Senior Thesis

Empirical Organization

Kyle Coombs (adapted from Tyler Ransom + Scott Cunningham)
Bates College | EC/DCS 456

Table of Contents

- Prolgue
- Literature
- Causation vs. Correlation
- Data
- How to start
- Collaboration

Prologue

What is a thesis?

- A thesis is a research project that you complete during your senior year.
- It is the culmination of your undergraduate education.
- It is arguably the most important thing you will do as an undergraduate
    - Other candidates include: meeting a spouse, persevering after failing a test, that one kegstand you did at the beginning of your junior year, etc.
- More importantly it is your chance to dig deep into an idea!

What will this class do?

- I am here to help you write the best thesis you can write
- I will not:
  - write your thesis for you.
  - write your code for you.
  - find your data
- I will provide scaffolding for you to build your thesis on.
- Specifically, I will
  - evaluate and sharpen your question
  - provide a useful structure for iteratively developing empirical work
  - suggest useful tools for data management and analysis
  - point you to literature you may have missed out, but is critical to your topic

What makes a strong thesis?

- **Originality**: A thesis should ask an original question. It should be a new idea, or a new way of looking at an old idea. It should not be a replication of an existing study. It should not be a summary of existing literature.

- **Importance**: A thesis should make a meaningful contribution to the literature. It should not be a trivial contribution, or a contribution that is only important to a small group of people. It should not be a contribution that is only important to you.

  - That does not mean you're inventing a new branch of economics

- **Feasibility**: A thesis should be feasible. It should not be a project that is too large to complete in the time you have available. It should not be a project that requires data that is not available. It should not be a project that requires skills that you do not have.

What makes a strong thesis? (cont.)

- **Organization**: A thesis should have a clear question, a clear methodology for answering that question, and a clear answer to that question. It should not be a collection of loosely related ideas. It should not be a collection of unrelated ideas.
- **Replicable**: A thesis should have an organized set of methods and code that can be replicated by someone else.

Literature

Literature

- How the heck do I do that?

- Well you'll want to check out the research literature on your topic

- Not just any literature though, you should be looking at academic articles published in peer-reviewed journals:

    - American Economic Review
    - Quarterly Journal of Economics
    - Journal of Political Economy
    - Econometrica
    - Review of Economic Studies
    - Review of Economics and Statistics
    - American Economic Journal: Applied Economics OR Economic Policy
    - Journal of Human Resources
    - Journal of Public Economics
    - National Bureau of Economic Research (NBER) Working Papers

- There are many more, but these journals are some of the most prestigious in economics

- The NBER is a working paper series that is often a precursor to publication in one of the journals listed above

    - I recommend you sign up for their email list and select articles specific to your topic

How do I engage with the literature?

- The biggest mistake you can make is to try to read any academic paper from start to finish

- Academic papers are often sprawling and dense
    - They are chock full of robustness checks and extensions that you don't need to read
    - Often these extra bits will make the paper more publishable, but do not advance the main point
    - (Sometimes the extra bits are hiding some critical new ideas or flaws)

- Instead, you should prioritize learning some fast facts about several papers and then decides which ones you invest in more deeply

- Fast facts:
    - What is the question?
    - What methodology do they use to answer it?
    - What's the intuition behind any identification strategies?
    - What data do they use?
    - What are the main findings?
    - Any limitations you have in mind?

- Most of this can be learned from the abstract and introduction followed by a skim of the rest of the paper

What should this look like?

- In practice, your thesis will likely ask an empirical question, which you will try to answer using data
- You'll need to:
  - Draw on the models you learned in your micro and macro courses to think through a conceptual framework
  - Draw on the descriptive and data wrangling skills you learned in econometrics and statistics
  - Draw on the empirical methods you learned in econometrics
  - Draw on the writing skills you learned in your writing-intensive courses
  - Draw on the economic topics you explored in your electives
- You will also need to learn a variety of new skills on the fly

Causation vs. Correlation

Correlation (or prediction) vs. causation

Most tasks in empirical economics boil down to one of two goals:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

1. **Prediction:** Accurately and dependably predict/forecast y using on some set of explanatory variables—doesn't need to be x_1 through x_k. Focuses on $\hat{y}$. β_j doesn't really matter.

2. **Causal estimation:**[†] Estimate the actual data-generating process—learning about the true, population model that explains how y changes when we change x_j—focuses on β_j. Accuracy of $\hat{y}$ is not important.

For the next few weeks, we will focus on **causally estimating** β_j.

[†] Often called *causal identification*.

The challenges

As you saw in the data-analysis exercise, determining and estimating the true model can be pretty difficult—both practically and econometrically.

Practical challenges

- Which variables?
- Which functional form(s)?
- Do data exist? How much?
- Is the sample representative?

Econometric challenges

- Omitted-variable bias
- Reverse causality
- Measurement error
- How precise can/must we be?

Many of these challenges relate to **exogeneity**, *i.e.*, $\boldsymbol{E}[u_i|X] = 0$.
Causality requires us to **hold all else constant** (*ceterus paribus*).

It's complicated

Occasionally, ***causal*** relationships are simply/easily understood, *e.g.*,

- What caused the forest fire?
- How did this baby get here?

Generally, ***causal*** relationships are complex and challenging to answer, *e.g.*,

- What causes some countries to grow and others to decline?
- What caused the capital riot?
- Did lax regulation cause Texas's recent energy problems?
- How does the number of police officers affect crime?
- What is the effect of better air quality on test scores?
- Do longer prison sentences decrease crime?
- How did cannabis legalization affect mental health/opioid addiction?

Correlation ≠ Causation

You've likely heard the saying

> Correlation is not causation.

The saying is just pointing out that there are violations of exogeneity.

Although correlation is not causation, **causation *requires* correlation**.

New saying:

> Correlation plus exogeneity is causation.

Data

Ideal Data

- Most reseearch questions are driven by data
- But often the data you want is not available
- So you have to work with the data that are available
- It is useful to imagine your ideal data set to benchmark the data you have
    - What variables would you have?
    - What time period would you cover?
    - How large is the sample?
    - How many observations per unit of analysis?
- You can then compare your ideal data to the data you have
    - What variables do you have?
    - What time period do you cover?
    - How large is the sample?
    - How many observations per unit of analysis?
- What problems do the deviations between the "ideal" and "actual" data cause?
- What assumptions do you need to make to use the "actual" data?
- What assumptions would you need to make even if you had the "ideal" data?

What if I don't want to use data?

- Get out of my classroom
- Just kidding
- It is possible that you will work on a theory-based thesis
- This will be harder than an empirical thesis because economic theory is based on some pretty complicated math
- Nevertheless, even a theory-based thesis should have a clear question, a clear methodology for answering that question, and a clear answer to that question.
    - It will also benefit from a smaller, even simulation-driven empirical component

How to start

What should I do first?

- Make a folder for your thesis project on your computer
- It should have a logical subfolder structure
    - Data
    - Code
    - Figures
    - Tables
    - Output
    - References
    - etc.
- Why?
- Because you will be working on this project for a long time, and you will need to be able to find things
- Also, other people (like me) will want to look at your project, and they (we) will need to be able to find things

Version Control with GitHub

- You should use GitHub to manage your thesis project
- This will be a new skill for some of you, but it is a skill that is in high demand in the job market
- It will also mean that you have an online backup of your thesis project that you can share with future employers, etc.
- It is also a great way to collaborate with other people, but it is a little... clunky at first
- Several people in this room actively used GitHub last semester, so ask each other for help
- Basic vocabulary:
  - Repository: A history of changes to a set of files stored in a single "folder"-like structure
  - Commit: A snapshot of the state of a repository at a particular point in time
  - Branch: A parallel version of a repository that can be edited without affecting the original version
  - Push/Pull: To send/receive changes to/from a remote repository
  - Clone: Copy a repository from a remote location (usually a webpage) to your local computer
  - Fork: Make a copy of repository that is hosted on your own GitHub account, so you can make changes without affecting the original version
  - Pull Request: Suggest changes to other repositories that you do not own

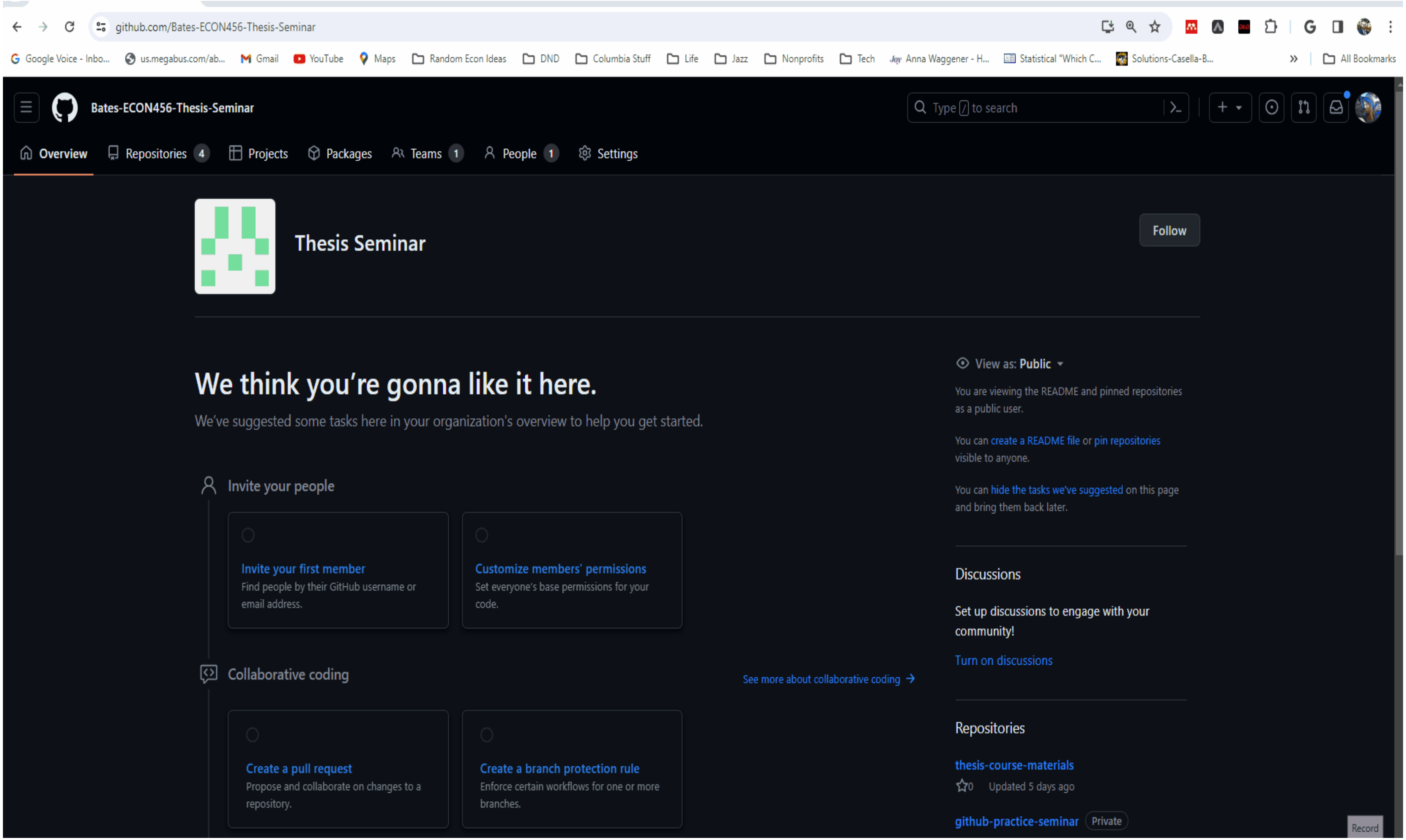
Why GitHub Repository?

- Imagine you want to rewrite some code in your thesis project to make it more efficient
- It will take several days to finalize
    - Do you make a new file named "thesis_code_v2.R" and save it in the same folder as "thesis_code_v1.R"?
- Then midway through you again realize that you need to make some other changes to the original code that have nothing to do with the efficiency changes
    - Do you make "thesis_code_v2_v2.R" and "thesis_code_v2.R"?
- But what about Dropbox? Or Google Drive? These backup file histories!
- Yes, but they are not designed for code -- they just save the entire file every time you make a change
- GitHub is designed for code, so you tell it what changes to save and when
    - When you have code that works, you "commit" it to the repository with a comment
    - Then you "push" it
    - If you need to undo something, you can "revert" to a previous commit

GitHub Templates

- One great use of GitHub is to create a template repositories to share with others
- Guess what? I made one for you!
- You can use this template to create your own thesis repository
  - Click `Use this template` in the upper right corner
  - Select `Create a new repository`
  - Give your repository a name
  - Make it private
  - Go to `Settings` > `Collaborators` > `Add people` and add me (`kgcsport`) as a collaborator
- I want you to have a private repository so you can share your code with me without sharing it with the world

Copy Template Gif



Thesis Seminar

We think you're gonna like it here.

We've suggested some tasks here in your organization's overview to help you get started.

Invite your people

○ Invite your first member
Find people by their GitHub username or email address.

○ Customize members' permissions
Set everyone's base permissions for your code.

Collaborative coding

See more about collaborative coding →

○ Create a pull request
Propose and collaborate on changes to a repository.

○ Create a branch protection rule
Enforce certain workflows for one or more branches.

You are viewing the README and pinned repositories as a public user.

You can create a README file or pin repositories visible to anyone.

You can hide the tasks we've suggested on this page and bring them back later.

Discussions

Set up discussions to engage with your community!

Turn on discussions

Repositories

thesis-course-materials
☆0 Updated 5 days ago

github-practice-seminar Private

Clone Repository

- Create it and clone it to your computer by next class
- I recommend you use GitHub Desktop if you're starting out
- Follow the instructions

  1. Click `Code` > `Open with GitHub Desktop`
  2. Click `Choose` and select the folder you created for your thesis project

- If you're more advanced, you can use the command line
- Or you can use RStudio with an SSH-key

Wait I already have a thesis repository!

- Great, you can turn that one into a GitHub repository using GitHub Desktop
  - **Note**: Step 1 involves using the command line in case this folder was already a repository, you can skip to Step 2 if that isn't relevant

Note: Lots of technical documentation in empirical work tries to be "catch-all" and cover all possible scenarios. This is a good example of that. Read directions fully before starting anything, but don't be afraid to skip steps that don't apply to you.

Collaboration

Lean on each other

- That last bullet point brings me to my next point:
- While your thesis is solo-authored, your work in this class should be collaborative
- You should be helping each other with your projects
- I'll build in formal and informal opportunities for you to do this

Example: Share research ideas

- Everyone go around and share your two research questions with the group
- It does not matter how preliminary they are
- Pair up with the person next to you and discuss your research questions for 10 minutes
- Please offer feedback:

1. How can the question be more specific?
2. Does it sound like material you learned about in one of your classes?
3. Any literature from those classes you think is relevant?
4. What ideal data do they need for this question?
5. Is the question causal or descriptive?

- Now that you've refined your research question, share it with the group

Next steps:

Next class:

- Create project repository on GitHub and clone to your computer
- Read through a previous economics thesis report from the literature folder and write up a Previous thesis report
- Check out the other readings on the syllabus

For January 25th

- Start working on two detailed question proposals

Next lecture: What makes research "good"?