90872 - Using R for Policy Data Analysis (6 units)

Adjunct Professor: M William Sermons (msermons@andrew.cmu.edu , 301-499-5018)
Teaching Assistant: Felix Tettey (jtettey@andrew.cmu.edu)
Carnegie Mellon University
Heinz College - School of Public Policy and Management
Fall 2021 Mini 1, Tuesday, 6:00 – 8:50 p.m.

Office Hours:

On the first day of class, we will come up with a plan for how to hold office hours. One option will be to create one-on-one Zoom breakout rooms at some point (e.g. beginning of class, end of class) during or just outside of our class meeting period. On our first meeting, we will also agree on a weekend time for R technical support from Felix. The first of those meetings will be by an R Primer on September 5th. The primer will be recorded so that students who can't attend can view it later.

Course Description/Objective:

Data analysis is an essential part of quantitative policy analysis; however, focused application of statistical methods is outside of the scope of what can be taught in classes such as Cost – Benefit Analysis (CBA) and Program Evaluation. In this course, students will apply a variety of data analysis techniques using R, a free open source statistics and graphical analysis environment that is increasingly used by data miners and analysts. Class sessions will include a combination of instruction on data analysis techniques, in-class application using R, and presentations by practicing policy data analysts. Applications will focus on analysis that is relevant to the social safety net, including cases that focus on consumer protection, affordable housing and homelessness.

Students will gain experience with data analysis that is critical to the successful execution of CBA and Program Evaluation studies. By the end of the course, students will be able to use R to:

- Estimate the size of a population impacted by a policy or program (e.g. number of people experiencing homelessness at a point in time)
- Estimate the incidence of a relevant condition (e.g. being housing cost burdened) or central tendency of relevant variables (e.g. average monthly rent).
- Illustrate differences in incidence measures across demographic groups. Calculate disproportionality and disparity indices.

- Use statistical tests to evaluate the statistical significance of differences in found across groups.
- Conduct multivariate analysis, including regression, cluster analysis, and classification and regression tree analysis.

Text Materials:

There is no text for this course; however, Canvas links to both videos and Data Carpentry pages will be included in Canvas.

Prerequisite Skills:

This course builds upon material taught in the MSPPM core curriculum and it is expected that students will have completed a course in statistics using R. Students are expected to have familiarity with basic statistics concepts, such as measures of central tendency, t-testing for differences in sample means, analysis of variance, and ordinary least squares regression analysis.

Because some students have not taken Stats with R, we will hold an R Primer on the first weekend of the course.

Weekly Schedule:

Each week, we will have three type of activities – asynchronous pre-work to be completed before class; homework assignments; and in-class activities. Students are expected to participate in all activities

Asynchronous Pre-Work:

The asynchronous portion of the course will be focused on the applications using R of our course concepts. Each week, a new R Notebook with explanatory text and executable R code will be made available for students to review before class. Students are expected to review these notebooks each week in preparation for both class and executing the next week's homework.

In-Class Activities:

Our scheduled meeting time each week is 6:30 – 9:40 each week. Because the asynchronous pre-work is work that would have normally taken place in class, we will meet for less than the scheduled time most weeks. In class, we will focus on the presentation and discussions of core course concepts, real-world examples of those core concepts, and review of common R application challenges and solutions. Examples of what we'll do in class:

- Mini-lectures introducing the courses core concepts and practical examples.
- Guest speakers: these will be practitioners doing high-quality policy data analysis.
- Discussions of what's being learned in homework and projects and on how course concepts are coming up in other courses, your fellowships.
- Demonstrations of solutions to common application problems that are appearing in homework solutions.

Homework Assignments:

- R Practice Assignments: Each week, students will be asked to replicate the analysis demonstrated in the asynchronous R Notebook and/or presented in class. Students will be asked to are asked to pick one dataset and to use it throughout the course, including your final project. One of the first assignment's in the course will be selecting a dataset.
- Online Discussion of Class Speakers: There will also be a Canvas discussion after each guest speaker with the following prompts:
    - Where does the presenter's work fit within the context of our course?
    - Was there a valuable take-away that you will incorporate into your work?

  Students are expected to contribute to the discussion by posting their own responses and commenting on colleague's responses.
- Progress on Course Project: Students are expected to complete a project where they apply the concepts of the course to their own independent analysis. There will be assignment's intended to ensure that students are making progress.

Course Requirements & Grading:

The requirements for this course are intended to reinforce the course objectives. Students are expected to attend all synchronous class sessions, to complete asynchronous preparation work before coming to class, to complete all in-class and homework assignments, and to participate in class discussions. Grading will be based on the completion of the following assignments:

- ***Weekly data analysis projects:*** These projects will give students experience using their selected public data source applying the analytical techniques taught in the course. Successful completion of these projects should adequately prepare students for the final project and move students along toward completion of their project. These projects are due by 6:30 pm on Tuesday. 20% will be deducted for late assignments turned in after the due date. No assignments will be accepted if they are more than a week late.
- ***Participation:*** There will be a variety of online discussions throughout the course conducted on Canvas. These will include small group discussions about peer project proposals and an ongoing discussion for each of our data sources to provide peer

support. There will also be a Canvas discussion after each guest speaker. Students will receive credit for participation in each online discussion.

- ***Final project****:* Students will complete a final project on a topic for their choice that uses R to analyze data and demonstrate of the techniques emphasized in the course. The project should be relevant to a policy issue, preferably one of interest to the student. The final project will consist of a paper that documents the problem being addressed, describes the methods applied, and presents and interprets the results. To make sure that projects are suitable and feasible, students will submit a short proposal of their project in the 3rd class period.

Percent (%)	Assignment
40 %	Weekly data analysis activities
20 %	Participation
40 %	Final Project

Grading Scale:

A+	99.0-100%	B+	88.0-90.9%	C+	78.0-80.9%
A	94.0-98.9%	B	84.0-87.9%	C	74.0-77.9%
A-	91.0-93.9%	B-	81.0-83.9%	C-	71.0-73.9%

Attendance Policy:

Students are expected to attend all classes without exception. I recognize, however, there can be unforeseen circumstances and emergencies that arise. Students may be granted one excused absence for the course which could include an illness or personal emergency (you need to contact me within 1-2 days of missing class if not sooner in order to be excused) or an apprenticeship-related travel/opportunity that is worked out with me in advance of the missed class. All absences due to reasons other than illness or personal emergency will be considered unexcused unless they are arranged with me in advance.

All unexcused absences will result in points being deducted from a student's final grade. Specifically, five (5) percentage points will be deducted for each unexcused absence.

Please note that even if a student misses a class (whether excused or unexcused), assignments due for that day must still be completed and handed in **on time**. Points will be deducted for late assignments. Under certain circumstances, such as illness of the student, the instructor may grant extensions to due dates.

Cheating & Plagiarism:

Students are expected to honor the letter and the spirit of the *Carnegie Mellon University Policy on Cheating and Plagiarism*. All activities cited in that policy will be treated as cheating in this course. Students are expected to familiarize themselves with this policy. Students are also encouraged to review the *Carnegie Mellon University Academic Disciplinary Actions Overview for Graduate Students,* which details penalties and sanctions, as well as students' rights. I will take a zero-tolerance policy on cheating and plagiarism and will consult with Departmental leadership on appropriate action for all instances of cheating and plagiarism. As the aforementioned policies indicate, penalties can include course failure, suspension, and dismissal from the program.

Carnegie Mellon University Policy on Cheating and Plagiarism:
http://www.cmu.edu/policies/documents/Cheating.html

Carnegie Mellon University Academic Disciplinary Actions Overview for Graduate Students:
http://www.cmu.edu/policies/documents/GradDisc.html

Data Sets for Exercises and Projects

The R Notebooks and in-class demonstrations using R will all use the American Housing Survey, which is a statistical survey funded by the United States Department of Housing and Urban Development (HUD) and conducted by the U.S. Census Bureau. It is the largest regular national housing sample survey in the United States and contains information on the number and characteristics of U.S. housing units as well as the households that occupy those units. Weekly application assignments will require that students use the same methodologies from the demonstrations on a new dataset using different variables. Students are asked to select one of the following sources for their weekly assignments and for their final project.

- ● American Community Survey (2018 1-year Public Use Microsample)
- ○ National Health Interview Survey (Public Use Files)
- ○ Household Pulse Survey (Public Use Files)

We will have ongoing discussions in Canvas for each of the data sets. The idea is for students to seek and offer peer support. Extra credit will be awarded to students who provide peer support to their classmates in that forum.

Course Schedule:

Dates	Topic	Assignments
Aug 31 – Sept. 6	• Course Overview • Reviewing R Skills Inventory • Demonstration: Using DPLYR to prepare data for analysis	• Assignment: Evaluating successful course projects. • Assignment: Preliminary project topic and dataset. • Notebook: Using DPLYR to prepare your data for analysis
Sept. 7 – Sept. 13	• What to Measure: Means, proportions and special measures. • Demonstration: Using DPLYR to obtain means, proportions and special measures.	• R Application Assignment: DPLYR Using DPLYR on your proposed project dataset. • Notebook: Using DPLYR to obtain means, proportions and special measures. • Notebook: Outliers and transformations. • Online Discussion: Peer feedback on project topics.
Sept. 14 – Sept. 20	• Documenting differences across groups. • Demonstration: Using DPLYR to identify differences in means and proportions across groups (geographic, demographic) • Demonstration: Using GGPLOT to visualize differences in means and proportions across groups	• R Application Assignment: Incidence and averages across groups. • Notebook: Quantifying and visualizing differences across groups.
Sept. 21 – Sept 27	• Significance testing: Which differences are meaningful. • Demonstration: Using OLS (LM) and logistic regression (GLM) to test for one-way differences across groups.	• R Application Assignment: Testing for the significance of differences across groups • Project update: Hypotheses about differences across groups. • Notebook: Regression for hypothesis testing (glm and lm). • Notebook: Crosstabs and Chi-Square
Sept. 28 – Oct 4	• Disparities indices. Visual representation of disparities. • Demonstration: Obtaining disparity and disproportionality indices.	• R Application Assignment: Disparity indices. Crosstabs and Chi-Square tests • Notebook: Disparity and disproportionality indices • Notebook: Visualizing disparities

	• Demonstration: Visualizing disparities	• Online Discussion: Guest speaker discussion.
Oct 5 – Oct 11	• Multivariate Analysis • Demonstration: Using OLS (LM) and logistic regression (GLM) to conduct multivariate analysis.	• R Application: Building and interpreting multivariate regression models. • Notebook: Regression
Oct 12 – Oct 18	• Reporting Results • How to write up your analysis results.	• Finish your project!