90872 - Using R for Policy Data Analysis
(6 units)
Adjunct Professors: M William Sermons (mwilliamsermons@gmail.com, 301-499-5018
& Trevor Filipiak (filipit@andrew.cmu.edu, 240-234-4692)
Carnegie Mellon University
Heinz College - School of Public Policy and Management
Fall 2017, Tuesday, 6:00 – 8:50 p.m.
CMU-DC Office, Hall of States

Office Hours:
Office Hours will be held upon request before or after our scheduled class. Please e-mail me in
advance of the day to schedule a time to meet. I am available to answer questions by e-mail
and will respond within 24 hours of all messages.

Course Description/Objective:
Data analysis is an essential part of quantitative policy analysis; however, focused application of
statistical methods is beyond the scope of what can be taught in classes such as Cost –Benefit
Analysis (CBA) and Program Evaluation. In this course, students who have completed CBA and
Program Evaluation 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
and in-class application using R. Applications will focus on analyses that support the execution
of CBA and Program Evaluation, including cases that focus on consumer protection, affordable
housing, and public health.

Students will gain experience with data analysis that is critical to the successful execution of
CBA and Program Evaluation studies. Specifically, students will gain experience:

- Using microdata to estimate the size of a population impacted by a policy or program.
- Estimating the per-unit impact of a policy change or program implementation.
- Understanding the demographics of impacted populations, including demonstrating
  which populations are disproportionately impacted.
- Accounting for uncertainty for sensitivity analysis.

We will use R to conduct all quantitative analysis. By the end of the course, students will be
able to use R to:

- Import data into R data structures and check for missing data
• Produce scatterplots, histograms, boxplots and other graphs to identify data problems and test data assumptions (e.g. normality, linearity).
• Produce descriptive statistics, including mean, median, mode, proportion, standard deviation and other measures of central tendency and dispersion.
• Calculate confidence intervals. Conduct t-test for differences in sample means and chi-square test to test for differences in categorical variables across groups.
• Conduct bivariate correlation and simple regression analysis.

Text Materials:
The following textbook is required for the Course.
• *The Book of R: A First Course in Programming and Statistics*, Davies, No Starch Press

The following optional resources may also be useful for students.
• *R for Data Science*, Grolemund and Wickham, O’Reilly Media: http://r4ds.had.co.nz/
• *Using the R Commander: A Point-and-Click Interface for R*, Fox, CRC Press

Prerequisite Skills:
This course builds upon material taught in the MSPPM core curriculum. In particular, 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. We will use statistical software (R) to conduct these analyses.

Weekly Schedule:
Our 3 hour block will be roughly broken up as follows:
6:00 – 7:00  Short lecture
7:00 – 7:15  Break
7:15 – 8:15  Hands-on data analysis
8:15 – 8:50  Final thoughts, Q&A, Housekeeping

Course Requirements & Grading:
The requirements for this course are intended to reinforce the course objectives. Students are expected to attend all class sessions, 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 various sources of public data and an opportunity to apply the analytical techniques taught in the course. Successful completion of these projects should adequately prepare
students for the final project. Students are expected to begin their weekly projects in class and complete the work and develop the write-up at home between sessions.

- **Final project:** Students will complete a final project on a topic for their choice that uses R to analyze data and demonstrates one of the techniques emphasized in the course, such as:
  - Estimating the size of an impacted population;
  - Estimating the unit impact of a policy or program change; or
  - Identifying populations that are disproportionately impacted;

The project should be relevant to a policy issue of interest to the student. The final project will consist of a 5-page 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 by the 3rd class period.

<table>
<thead>
<tr>
<th>Percent (%)</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 %</td>
<td>Weekly data analysis activities (begun in class, completed and written-up at home)</td>
</tr>
<tr>
<td>40 %</td>
<td>Final Project</td>
</tr>
</tbody>
</table>

**Grading Scale:**

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

We will use a variety of public and proprietary data sources, with most exercises using one of the following data sets:

- American Community Survey, 2014 1-year Public Use Microsample (PUMS)
- Self-Help Fresh Start Loan customer data
- 2013-2014 National Health and Nutrition Examination Survey

**Course Schedule:**

**Week 1 (Introduction to R, Part 1)**

Lecture and live practice: Course overview. Examples of quantitative analyses that support and inform CBA and Program Evaluation

Introduction to the R Programming Language, strengths and weaknesses; Introduction to the RStudio environment; Importing data into R; Basic R syntax; Using R packages.

Reading: Davies, Chapters 1-8.
**Week 2 (Introduction to R, Part 2)**

**Lecture and live practice:** Programming with R. Understanding variables, data types, and data structures; Vectors and data frames; Loops, Functions, and objects; Debugging; R Markdown for

**Reading:** Davies, Chapters 9-12

**Week 3**

**Lecture**

Critical issues in data screening (accuracy of data file, dealing with missing data, univariate and multivariate outliers, testing for normality, linearity).

**Data Analysis:** Producing scatterplots, histograms, boxplots and other graphs that allow students to explore and characterize data, identify data problems and test data assumptions (e.g. normality).

**Reading:** Davies, Chapters 14

**Data Source(s):** Pew Overdraft data from 50 largest banks.

**Week 4**

**Lecture**

Estimating the size of the impacted population. Examples, including homelessness, where population counting is an importation public policy metric. Descriptive statistics, including mean, median, mode, proportion and other measures of central tendency for continuous (interval and ratio variables).

**Data Analysis:** Using R for descriptive statistics, including mean, median, and standard deviation for continuous variables. Using ACS microdata to count severe housing cost-burdened households.

**Reading:** Davies, Chapters 13

**Data Source(s):** American Community Survey, 2015 1-year Public Use Microsample (PUMS)

**Week 5**

**Lecture**

Understanding the demographics of impacted populations, including demonstrating which populations are disproportionately impacted. Descriptive statistics
(proportions, distributions) for categorical data. Contingency tables. Hypothesis testing. Using the T-test and Chi-square test to variables across groups.

Data Analysis: Using R to obtain provide descriptive statistics for categorical variables and descriptive statistics (of continuous variables) by category. Using R for confidence intervals and simple hypothesis testing. Using R to conduct T-tests and Chi-square tests. Estimating the degree of gluten sensitivity in the US population.

Reading: Davies, Chapters 18
Data Source(s): 2013-2014 National Health and Nutrition Examination Survey

Week 6
Lecture Correlation. Using the correlation matrix to understand relationships between variables. Correlation versus causation.
Data Analysis: Using R to identify correlation between variables
Reading: Davies, Chapter 13
Data Source(s): Students should use the data they are using for their final project.

Week 7
Lecture Analysis of Variance and Linear Regression.
Data Analysis: Using R for ANOVA and Linear Regression
Reading: Davies, Chapter 19-20
Data Source(s): Students should use the data they are using for their final project.