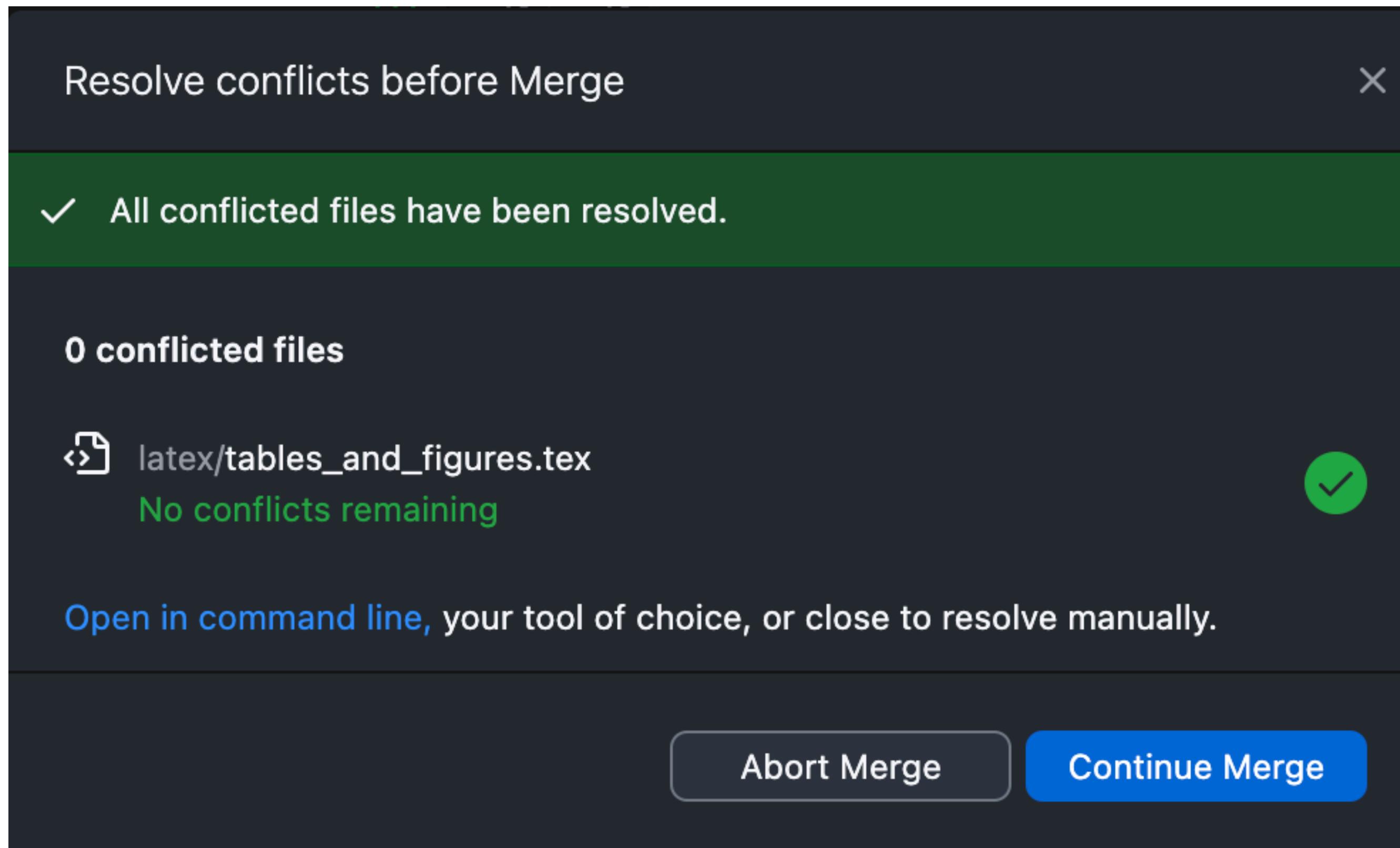
- Say you and someone else edited the same document. Sound familiar?



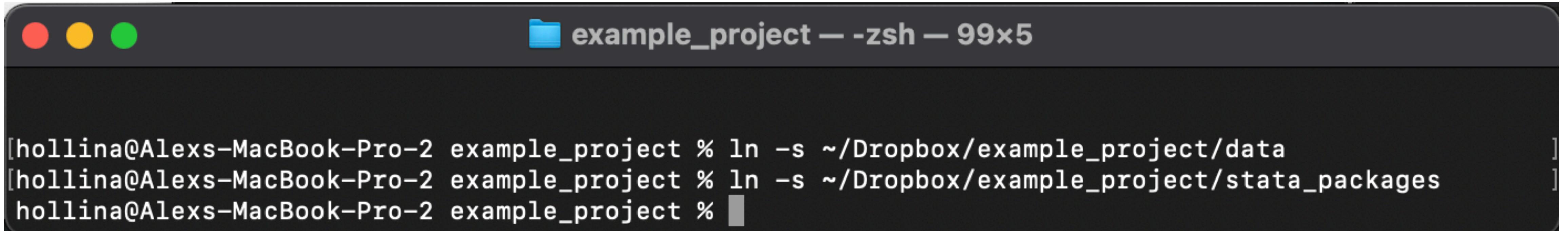
Second benefit: merge conflicts

```
40
Accept Current Change | Accept Incoming Change | Accept Both Changes | Compare Changes
41 <<<<< HEAD (Current Change)
42     \subcaption*\{ \emph{Note:} A very important note. Oh
43     no! A conflict. We edited the same text!}
44 =====
45
46     \subcaption*\{ \emph{Note:} A very important note that
47     now has a conflict!!.}
48
49 >>>> 780314ad7e8a3ea1a38f3cb69dea893a641c5c2a (Incoming Chang
50
51     \label{fig:simple_scatter}
52
53     \end{figure}
54
55     ; Coef Plot
```

Second benefit: merge conflicts

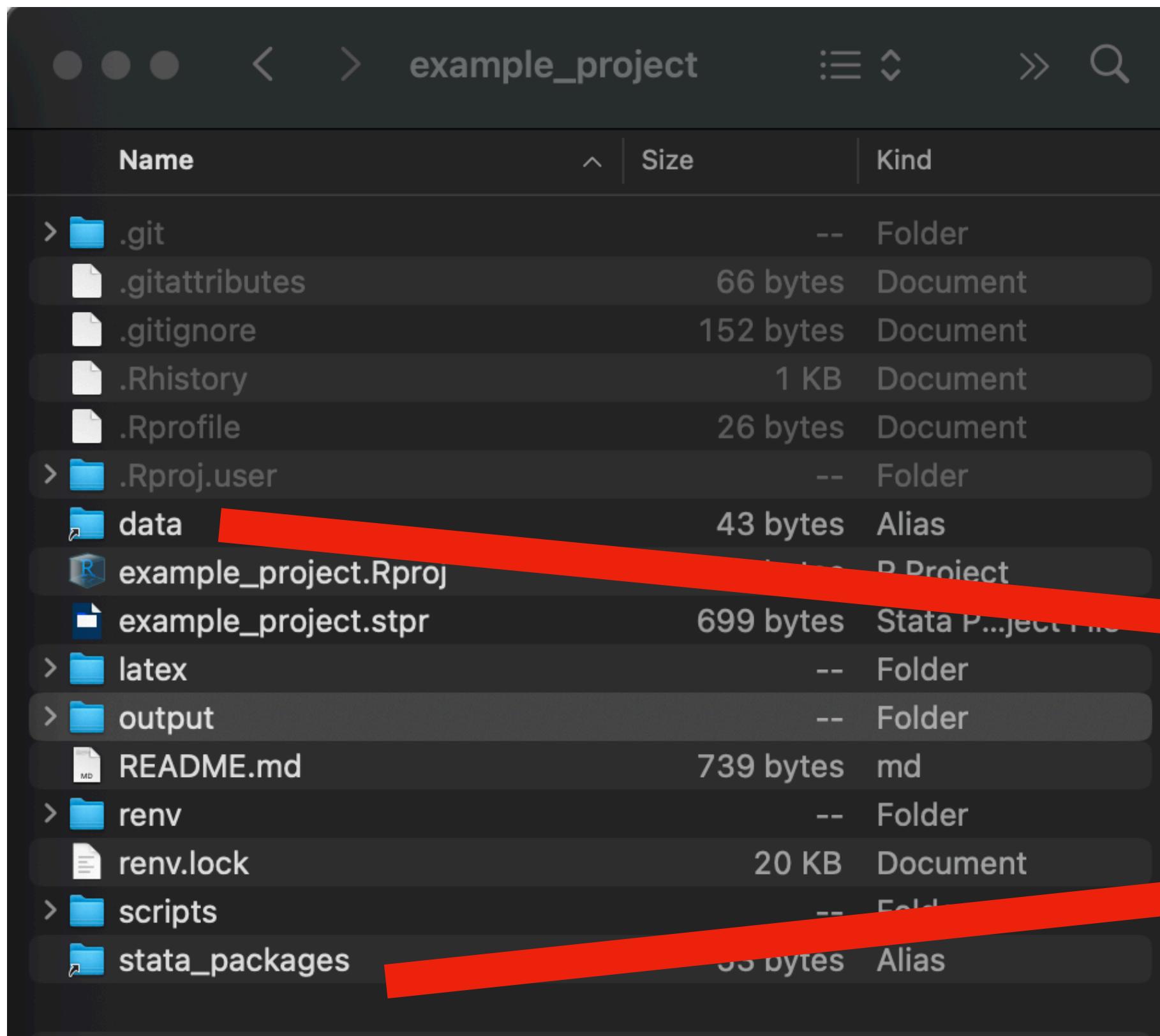


Symbolic links

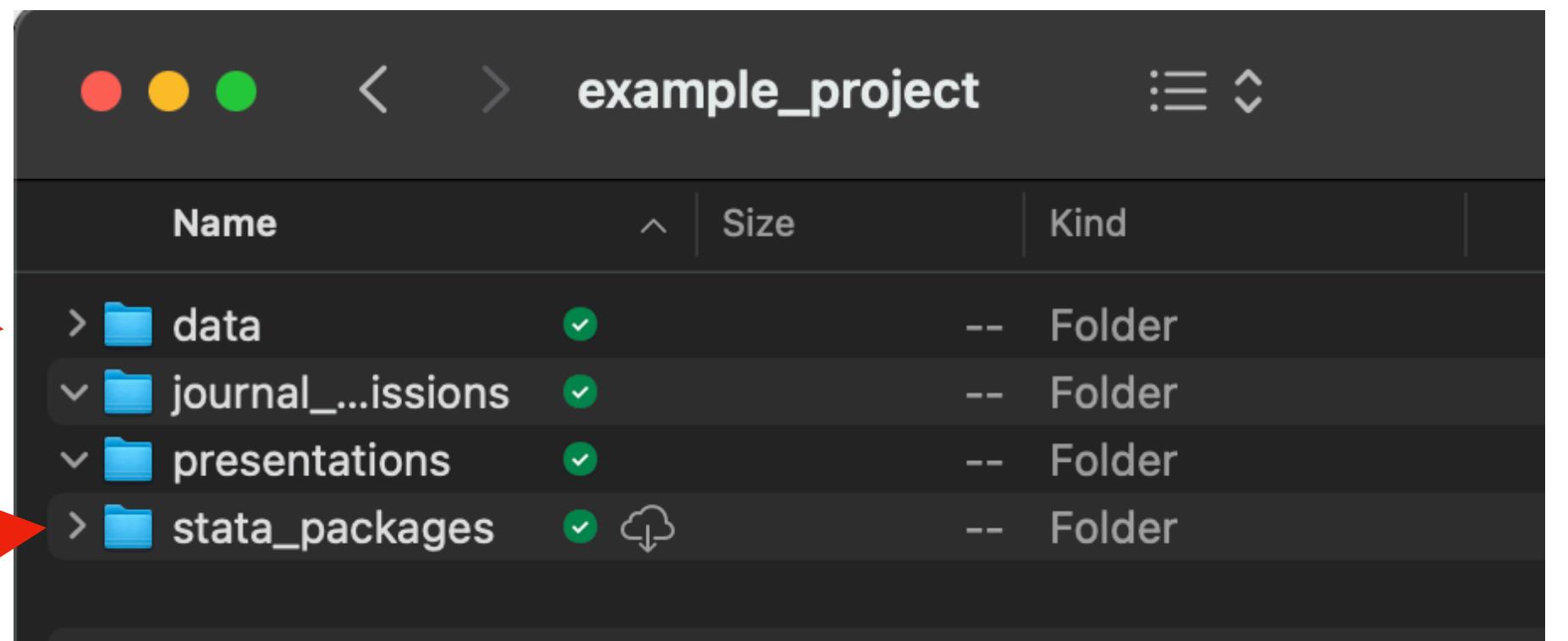


```
[hollina@Alexs-MacBook-Pro-2 example_project % ln -s ~/Dropbox/example_project/data  
[hollina@Alexs-MacBook-Pro-2 example_project % ln -s ~/Dropbox/example_project/stata_packages  
hollina@Alexs-MacBook-Pro-2 example_project %
```

- This is basically a shortcut from your Dropbox to your GitHub

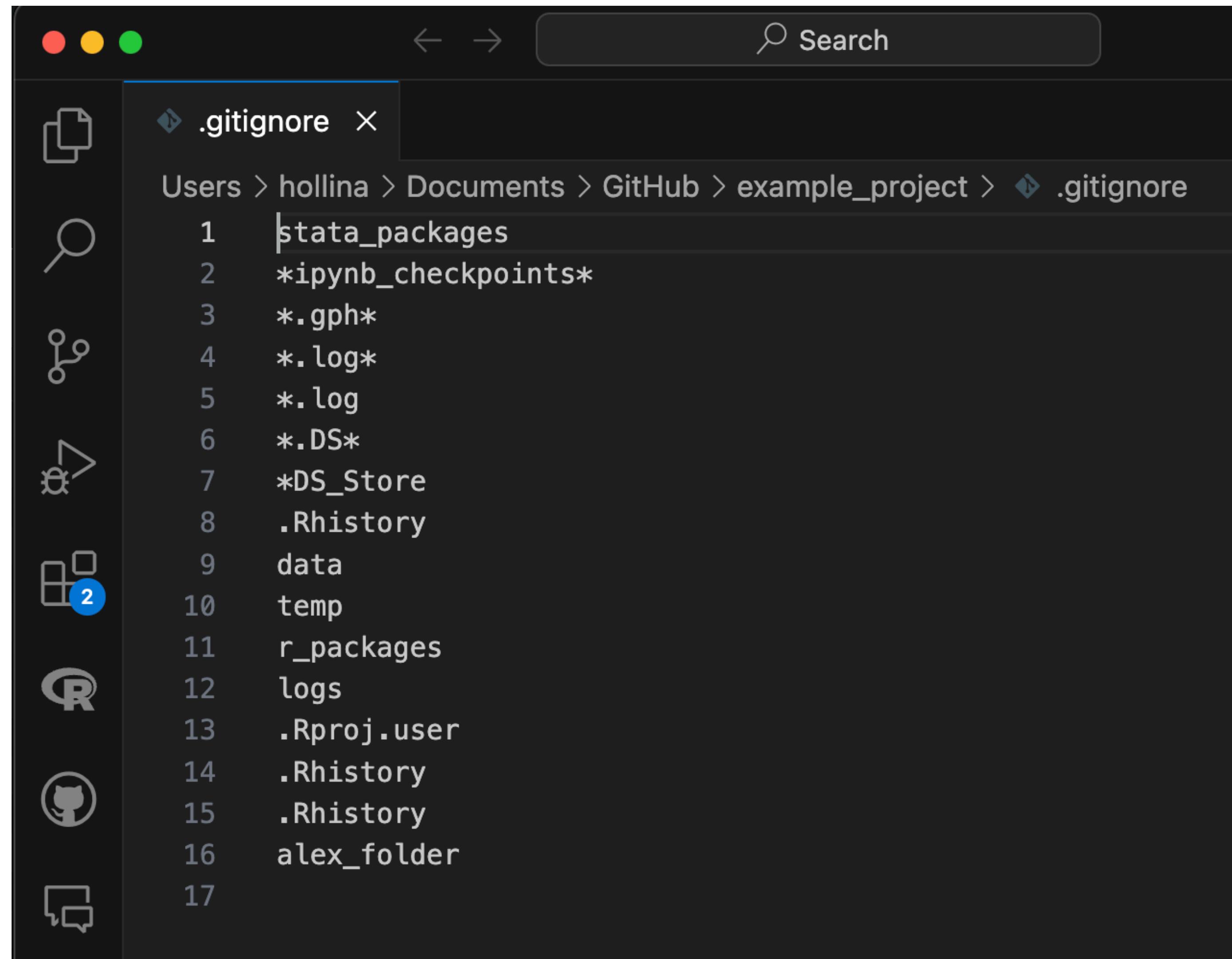


Dropbox for data/other things



GitIgnore

- List of files you don't want GitHub to “track”
 - If you use symbolic links they can still be backed up via DropBox

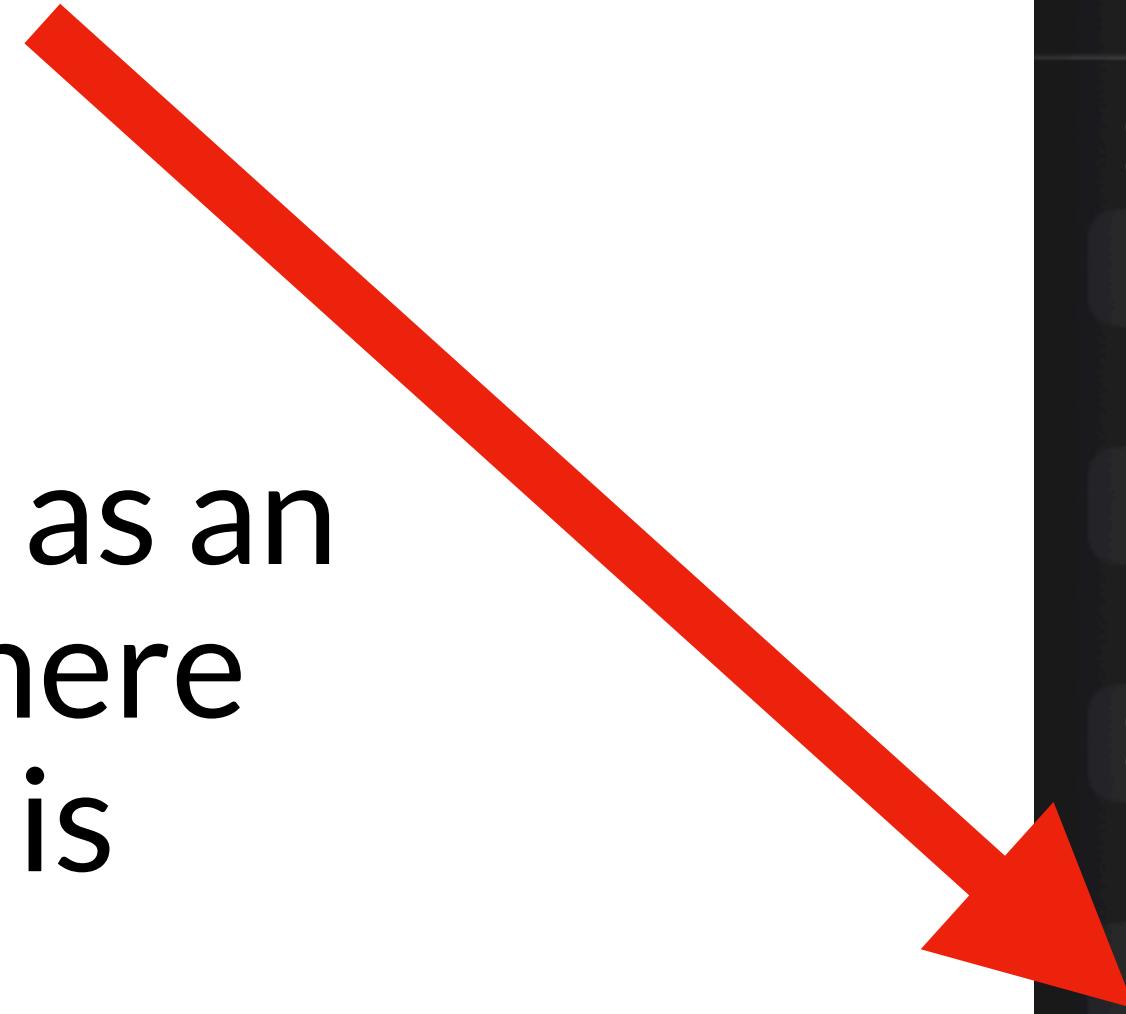


A screenshot of a dark-themed file explorer window. The title bar shows the file name ".gitignore". The path is listed as "Users > hollina > Documents > GitHub > example_project > .gitignore". The main pane displays a list of file patterns to ignore:

```
1 stata_packages
2 *ipynb_checkpoints*
3 *.gph*
4 *.log*
5 *.log
6 *.DS*
7 *DS_Store
8 .Rhistory
9 data
10 temp
11 r_packages
12 logs
13 .Rproj.user
14 .Rhistory
15 .Rhistory
16 alex_folder
17
```

Projects

- Think of this as an anchor to where your project is located
- Key benefit:
 - Can you relative file paths
 - No more hard-coding



Name	Size	Kind
> .git	--	Folder
.gitattributes	66 bytes	Document
.gitignore	152 bytes	Document
.Rhistory	1 KB	Document
.Rprofile	26 bytes	Document
> .Rproj.user	--	Folder
data	43 bytes	Alias
example_project.Rproj	205 bytes	R Project
example_project.stpr	699 bytes	Stata P...ject File
> latex	--	Folder
> output	--	Folder
README.md	739 bytes	md
> renv	--	Folder
renv.lock	20 KB	Document
> scripts	--	Folder
stata_packages	53 bytes	Alias

/Users/hollina/Documents/GitHub/example_project/output/important_table.tex

vs. output/important_table.tex

This really simplifies coding

A screenshot of the Stata software interface. The main window shows a do-file named "1.test_script.do" with the following content:

```
1 // Clear memory
2 clear all
3
4 // Open Auto Dataset
5 sysuse auto
6
7 /////////////////////////////////
8 // Summary Statistics
9 eststo clear
10
11 estpost summarize price mpg foreign headroom trunk weight length turn
12     esttab using "output/summary_statistics.tex" , replace ///
13         cells("mean(fmt(%20.2f) label(\multicolumn{1}{c}{Mean})) sd(fmt(%20.2f) \
14             label(\multicolumn{1}{c}{S.D.})) min(fmt(%20.2f) label(\multicolumn{1}{c}{Min.})) \
15             max(fmt(%20.2f) label(\multicolumn{1}{c}{Max.})) count(fmt(%3.0f) \
16             label(\multicolumn{1}{c}{N}))") ///
17             nomtitle nonum label f alignment(S S) booktabs nomtitles b(%20.2f) se(%20.2f) eqlabels( \
18             none) eform ///
19             noobs substitute(\_ _) ///
20             refcat(foreign "\emph{Mortality Sample}" pop "\hspace{0.5cm} \emph{All}" popw \
21             "\hspace{0.5cm} \emph{White}" popb "\hspace{0.5cm} \emph{Black}" poph "\hspace{0.5cm} \
22             \emph{Hispanic}" , nolabel)
```

The Project Manager on the right side of the interface lists "example_project" which contains "0.run_file.do" and "1.test_script.do". "1.test_script.do" is highlighted with a blue background.

And writing

The screenshot shows a LaTeX editor interface with two tabs: 'tables_and_figures.tex' and 'tables_and_figures.pdf'. The left pane displays the LaTeX source code, and the right pane shows the resulting PDF document.

LaTeX Source Code (tables_and_figures.tex):

```
72
73 \begin{table}[ht]
74 \centering
75 \begin{tblr}{}
76 \hline
77 \textbf{Tables} & \\
78 \hline
79 \end{tblr}
80 \FloatBarrier
81 \thispagestyle{empty}
82 \newpage
83 \begin{table}[ht]
84 \centering
85 \begin{tblr}{}
86 \hline
87 \textbf{Summary Statistics Table} & \\
88 \hline
89 \begin{threeparttable}
90 \caption{Summary Statistics}
91 \label{tab:summary_statistics}
92 \text{Mortality Sample}
93 \text{Car type}
94 \text{Headroom (in.)}
95 \text{Trunk space (cu. ft.)}
96 \text{Weight (lbs.)}
97 \text{Length (in.)}
98 \text{Turn Circle (ft.)}
99 \end{threeparttable}
100 \end{tblr}
101 \end{table}
102 \begin{table}[ht]
103 \centering
104 \begin{tblr}{}
105 \hline
106 \textbf{Mileage (mpg)} & \\
107 \hline
108 \end{tblr}
109 \end{table}
```

PDF Output (tables_and_figures.pdf):

Table 1. Summary Statistics

	Mean	S.D.	Min.	Max.	N
Price	6165.26	2949.50	3291.00	15906.00	74
Mileage (mpg)	21.30	5.79	12.00	41.00	74
<i>Mortality Sample</i>					
Car type	0.30	0.46	0.00	1.00	74
Headroom (in.)	2.99	0.85	1.50	5.00	74
Trunk space (cu. ft.)	13.76	4.28	5.00	23.00	74
Weight (lbs.)	3019.46	777.19	1760.00	4840.00	74
Length (in.)	187.93	22.27	142.00	233.00	74
Turn Circle (ft.)	39.65	4.40	31.00	51.00	74

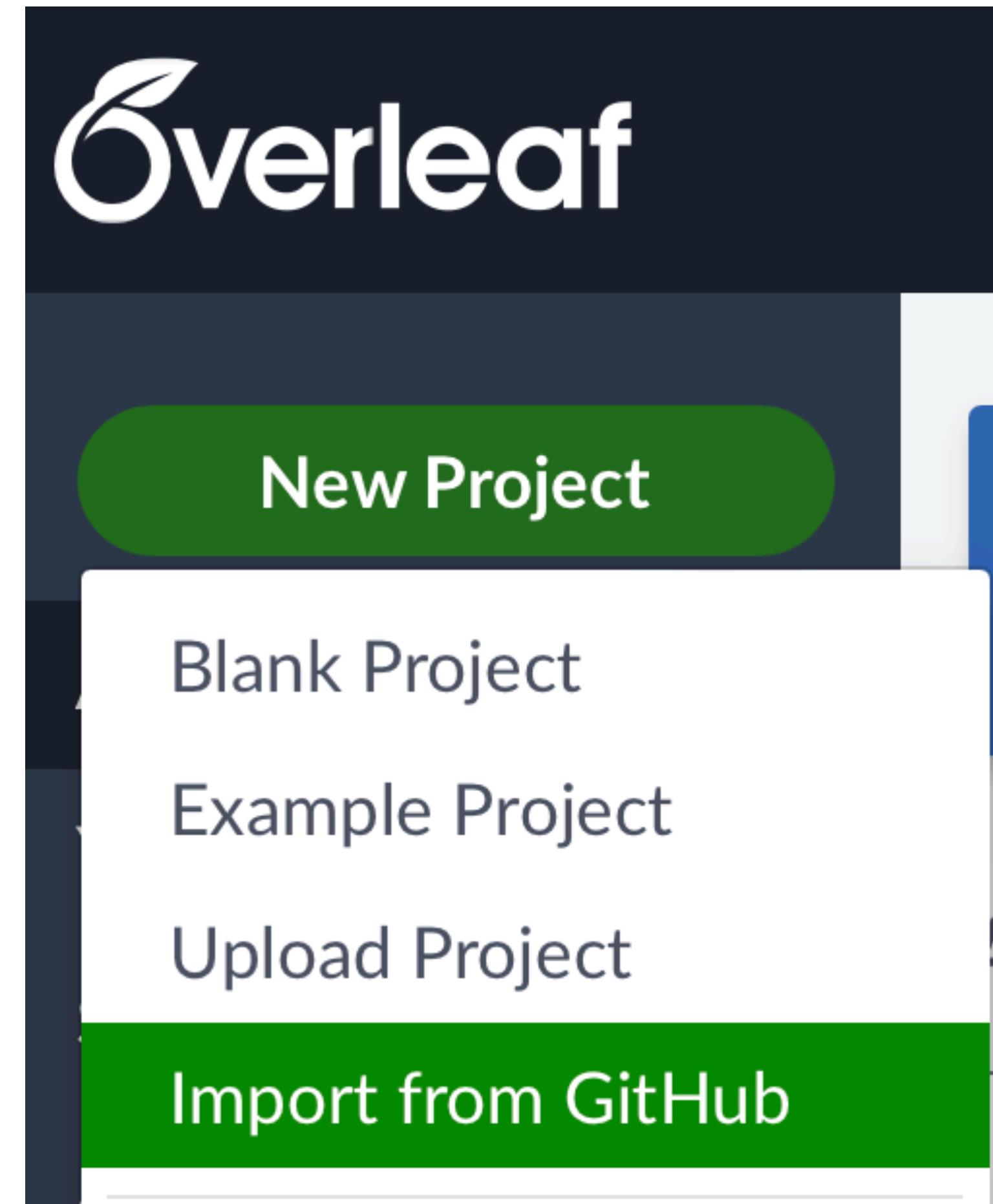
Source: Important Stuff

A red arrow points from the line '\label{tab:summary_statistics}' in the LaTeX code to the caption 'Table 1. Summary Statistics' in the PDF output.

Speaking of writing

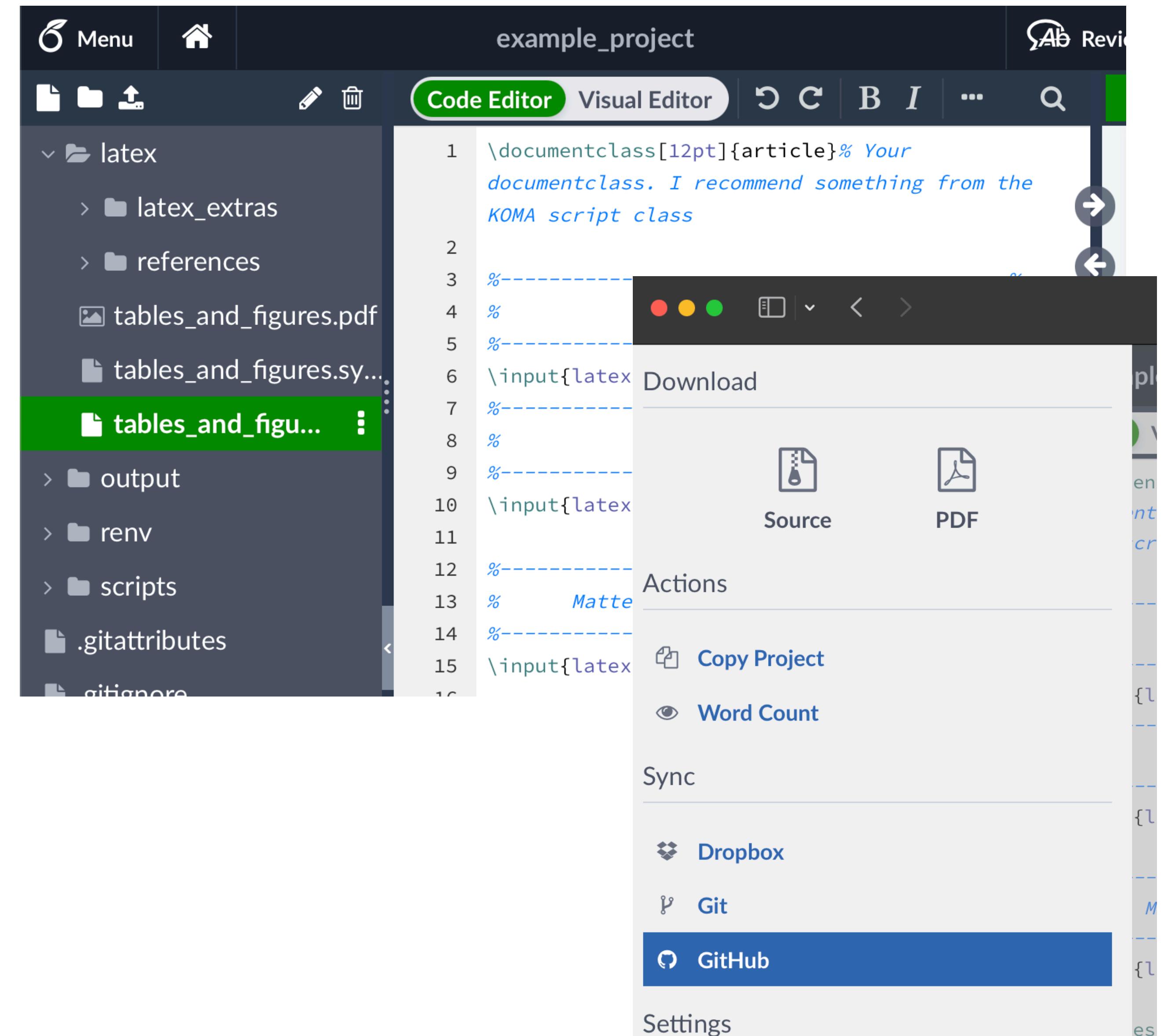
Integrating with overleaf is relatively pain free

- First, just import your GitHub project into overleaf



Integrating with overleaf is relatively pain free

- To update
 - click Menu
 - And then GitHub
 - And then write your message and click save
 - Key annoyance : It doesn't happen automatically



Packages are code too

Be sure to save the user-written packages

- Stata packages are small
 - Recommend just creating your own folder
- R packages can be large
 - `renv` is great
 - So is `groundhog`

```
37 ///////////////////////////////////////////////////////////////////
38 // Use included packages
39 cap adopath - PERSONAL
40 cap adopath - PLUS
41 cap adopath - SITE
42 cap adopath - OLDPLACE
43 adopath + "stata_packages"
44 net set ado "stata_packages"
45
46 // Download packages
47 if $downloads == 1 {
48     // Install Packages
49     ssc install estout, replace
50     ssc install blindschemes, replace
51     ssc install coefplot, replace
52 }
53
54
55
56
```

Quick example

What is is?

- Download this project here: https://github.com/hollina/example_project

Table 2. Regression Results

- Stata code
 - 0.run_file.do
 - Sets up project then runs
 - 1.test_script.do
 - 2.test_script.R
 - Makes a nice latex document

	(1)	(2)	(3)	(4)
Mileage (mpg)	-294.20*** (55.69)	-294.20*** (60.34)	-24.22 (71.71)	-24.22 (79.80)
Car type	1767.29** (700.16)	1767.29*** (607.74)	3212.15*** (685.58)	3212.15*** (668.88)
Headroom (in.)			-625.42* (373.69)	-625.42** (279.58)
Trunk space (cu. ft.)			60.94 (92.62)	60.94 (70.66)
Weight (lbs.)			5.93*** (1.02)	5.93*** (1.67)
Length (in.)			-74.98** (37.54)	-74.98 (51.16)
Turn Circle (ft.)			-157.74 (116.02)	-157.74 (119.89)
Mean of Dependent Variable	6165.26	6165.26	6165.26	6165.26
Observations	74	74	74	74
Adjusted R sq.	0.26	0.26	0.53	0.53
Year FE	No	No	No	No
County FE	No	No	No	No
Some Type of Linear Time Trend(s)	No	No	No	No
Robust Standard Errors	No	Yes	No	Yes

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Tools/Resources

- GitHub: <https://github.com/>
- GitHub Desktop: <https://desktop.github.com/>
- VSCode: <https://code.visualstudio.com/>
- GitHub CoPilot: <https://github.com/features/copilot>
- Julian Reif's coding guide: <https://julianreif.com/guide/>