AFRE 891: Data Analytics and Emerging Methods in Applied Economics

(3 credits: Tue Thu: 1:00 PM-2:20PM – Ag Hall 119A)
Department of Agricultural, Food, and Resource Economics

Instructor: Dr. James Sears (he/him/his)
Department of Agricultural, Food, and Resource Economics
Office: 306 Agricultural Hall
Phone: +1-517-353-4518
Email: searsja1@msu.edu
Office Hours: Thursdays 10:15-11:45AM and upon request (in person and Zoom)

Course Description:

Training in data analytics for applied economists and social scientists with emphasis on data acquisition, manipulation, and visualization, and spatial data methods. Coverage of recent empirical methods including machine learning and synthetic control methods.

Prerequisites:

AFRE 802 and AFRE 835 or equivalents. In order to focus energy on data science skills and forefront methods, this course assumes graduate-level knowledge of econometric methods and statistics.

Course Objectives:

This course is designed as a primer on modern data science and applied economic analytical techniques for Ph.D. and advanced masters students. At a high level, this course aims to provide a core competency in practical data analytics skills at all stages of the research supply chain: from data acquisition to data cleaning, manipulation, and visualization, as well as forefront synthetic control methods and techniques like machine learning, spatial analysis, and big data analysis that lie outside the scope of other quantitative methods courses. Many of these skills are essential components of the applied economists' research workflow but are nevertheless often left out of the core graduate program curriculum. My intention is to provide training in the empirical tools that will help students make the transition from consumers of graduate coursework to producers of high-quality, modern, empirical economic research.

In short, this is the class I wish I had could have taken before starting my dissertation research.

Learning Outcomes:

Upon successful completion of this course, students should be able to

- Utilize modern statistical programming methods for data cleaning, merging, and manipulation of complex data structures (including geospatial data)
- Effectively structure research project workflows and generate reproducible code
- Create professional-quality data visualizations
- Use web scraping techniques to compile datasets from websites with wide ranges of webpage structures
- Estimate fixed effects and instrumental variable models quickly in R, compile high-quality results tables
- Effectively utilize synthetic control and machine learning methods, understand how these empirical methods differ from traditional econometric models, and when these techniques are appropriate

You will meet the objectives listed above through the combination of the following activities in this course:

- Lectures with students actively replicating code and analyses
- Homework assignments in which students will replicate and extend existing research papers
- Development of a research project. This process has two main steps: (1)

Delivery Format

Lectures will be taught in-person on Tuesdays and Thursdays from 1:00 – 2:20am in Ag Hall 119A. I will also record all lectures and upload recordings to the class D2L for later viewing (for those who are unable to make class on a given day or for later review) within 48 hours of class time.

Required Software

Please plan to have all required software (R, RStudio, and Windows Rtools/macOS R toolchain) installed before our first lecture as we will jump right into things. If you run into installation issues, please reach out to me as soon as possible. I will also be available in office hours during the first week of the semester for additional troubleshooting.

1. R

The class will be taught primarily using the statistical programming language R. While R can be used through the console, I will be teaching with the RStudio integrated development environment (IDE), and highly recommend you install it as well.

Why R? Beyond the immediate benefits of being free and open-source, R has been designed from the ground-up for data science and accordingly is heavily utilized in industry, increasingly used for academic research, and has a rich user community. When it comes to emerging methods, many tools appear in R before other platforms. R also has tremendous flexibility in the types of data it can read and the analytical methods it can employ, allowing for simplified workflows. As well, R can be much faster than Stata in typical econometrical use cases.

You can download R for free from the Comprehensive R Archive Network (CRAN) for Linux, macOS, and Windows.

RStudio Desktop (also free) is available here for a wide range of operating systems.

2. Tools for R Packages

Throughout the course we will be making use of R packages, bundles of functions and documentation that build on base R and allow R to do… a lot of stuff. While many are available centrally through CRAN and can be installed directly, some packages that we will use in this class are only available through Git repositories and require additional tools for installation. Please download and install the following tools for your OS:

- Windows: install Rtools.
- macOS: install the macOS R toolchain.
- Linux: no additional software needed
- Jailbroken TI-83 calculator: I have some bad news…

A note: the main goal of this class, data analytical literacy, is not specific to R. While we're developing our skills this semester through R, the ability to conceptualize analytical workflows and solve data challenges is core to data analysis in Stata, Python, ArcGIS, and any other platform you may use in the future or may deem optimal for a given task.

Git and Github Classroom

We will also be making use of the version control system **Git** for version control and ease of collaboration. In addition, I will be running our course through **Github Classrooms** for resource distribution, assignment submission, grading, and feedback provision. Early in the course you should receive an email invitation with instructions on how to join the course repository.

I will indicate in-class and on lecture slides when any additional software needs to be installed.

Optional Textbooks and References

There is no required textbook for this class. The lecture notes are designed to be as self-contained as possible and draw from various resources (huge shoutouts to Nick Hagerty at Montana State, and Ed Rubin and Grant McDermott at U Oregon). A few of my inspiration sources are provided below as well as other recommendations for expanded reading – all are available for free online.

- R for Data Science by Hadley Wickham and Garrett Grolemund
- Advanced R by Hadley Wickham
- Data Visualization: A Practical Introduction by Kieran Healy
- Geocomputation with R by Robin Lovelace, Jakub Nowosad, and Jannes Muenchow
- Spatial Data Science with Application in R by Edzer Pebesma and Roger Bivand
- An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani

As the above shows, there is a wealth of incredible resources available online and I encourage you to make use of any additional resources that you want. If you find a specific resource of particular benefit, I would love to hear about it to hopefully integrate more of its perspective in the future.

Assessments and Grading

Assessments for the course fall into one of two types:

1. Homework Assignments
   Homework assignments are designed to provide hands-on practice with the analytical skills covered throughout the semester. Many will take the form of replicating and extending existing studies, manipulating and analyzing a provided dataset, or compiling an original dataset. I encourage you to work with your fellow students to complete these assignments, but every student is expected to submit their own unique version of the assignment; the nature and extent of any collaboration with other students should be thoroughly documented at the start of a submission.

   Unexcused late assignments will be given a grade of zero. If you are worried that you will be unable to submit an assignment on time, make sure to communicate that to me as far in advance as possible.

2. Research Project and Final Presentation

Throughout the semester, students will work on an independent research project. Partway through the semester students will submit a brief prospectus outlining the proposed project and how it relates to both the course content and your graduate research portfolio. This can take the form of developing a novel dataset, replicating a paper in your field and extending it with the addition of new data or empirical techniques, or using covered course methods to tackle a desired research question. Consider this an opportunity to incubate one of your ideas for a potential third-year paper, or make progress on a new research question.

At the end of the semester, students will give a 20-minute, conference-style presentation on your research project. Each presenter will be assigned a discussant, who will review the final project submission and provide a brief 5-minute oral feedback following the project's presentation.

While the final presentation is structured to mimic the format of an economics conference, I will only be grading on the portion of content that is relevant to our course content (I enjoy separating hyperplanes as much as the next person, but the emphasis of your presentation should be on the empirical work). However, effectively communicating your research with others is an essential skill as an applied economist; at various stages during the course I will provide guidance on structuring an economics research presentation and how to deliver it in front of an audience.

In addition to the final research presentation, students will submit replication packages (code and utilized data) that will allow for the replication of all figures and tables presented in the research presentation. If there are privacy or data sharing concerns regarding your data, students should make arrangements with me in advance of the final presentation.

Course Grade Determination

Final course grades will be calculated as follows:

Homework Assignments	4 at 20% each = 80%
Research Project: Presentation	7.5%
Research Project: Code and Replication	7.5%
Research Project: Discussant	5%

Percentage grades will be converted to the university 4-point grading scale based on the following:

4.0 = 92.0% to 100%	**2.0** = 70.0% to 74.9%
3.5 = 85.0% to 91.9%	**1.5** = 65.0% to 69.9%
3.0 = 80.0% to 84.9%	**1.0** = 60.0% to 64.9%
2.5 = 75.0% to 79.9%	**0** = Less than 60.0%

Office Hours:

I will hold office hours from 10:15-11:45am on Thursdays to assist with questions. I will be available both in person and over Zoom (https://msu.zoom.us/j/97344840516, pass: gogreen). If that time does not work for you, please send me an email and we can schedule a different time to meet.

Participation and Engagement:

During all classes, I expect students to be fully engaged and adequately prepared to follow along with coding and covered methods. Bringing an appropriately set up laptop with you is essentially required to keep up with course materials. I encourage students to ask questions of the instructor and their peers.

Note that, while I will not take attendance, the course moves very quickly and builds upon itself. Attendance and engagement is necessary for success in this class.

OTHER COURSE POLICIES:

Diversity, Equity, and Inclusion: Diversity, Equity and Inclusion are important, interdependent components of everyday life in the College of Agriculture and Natural Resources (CANR) and are critical to our pursuit of academic excellence. Our aim is to foster a culture where every member of CANR feels valued, supported and inspired to achieve individual and common goals with an uncommon will. This includes providing opportunity and access for all people across differences of race, age, color, ethnicity,

gender, sexual orientation, gender identity, gender expression, religion, national origin, migratory status, disability / abilities, political affiliation, veteran status and socioeconomic background.

Academic Honesty: Article 2.3.3 of the Academic Freedom Report states that "The student shares with the faculty the responsibility for maintaining the integrity of scholarship, grades, and professional standards." In addition, the (insert name of unit offering course) adheres to the policies on academic honesty as specified in General Student Regulations 1.0, Protection of Scholarship and Grades; the all-University Policy on Integrity of Scholarship and Grades; and Ordinance 17.00, Examinations. (See Spartan Life: Student Handbook and Resource Guide and/or the MSU Web site: www.msu.edu.) Therefore, unless authorized by your instructor, you are expected to complete all course assignments, including homework, lab work, quizzes, tests and exams, without assistance from any source. You are expected to develop original work for this course; therefore, you may not submit course work you completed for another course to satisfy the requirements for this course. Also, you are not authorized to use the www.allmsu.com Web site to complete any course work in this course. Students who violate MSU academic integrity rules may receive a penalty grade, including a failing grade on the assignment or in the course. Contact your instructor if you are unsure about the appropriateness of your course work. (See also the MSU Academic Integrity webpage.)

Limits to Confidentiality. Assignments, code, and other materials submitted for this class are generally considered confidential pursuant to the University's student record policies. However, students should be aware that University employees, including instructors, may not be able to maintain confidentiality when it conflicts with their responsibility to report certain issues to protect the health and safety of MSU community members and others. As the instructors, we must report the following information to other University offices (including the Department of Police and Public Safety) if you share it with one of us:

- Suspected child abuse/neglect, even if this maltreatment happened when you were a child,
- Allegations of sexual assault or sexual harassment when they involve MSU students, faculty, or staff, and
- Credible threats of harm to oneself or to others.

These disclosures may trigger contact from a campus official who will want to talk with you about the incident that you have shared. In almost all cases, it will be your decision whether you wish to speak with that individual. If you would like to talk about these events in a more confidential setting you are encouraged to make an appointment with the MSU Counseling Center.

Accommodations for Students with Disabilities (from the Resource Center for Persons with Disabilities (RCPD): Michigan State University is committed to providing equal opportunity for participation in all programs, services and activities. Requests for accommodations by persons with disabilities may be made by contacting the Resource Center for Persons with Disabilities at 517-884-RCPD or on the web at rcpd.msu.edu. Once your eligibility for an accommodation has been determined, you will be issued a Verified Individual Services Accommodation ("VISA") form. Please present this form to the instructor at the start of the term and/or two weeks prior to the accommodation date. Requests received after this date may not be honored.

Commercialized Lecture Notes: Commercialization of lecture notes and university-provided course materials is not permitted in this course without expressed permission by the lecturer.

Tentative Course Outline

I. Intro to Data Science

Week 1: Course Overview and Introduction to R

1. Course Introduction
2. R Basics

Week 2: Productivity

3. Version Control with Git(Hub)
4. GitHub Integration with RStudio
5. GitHub Desktop

Week 3: Data Wrangling with tidyverse

6. Data Wrangling with dplyr
7. Data Tidying with tidyr

Week 4: Data Cleaning and Data Visualization

8. Relational Keys and Troubleshooting Joins
9. Cleaning Strings
10. Number Storage
11. Data Cleaning Checklist
12. Principles of Data Visualization
13. Data Visualization with ggplot2

II. Data Acquisition

Week 5: Data Acquisition and Web Scraping

14. Finding and Acquiring Data
15. Considerations for Administrative or PII data
16. Scraping Static Web Pages

Week 6: Web Scraping Part 2

17. Scraping Dynamic Web Pages
18. Respectful Web Scraping

Week 7: Programming

19. Functions
20. Iteration and Parallelization

Week 8: SPRING BREAK

III. Analysis and Programming

Week 9: Fast Econometric Analysis in R

21. Faster Regression Analysis and Event Study Methods in R
22. Reporting Analytical Results in Tables and Figures

Week 10: Synthetic Control Methods

23. Canonical Synthetic Control
24. Synthetic Difference-in-Differences
25. Partially Pooled SCM

VI. Spatial Analysis

Week 11: Intro to Geospatial Data and Vector Data

26. Intro to Geospatial Data
27. Vector Data and Spatial Operations

Week 12: Raster Data and Mapping

28. Raster Data and Integration with Vector Data or Empirical Analyses
29. Static and Interactive Mapping

VII. Machine Learning

Week 13: Fundamentals of Machine Learning, Prediction Methods

30. Fundamentals of Machine Learning
31. Prediction Methods

Week 14: Classification Methods and Machine Learning for Causal Inference

32. Classification Methods
33. Machine Learning for Causal Inference

VIII. Intro to Big Data Tools

Week 15: Big Data Tools

34. Storage and Memory-Efficient Methods
35. High-Performance and Cloud Computing Tools

Week 16: Research Presentations