FDR Pre-Doctoral Training Curriculum[1]

Academic Year 2021–2022

This version 2021-11-11

Introduction

This 10-week training program is designed to prepare incoming pre-doctoral research fellows at the Princeton Empirical Studies of Conflict (ESOC) lab with the skills needed to support faculty research projects within ESOC, SPIA, and associated departments. The program will draw from online courses for core data processing and visualization skills, research training materials from partnering organizations, and materials prepared by the team at ESOC.

The goal is to expose FDR fellows to all aspects of the data-driven research process. We will touch upon topics such as best practices for data management, research methodologies used in the social sciences, and production-related skills like optimizing code for publication and working with LaTeX. Ultimately, these skills will provide fellows with a strong analytical foundation for careers in public service and future PhD study in Political Science, Economics and related fields.

The syllabus should be viewed as a resource for the long run. The goal is not to have all the answers in 10 weeks, it's to have confidence that you can implement and know where to go to find answers.

Bootcamp Details

All fellows should already have access to the "FDR Predoctoral Training" Dropbox (DB) folder that hosts all relevant course materials (including instructions for the three data

[1]This curriculum was developed by Alicia Chen and Samikshya Siwakoti with guidance from Jacob N. Shapiro and feedback from many colleagues. Gigs Banga provided invaluable beta testing and Yining Sun and Nilima Pisharody completed post-testing revisions. The first cohort of FDR fellow, Chris Buckley and Hanjatiana Nirina Randrianarisoa, provided excellent feedback and made many changes from their experience with the curriculum. Throughout, we draw on others' outstanding resources, and owe a particular debt to Scott Cunningham's *Causal Inference: The Mixtape*. Developing this curriculum was a an interdisciplinary team effort.

exercises you'll work on in the last weeks of the bootcamp).[2] You should start by reading the ESOC Research Production Guide handout to learn about the lab's research practices. Each fellow also has their own subfolder and you should set your working directory to your designated folder. This also allows you to start thinking about best practices when it comes to project and data management discussed in the handout, which we will expand upon in the last week.

Over the course of the bootcamp, we recommend placing all files in an online storage service and importing all materials from the curriculum into said-cloud storage. Each user should have their own subfolder and set working directories to their designated folder.

Users from other institutions should seek their organizations' equivalent to the *Research Production Guide* to learn their specific standards and practices. If users are not affiliated with any organizations, we recommend looking for good research practices resources online, including those listed later in this document, or consult ESOC's own Research Production Guide.

The majority of this program is designed to be completed asynchronously using online resources and data exercises we have created. However, we hope that users will interact and work together to better understand the course materials, when and if possible.

Prerequisites

Though the course is designed in part to develop your programming skills. Basic familiarity with languages such as STATA, R, or Python is highly desirable. If you are more versed in coding with Stata, we recommend familiarizing yourself with and/ or reviewing both Python and R resources before starting the course. If you are more versed in R or Python, we recommend going over the other language. We also recommend that you review basic functions and common packages used in your program of choice prior to the start of this bootcamp. Some useful resources to review are:

- `STATA`. There are many outstanding set of introductory modules for STATA online:
  - Data Carpentry, "Economics Lesson with STATA", `https://datacarpentry.org/stata-economics/`
  - Oscar Torres-Reyna, "Getting Started in Data Analysis using STATA", `https://www.princeton.edu/~otorres/StataTutorial.pdf`

[2] We encourage organizations using this curriculum to set up analogous folder structures.

- Princeton DDS, "Online STATA tutorial", https://www.princeton.edu/~otorres/Stata/
- Alexander C. Lembcke, "Introduction to STATA", https://personal.lse.ac.uk/lembcke/ecStata/2010/MResStataNotesOct2010PartA.pdf
- UCLA's Statistical Consulting group STATA page, https://stats.idre.ucla.edu/stata/modules/

- `R`. We highly recommend going through Datacamp's Data Scientist track with R, at least finishing the Introduction and Visualization in R courses:
  - Data Camp, "Data Scientist with R", https://app.datacamp.com/learn/career-tracks/data-scientist-with-r
  - Simon Ejdemyr, "R Tutorials: Basics," https://sejdemyr.github.io/r-tutorials/basics/
  - Hadley Wickham and Garrett Grolemund's *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*: https://r4ds.had.co.nz/
  - Getting started with R Markdown: https://www.rstudio.com/resources/webinars/getting-started-with-r-markdown/
- `Python`. We highly recommend going through the Datacamp's Data Scientist track with Python, at least finishing the introductory and visualization courses:
  - Data Camp, "Data Scientist with Python" https://app.datacamp.com/learn/career-tracks/data-scientist-with-python?version=5
  - NYU Data Bootcamp, https://nyudatabootcamp.gitbook.io/thebook/
  - "Introduction to Python for Science": https://physics.nyu.edu/pine/pymanual/html/pymanMaster.html
  - Getting started with Jupyter notebooks: https://www.dataquest.io/blog/jupyter-notebook-tutorial/

We will be focusing heavily on the Potential Outcomes framework for Causal inference in this bootcamp. As such we also recommend that prior to the start of bootcamp, you read the first chapter of *Causal Inference: The Mixtape* by Scott Cunningham and review basic math and statistics concepts provided in the resources below:

- Scott Cunningham, *Causal Inference: The Mixtape, Chapter 1: Introduction*, https://mixtape.scunning.com/

- Khan Academy, "Statistics and Probability", https://www.khanacademy.org/math/statistics-probability
- Probability cheatsheets: http://www.wzchen.com/probability-cheatsheet/

Finally, here are Princeton University resources you could explore if needed along with other additional resources you could review prior to the bootcamp

University Resources:

- Princeton Data and Statistical Services (DSS) offers data and statistical consulting: https://dss.princeton.edu/
- Initiative for Data-Driven Social Science (DDSS) offers research consultations: https://ddss.princeton.edu/resources-services/our-services
- Princeton Research Computing Help Sessions are great for dealing with technical issues that come up with using HPC clusters: https://researchcomputing.princeton.edu/support/help-sessions

Course Outline

This schedule is tentative and subject to change. The learning goals outlined are key concepts you should grasp and can start to implement in practice by the end of the week.

In some weeks, your assignment will be a week-long course that includes both readings and exercises from an online resource. In other weeks, we have created assignments that require you to submit something to us. For those weeks, make sure to read the required readings first before completing the exercise as they are designed to implement concepts covered in the readings. There are also additional resources listed for each week that are not required but are there for reference.

Note on the Capstone Project: We recommend users to begin thinking about potential research ideas and questions within the first few weeks to minimize any complications such data unavailability during the final weeks of the training.
For example: brainstorm potential topics within research interests in week 1 and 2, outline research questions for two or three topics in week 3 and 4, and check data availability to develop research project in following weeks.

Week 01: Data Cleaning, Descriptive Statistics, and Visualizations

Learning Objective:

- Learn/review effective data cleaning and management techniques.
- Understand fundamental concepts in statistics in order to describe data (e.g., distributions, central tendency theory, bi/multivariate data, etc.).
- Implement descriptive analysis in R/STATA/Python (e.g., learn to extract key summary information from data, etc.).
- Understand how and why different visualisation tools are used in descriptive data analysis.

Assignment: Instructions are in the "Week 1 Assignment" folder on Dropbox.

Readings:

(a) Data cleaning:

- `STATA users`: World Bank's Development Impact Evaluation (DIME):
  - Tidying Data: https://osf.io/eznjf/
  - Data Cleaning: https://osf.io/q54uy/
  - Data Construction (includes constructing panel data): https://osf.io/k4tr6/
- `STATA users`: J-PAL, "Data cleaning and management", https://www.povertyactionlab.org/resource/data-cleaning-and-management
- "Missing-data imputation", http://www.stat.columbia.edu/~gelman/arm/missing.pdf
- Jason Osborne, "Notes on the use of data transformations", *Practical Assessment, Research, and Evaluation* 8(8):2002, https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1115&context=pare
- "Statistical Data Types" chapter of Andrius Buteikis, *Practical Econometrics and Data Science*, http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/1-2-data-types.html

(b) Descriptive analysis:

- World Bank DIME, Descriptive Analysis in R (https://osf.io/7jbfg/) and in STATA (https://osf.io/6be2t/)

(c) Visualizing data

- J-PAL, "Data Visualization", https://www.povertyactionlab.org/resource/data-visualization
- United Nations' guide to making data meaningful, https://ec.europa.eu/eurostat/documents/64157/4374310/33-UNECE-making-data-meaningf pdf/d5b954e2-b110-469b-a2b5-78aa7c12ab62
- R Graph Gallery (useful for getting ideas and example code in R), https://https://www.r-graph-gallery.com/index.html

Additional Resources:

- Claus O. Wilke, *Fundamentals of Data Visualization*, https://clauswilke.com/dataviz/
- Kieran Healy, *Data Visualization: A Practical Introduction*: https://socviz.co/
- Additional resources for graphics
  - https://www.stata.com/support/faqs/graphics/gph/stata-graphs/
  - https://www.stata.com/features/overview/example-graphs/
  - https://www.data-to-viz.com/

Week 02: Probability & Regressions

Learning Objective:

- Understand the basics of linear regression and discrete choice models, and learn to run specific kind of regression models in R/STATA/Python.
- Understand what a DAG is and be able to construct one.
- Be able to calculate and interpret coefficients and different standard errors in regression models.
- Be prepared to run and plot interaction terms in linear and discrete choice models.

Assignment:

- Complete Data Camp's "Introduction to Statistics" course in R https://learn.datacamp.com/courses/introduction-to-statistics-in-r or Python https://learn.datacamp.com/courses/introduction-to-statistics-in-python. Email the certificate of completion to ESOC Research Specialist copying Prof. Shapiro.
- From Cunningham's book: any data exercises from the assigned readings & "Probability and Statistics" exercises here: https://mixtape.scunning.com/teaching-resources.html

Readings:

- Chapter 2 and 3 of Scott Cunningham's *Causal Inference: The Mixtape*, https://mixtape.scunning.com/probability-and-regression.html
- "Binary Choice Models", https://faculty.utrgv.edu/diego.escobari/teaching/Econ3341/Handouts/Chapter09.pdf
- UCLA Statistical Consulting Group's introduction to interaction terms *Decomposing, Probing, and Plotting Interactions in R/Stata*, https://stats.idre.ucla.edu/r/seminars/interactions-r/#s3c

Additional Resources:

- Probability cheatsheets: http://www.wzchen.com/probability-cheatsheet/

- Khan Academy, "Statistics and Probability", https://www.khanacademy.org/math/statistics-probability
- J-PAL MicroMasters' Data Analysis for Social Scientists course: https://www.edx.org/course/data-analysis-for-social-scientists
- Konrad Menzel's Introduction to Statistical Methods in Economics course: https://ocw.mit.edu/courses/economics/14-30-introduction-to-statistical-methods-in-economics-sp index.htm

Week 03: Causal Inference

Learning Objective:

- Understand potential outcomes framework, develop familiarity with one way of approximating the ideal experiment (e.g. making $[X'X]^{-1}X'\epsilon = 0$).
- Be able to identify and utilize methods for estimating causal effects using matching and subclassification.
- Think about potential research designs to improve causal inference in your capstone project.

Assignment: From Cunningham's book: any data exercises from the assigned readings & the chapter-specific exercises from: https://mixtape.scunning.com/potential-outcomes.html

Readings: Chapters 4 and 5 of Scott Cunningham's *Causal Inference: The Mixtape*, https://mixtape.scunning.com/potential-outcomes.html

Additional Resources:

- Josh Angrists and Victor Chernozhukov, "Applied Econometrics: Mostly Harmless Big Data" (companion course to *Mostly Harmless Econometric*): https://ocw.mit.edu/courses/economics/14-387-applied-econometrics-mostly-harmless-big-data-fall-2014/index.htm
- MIT 14.771/ Harvard 2390b "Empirical Methods" handout: http://web.mit.edu/14.771/www/emp_handout.pdf

Week 04: Operationalizing Regressions pt.1

Learning Objective:

- To be able to identify and utilize methods for estimating causal effects: regression discontinuity, instrumental variables
- Think about potential research designs to improve causal inference in your capstone project.

Assignment: From Cunningham's book: any data exercises from the assigned readings & any other exercises from: https://mixtape.scunning.com/teaching-resources.html

Readings:

- Chapters 6 and 7 of Scott Cunningham's *Causal Inference: The Mixtape*, https://mixtape.scunning.com/regression-discontinuity.html

Week 05: Operationalizing Regressions pt.2

Learning Objective:

- Be able to understand Panel Data structure and why fixed effects are used in panel data
- Be able to understand when and why a Difference-in-Differences model is used
- Think about potential research designs to improve causal inference in your capstone project.

Assignment: From Cunningham's book: any data exercises from the assigned readings any other exercises from: https://mixtape.scunning.com/teaching-resources.html

Readings:

- Chapters 8 and 9 of Scott Cunningham's *Causal Inference: The Mixtape*, https://mixtape.scunning.com/panel-data.html
- Josh Blumenstock's lecture notes on fixed effects models: http://www.jblumenstock.com/files/courses/econ174/FEModels.pdf

Week 06: Operationalizing Regressions pt.3 & Panel Data Exercise #2

Learning Objective:

- Be prepared to work with clustered standard error.
- Understand the basics of power analysis and identify what factors could affect statistical power.
- Be able to interpret interaction terms in logit and probit models.
- Practice panel data skills.

Readings:

- On non-standard errors:
  - Summary of Athey, Abadie, Imbens and Wooldridge (2017): `https://blogs.worldbank.org/impactevaluations/when-should-you-cluster-standard-errors-new-wisdom-`
  - Nonstandard errors in code: `https://lost-stats.github.io/Model_Estimation/Statistical_Inference/Nonstandard_Errors/nonstandard_errors.html`
- Intro to power analysis: `https://stats.idre.ucla.edu/other/mult-pkg/seminars/intro-power/`
- Thomas Brambor, William Roberts Clark and Matt Golder, "Understanding Interaction Models: Improving Empirical Analyses", *Political Analysis* 14(1):2006.
- Chunrong Ai and Edward C. Norton, "Interaction terms in logit and probit models", *Economics Letters* 80(1): 2003.

Assignment: Instructions for Panel data exercise 2 are in the "Panel Data Exercise" folder on Dropbox. This exercise will implement many of the tools and methods we covered in earlier weeks. You should feel free to review those notes as you complete this data challenge, in particular the assigned readings for the empirical methods and operationalizing regressions weeks. As you do the exercise, remember to think about substantive effects, not just statistical significance

Week 07: Text-as-Data & Panel Data Exercise #1

Learning Objective:

- Build basic skills for Natural language Processing
- Learn to clean and process unstructured text data
- Learn to create document term matrices
- Explore dictionary and topic modeling methods for text analysis
- Solidify understanding of Panel Data regressions

Assignment:

- Instructions are in the "Text-as-Data Exercise" folder on Dropbox
- Instructions for Panel data exercise 1 are in the "Panel Data Exercise" folder on Dropbox.

We encourage you to use Python to handle Natural Language Processing. While there are many programs out there, Python has an impressive library of packages for working with text data and conducting Natural Language Processing commonly used in the social sciences. You are welcome to find your own set of resources if you prefer using a separate program (eg. R, Stata, etc.) for this section.

Readings:

- Data Camp skill track, "Natural Language Processing in Python":
  - We recommend completing the courses Introduction to NLP in Python, Sentiment Analysis in Python, and Advanced NLP with spaCy. Additional courses in skill track are optional. https://app.datacamp.com/learn/skill-tracks/natural-language-processing-in-python
- Learning `regex`: https://regexone.com/
- `regex` cheatsheet: https://github.com/justingrimmer/tad_19/blob/master/regex.pdf

- David M. Blei, “Surveying a suite of algorithms that offer a solution to managing large document archives,” *Communications of the ACM* 55(4):2012, http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf

Additional NLP Resources:

- Chapter 17 (Collecting Data from the Web) of Rochelle Terman’s PLSC 31101: Computational Tools for Social Science course notes: https://plsc-31101.github.io/course/collecting-data-from-the-web.html#web-apis (uses R)
- Ryan Mitchell, *Web Scraping with Python: Collecting Data from the Modern Web.*

Robustness checks Resources:

- Abadie, A. (2005): “Semiparametric difference-in-differences estimators” The Review of Economic Studies, 72, 1–19.
- Belloni, A., V. Chernozhukov, and C. Hansen (2014): “High-dimensional methods and inference on structural and treatment effects” Journal of Economic Perspectives, 28, 29–50.
- Borusyak, K., X. Jaravel, and J. Spiess (2021): “Revisiting event study designs: Robust and efficient estimation” Tech. rep., Working Paper.
- Callaway, B. and P. H. Sant’Anna (2020): “Difference-in-differences with multiple time periods” Journal of Econometrics.
- Marcus, M. and P. H. Sant’Anna (2021): “The role of parallel trends in event study settings: An application to environmental economics” Journal of the Association of Environmental and Resource Economists, 8, 235–275.
- Roth, J. (2021): “Pre-test with caution: Event-study estimates after testing for parallel trends” Working paper.
- Sant’Anna, P. H. and J. Zhao (2020): “Doubly robust difference-in-differences estimators” Journal of Econometrics, 219, 101–122.
- Wooldridge, J. (2021): “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators” Available at SSRN 3906345.
- De Chaisemartin, C. and X. d’Haultfoeuille (2020): “Two-way fixed effects estimators with heterogeneous treatment effects” American Economic Review, 110, 2964–96.

Week 08: Working with GIS Data & Spatial Data Exercise

Learning Objective: Basic skills for visualizing geo-spatial data.

- Understand the basics of working with GIS data.
- Be able to visualize the geo-spatial data with layers.
- Practice the basics of geo-coding with google APIs.

Assignment: Instructions are in the "GIS Data Exercise" folder on Dropbox.

We encourage you to use R to handle geospatial data. While there are many programs out there, R has an impressive library of packages for working with spatial data and conducting spatial analysis commonly used in the social sciences. You are welcome to find your own set of resources if you prefer using a separate program (eg. Python, ArcGIS, Stata, etc.)for this section. While we recommend using the *sf* package to complete the assignment, the resources below uses a mix of both *sp* and *sf*.

Readings:

(a) Introduction to GIS basics in R:

- Chapter 1 and 2 of Manuel Gimond's *Intro to GIS and Spatial Analysis* https://mgimond.github.io/Spatial/introGIS.html
- Chapters 2–4 of Robert J. Hijmans' "Spatial Data in R": https://rspatial.org/spatial/Spatialdata.pdf
- GIS data types cheatsheet: https://geol260.academic.wlu.edu/course-notes/introduction-and-data-types/3-types-of-data/
- Geospatial data formats: https://www.lib.ncsu.edu/gis/formats

(b) GIS in R:

- Complete Datacamp's Spatial Analysis with sf and raster in R. https://app.datacamp.com/learn/courses/spatial-analysis-with-sf-and-raster-in-r
- Nick Eubank's "GIS in R" has more advanced tutorials for using R to handle spatial data, including cheatsheets of common R commands: https://www.nickeubank.com/gis-in-r/

  - For Tutorial 4: Geo-Coding and Principles of Web-APIs, please register your API key on Google Maps Platform. Google offers a 90-day free trial. https://www.rdocumentation.org/packages/ggmap/versions/3.0.0/topics/register_google

- An application of spatial analysis: Simon Ejdemyr, "Segregation Measures in R," https://sejdemyr.github.io/r-tutorials/segregation/
- Creating a Point Map From a CSV File in R by Michael Minn https://michaelminn.net/tutorials/r-csv-point-map/index.html
- Li, X., Zhou, Y., Zhao, M. et al. A harmonized global nighttime light dataset 1992–2018. Sci Data 7, 168 (2020). https://doi.org/10.1038/s41597-020-0510-y

(c) Additional GIS resources:

- Fick, S. E. and R. J. Hijmans (2017): "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas," International Journal of Climatology, 37, 4302–4315. https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/joc.5086
- Goldblatt, R., M. F. Stuhlmacher, B. Tellman, N. Clinton, G. Hanson, M. Georgescu, C. Wang, F. Serrano-Candela, A. Khandelwal, W.-H. Cheng, and R. C. Balling Jr (2018): "Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover," Remote Sensing of Environment, 205. https://gps.ucsd.edu/_files/faculty/hanson/hanson_publications_landsat.pdf
- Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, et al. (2013): "High-resolution global maps of 21st-century forest cover change," Science, 342, 850–853. https://pubmed.ncbi.nlm.nih.gov/24233722/
- Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem (2015): "Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data," PloS one, 10, e0107042. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107042

Week 09 & 10: Production

Learning Objective:

- Draw on theories and methods canvased throughout the course and see how all these skills come together.
- Define your research question and hypotheses.
- Explore and justify your research designs.
- Incorporate reproducible research practices and implement the techniques.

Assignment: Finish capstone project! Questions that can be used as inspiration for capstone projects can be found in the GIS Data Exercise instructions.

Creating Efficient and Reproducible Code:

- Re-read ESOC's Research Production Guide
- Matthew Gentzkow and Jesse M. Shapiro, "Code and Data for the Social Sciences: A Practitioner's Guide," `http://web.stanford.edu/~gentzkow/research/CodeAndData.pdf`
- Garret Christensen and Edward Miguel, "Transparency, Reproducibility, and the Credibility of Economics Research", *Journal of Economic Literature* 56(3):2018, `https://www.aeaweb.org/articles?id=10.1257/jel.20171350`
- This was written as a STATA coding guide, but goes over best code/data management practices that are applicable for publications in general: Julian Reif, "Stata Coding Guide", `https://julianreif.com/guide/#setting-up-the-environment`
- World Bank DIME, Research Standards, `https://github.com/worldbank/dime-standards`

Working with LaTeX:

Overleaf is a commonly used collaborative writing tool in academia (think of it as Google Docs but with a LaTeX editor). Though you will likely use Overleaf for most of your work as a FDR fellow, you may also want to consider installing a TeX distribution on your computer as well. The common ones are MacTex for Mac and MikTex for Windows.

- A quick guide to : `http://users.dickinson.edu/~richesod/latex/latexcheatsheet.pdf`
- Andrew Roberts' LaTeX guide: `https://www.andy-roberts.net/writing/latex`
- Overleaf has many guides on their website. You can start with "Learn LaTeX in 30 minutes": `https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes`
- There are many references package available, like `natbib` and `biblatex` A reference sheet for the `natbib` package is available here: `https://gking.harvard.edu/files/natnotes2.pdf`
- Beamer is a LaTeX document class for creating presentation slides. The basics are covered here and here

One of the most important skills to learn is to automatically convert your code results into nicely-formatted tables for publication. You should start by understanding how the "table" environment works in : `http://www1.maths.leeds.ac.uk/LaTeX/TableHelp1.pdf`. Luckily, there are many packages available out there that can format your results into tables automatically. Here are some resources with code samples to get started:

- STATA workflow:
  - `https://lukestein.github.io/stata-latex-workflows/`
  - `https://medium.com/the-stata-guide/the-stata-to-latex-guide-6e7ed5622856`
- R workflow:
  - `https://www.jakeruss.com/cheatsheets/stargazer/` (`Stargazer` is the most frequently used package for R)
  - `https://haozhu233.github.io/kableExtra/awesome_table_in_pdf.pdf` (`kableExtra`)
  - `https://cran.r-project.org/web/packages/xtable/vignettes/xtableGallery.pdf` (`xtable`)

- Python workflow:
  - https://tug.org/tug2019/slides/slides-ziegenhagen-python.pdf
  - https://github.com/mwburke/stargazer/blob/master/examples.ipynb (`Stargazer`)
  - https://medium.com/@vince.shields913/econometrics-with-python-pt-4-20b7842f01df (also `Stargazer`)

Additional Resources

Regressions:

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print10.pdf
- Joshua Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics*, https://jonnyphillips.github.io/FLS6415/Class_3/Angrist%20&%20Pischke.pdf

Data visualization:

- Andrew Heiss' Data Visualization class (using R): https://datavizm20.classes.andrewheiss.com/
- Descriptive Statistics and Visualizing Data in STATA: http://courses.washington.edu/b517/Misc/Disc2.pdf
- "Visualization" chapter of Jake VanderPlas, *Python Data Science Handbook*, https://jakevdp.github.io/PythonDataScienceHandbook/04.00-introduction-to-matplotlib.html

Implementing empirical models in code:

- Princeton DSS has slides on how to implement all types of models in R and STATA: https://dss.princeton.edu/training/
- Model Estimation chapter of the Library of Statistical Techniques (LOST) has R and STATA code as well: https://lost-stats.github.io/Model_Estimation/Model_Estimation.html
- STATA specific:
  - Xiao Chen, Philip B. Ender, Michael Mitchell and Christine Wells, *Regression with Stata*, https://stats.idre.ucla.edu/stata/webbooks/reg/
  - Xiao Chen, Phil Ender, Michael Mitchell and Christine Wells, *Logistic Regression with Stata*, https://stats.idre.ucla.edu/stata/webbooks/logistic/
  - Handouts from BerkeleyX's CEGA101AIE Applied Impact Evaluation course:

        * Diff-in-diff designs: https://edge.edx.org/assets/courseware/v1/b8d2a8030b7aa5f2762a4 c4x/BerkeleyX/CEGA101AIE/asset/Module_2.5_Difference_in_Differences.pdf
        * Regression discontinuity: https://edge.edx.org/assets/courseware/v1/fe1cb61a45c21910d8981c75298484d2/c4x/BerkeleyX/CEGA101AIE/asset/Module_2.4_Regression_Discontinuity.pdf

- R specific:
    - Christoph Hanck, Martin Arnold, Alexander Gerber, and Martin Schmelzer, *Introduction to Econometrics with R*, https://www.econometrics-with-r.org/
    - Constantin Colonescu, *Principles of Econometrics with R*, https://bookdown.org/ccolonescu/RPoE4/
    - Anthony Schmidt, *Causal Inference in Education*, https://bookdown.org/aschmi11/causal_inf/
- Python specific:
    - Chapters 18–20 of Kevin Sheppard, *Introduction to Python for Econometrics*, kevinsheppard.com/teaching/python/notes/ (covers LM/GLM)
    - "Causal Inference for The Brave and True" is an open source causal inference guide in Python: https://matheusfacure.github.io/python-causality-handbook/landing-page.html