# RSOC 573: Methods of Survey Data Analysis

## Penn State University – Spring 2022

**Professor**: Heather Randell
**Email**: hrandell@psu.edu

### I. Course Overview and Objectives

This graduate level course is intended to provide you with familiarity with developing, interpreting, and presenting the results from theoretically-informed multivariate regression analyses. We will focus on the quantitative analysis of secondary survey data from probability samples. We will begin by reviewing basic statistical concepts, descriptive and inferential statistics, tabular analysis, ways to handle missing data, bivariate ordinary least squares (OLS) regression, and basic features of the Stata statistical software program. The course will proceed with concepts of multivariate regression, and we will cover issues that we need to pay attention to, such as nonlinearity, outliers, and interaction terms. The final section of the course will cover methods for handling complex survey data and estimation techniques for binary, nominal, ordinal, and count outcomes. We will end the course by discussing how best to present results from quantitative analyses.

We will discuss the uses of these techniques and the assumptions that we make when using them. We will also spend time discussing how to interpret results and how to choose the best method for the research question. We will conduct data analysis using Stata, a statistical analysis software package. During class, we will go through the basics of using Stata and talk through issues that come up such as dealing with missing data; saving data, code, and output; and making tables and graphs.

Throughout the course, we will cover several examples of social science research using regression analysis, where to locate and how to extract secondary survey data, and how to accurately, clearly, and succinctly describe your analytic methods and report your findings. You will become familiar with Stata, one of the most widely used statistical software programs among sociologists and economists. You will have the opportunity to analyze real survey data, and the course will be focused on selecting proper methodological techniques, accurately interpreting findings, and clearly presenting results rather than statistical theory.

OBJECTIVES: By the end of the course, you should be able to (1) describe the most common techniques currently used for survey data analysis in the social sciences; (2) recognize the assumptions of regression methods and identify how to accommodate violations of those assumptions; (3) conduct various intermediate and advanced statistics tests and draw informed conclusions about research questions; (4) apply these techniques to your own research; and (5) clearly and succinctly discuss and write about findings from simple and complex quantitative data procedures.

### II. Prerequisites

It is expected that you have completed an undergraduate statistics course and at least one introductory graduate statistics course that, together, have introduced you to basic-to-intermediate statistical concepts and methods. You should be comfortable with levels of measurement (e.g., nominal, ordinal, interval-ratio), measures of central tendency (e.g., mean, median, mode), samples, populations, and sampling distributions, measures of variability (e.g., standard deviation, variance, range), tools for understanding distributions (e.g., frequency distributions, cross-tabulations, histograms), estimation and hypothesis testing (e.g., p-values, t-test, confidence intervals, type I and type II error), measures of association (e.g., R-square, correlation), and bivariate ordinary least squares regression.

## III. Required Texts and Course Readings

Assigned readings will be a combination of chapters from a standard social science statistics textbooks and journal articles.

Required texts:

Mehmetoglu, Mehmet and Georg Jakobsen. 2017. *Applied Statistics Using Stata: A Guide for the Social Science*s. Washington, DC: Sage.

Long, J. Scott and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata.* Third Edition. College Station, TX: Stata Press.

Additional books that may be useful:

Allison, Paul. 1999. *Multiple Regression: A Primer*. London: Sage.

Miller, Jane E. 2005. *The Chicago Guide to Writing about Multivariate Analysis*. Chicago: University of Chicago Press.

## IV. Stata

We will be using Stata as our statistical software package in this course. You can access Stata virtually through Penn State's "Weblabs" service that allows students to remotely access computers in the PSU public labs. See https://it.psu.edu/news/weblabs-give-students-remote-access-computer-lab-environment . No special software is needed. Students using a browser on their personal computers can remotely connect to a PSU lab computer and get access to all the software on them, including SAS, SPSS, Stata, ArcGIS, Matlab, R and more. The direct link to weblabs is - https://weblabs.psu.edu/

On those lab computers, you may save your files to your personal PASS. That is free space PSU provides to you on their file servers. It's commonly mapped as the X drive when you login to any PSU lab computer. However, PSU only allocates a small amount of space to start. To get the maximum amount of space that PSU provides, login to https://www.work.psu.edu . On the left you should see "quota". It's probably at the default 500mb. Click on the drop down and select the maximum 10gb.

Rural Sociology grad students can also access Stata through remote access to the Armsby grad lab or on the lab computers.

Another option is to purchase a student license so that you can install Stata on your personal computer. There are 6-month, annual, and perpetual licenses available. Information on purchasing Stata is here: https://www.stata.com/order/new/edu/gradplans/student-pricing/.

Stata Tutorials: I have created Stata tutorials as step-by-step guides with examples of the coding and analysis methods we will learn in the course. The intent of these tutorials is to provide a resource for you for when you are running your own data analysis. These are works in progress, so please report to me any errors you find or areas where more detail would be helpful as we make our way through them.

Below are some useful resources to help you as you learn Stata and inevitably come across challenges and road blocks:

- ◊ Stata Cheat Sheets: https://www.stata.com/bookstore/statacheatsheets.pdf
- ◊ UCLA Stata Learning Modules: https://stats.idre.ucla.edu/stata/modules/
- ◊ Statalist: https://www.statalist.org/

- ◊ Stata Video Tutorials: https://www.stata.com/links/video-tutorials/

## V. Course Format

Classes will be mixed with lectures that cover the introduction and instruction of new concepts and methods, review and discussion of examples of existing social science research, student presentations of class readings, and Stata tutorials to help you learn the software. At least one break will be given throughout the class period.

## VI. Course Expectations

Methods for statistical analyses can be difficult, especially when you are learning new content along with new data sets and software programming. The difficulty is alleviated when you invest proper time and energy into this course. This includes attending all classes, completing all assigned readings, taking good course notes, paying attention in class, coming to see me if you feel like you are falling behind, and putting proper thought into your final paper. Please note the following expectations for this class:

1. Attend all class sessions: I do not take attendance, but part of your final grade is based on participation. More importantly, it is very easy to fall behind on this material. Missing class will negatively affect your grade and your ability to fully take advantage of this class.

2. Arrive on time: Arriving to class late is inconsiderate and distracting to your instructors and classmates.

3. Be prepared: Before each class, read the required texts. Be prepared to fully participate in discussions and activities. Engaged participation is included as part of your final course grade.

4. Web-surfing: Avoid checking emails, social media, the news, etc. during lecture. It is very easy to get distracted and miss something important from class lecture. I will integrate adequate breaks into the class when you can respond to emails, texts, etc.

5. You earn your own grade: I do not "give" you a grade. You earn your grade by successfully completing the course requirements. I will provide feedback on homework assignments and your paper proposal. You can make appointments with me to help you with class material and discuss your paper. Simply showing up for class and doing the work does not automatically equate to an 'A'. Before you submit a product, ask yourself: "Is this the best work I can possibly do?" If the answer is no, then do not expect me to award that work with an A.

## VII. Canvas

The course Canvas site (http://canvas.psu.edu/) contains the syllabus, some of the required readings, course data, codebooks, Stata code we will use in class, Stata tutorials, PowerPoints for class lecture, homework assignments, and locations for submitting assignments. I will aim to post your grades to the Canvas gradebook within one week of submission.

## VIII. Grading and Assignments

**Class Participation:** You are expected to attend all classes. It is extremely difficult to catch up on this type of subject matter, even when missing just one class. You are also expected to participate in class by answering questions and completing practice examples during the class period. One time during the semester, everyone will make a short presentation to the class on the journal article for the week that uses the statistical methods we discussed during the week prior. I will provide more details on this during our first class meeting. Class attendance and participation are worth 10% of your final grade.

**Homework Assignments:** To practice the material you have learned, you will complete six homework assignments starting in Week 3 and ending in Week 14. I will aim to post each homework assignment on the Canvas site by 5 p.m. the Wednesday before it is due. You may work collaboratively with other students on these homework assignments, but it must be clear that you have not simply copied and pasted answers from each other or duplicated a document to submit. You must submit your answers through the appropriate Canvas assignment page no later than 2:00 PM on Tuesday (right before class on the day the assignment is due). I will grade the assignments on a 10-point scale, with the following benchmarks: perfect = 10; very good = 8; adequate = 5; inadequate = 3; not submitted = 0. If you turn in your assignment late, you will lose one point for each day that it is late. At the end of the semester, I will drop your lowest grade. Collectively, the homework assignments are worth 50% of your final grade.

**Final Paper:** You are required to complete a publication-quality research manuscript using survey data and the techniques we use throughout the course. You are allowed to select any existing data set, but you should be aware that I may not have expertise with those data and may not be able to provide you with the same level of feedback and assistance if you use one of the recommended data sets listed below. If you choose to use data that are not listed below, you should speak with me about those data, explain how you are going to access them, and tell me about your previous experiences with those data. Your paper should follow the format of a published research manuscript. I recommend that you find a target journal where you may be interested in submitting your paper, locate the *Guidelines for Authors* on the journal's website, and use those guidelines (including the word count) to format your paper. Below are the four paper due dates:

- You must submit a short paragraph listing your proposed paper topic and data you plan to use by **February 1**.
- You must submit an extended abstract by **March 1**. The extended abstract should be no more than 2 pages, single spaced, and should indicate your research question, the motivation for this research (i.e., why should we care about the answers to this question?), your hypotheses, the data you will use, the current status of those data (do you have them in your possession in Stata format?), your outcome(s) of interest, your main independent variables and control variables, and one or more target journals. The extended abstract is worth 5% of your final grade.
- You must submit preliminary results (descriptive statistics and regression models) by **April 19**. The preliminary results are worth 5% of your final grade.
- The final paper is due by **May 3** and is worth 30% of your final grade. I will not accept late papers, so please plan accordingly.

Recommended Data Sets: Note that you must use survey data for your paper in this course. You may use any survey data you wish, but my assistance will be limited for data sets with which I have no experience. Here are the survey data sets with which I have the most expertise and can provide you with insight about variables, recommended analytic strategies, and data challenges:

- ◊ Demographic and Health Surveys Program (https://www.dhsprogram.com/) and IPUMS-DHS (https://www.idhsdata.org/idhs/)
- ◊ IPUMS-International (https://international.ipums.org/international/)

- ◊ Young Lives Survey (http://www.younglives.org.uk/)
- ◊ World Bank Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) (https://www.worldbank.org/en/programs/lsms/initiatives/lsms-ISA)

Here are other sources of data, though I have not worked with them before in my own research so I can only be of limited assistance:

- ◊ IPUMS (https://www.ipums.org/)
- ◊ ICPSR – repository for many different datasets (https://www.icpsr.umich.edu/icpsrweb/ICPSR/)
- ◊ General Social Survey (http://www3.norc.org/GSS+Website/)
- ◊ National Survey on Drug Use and Health (https://nsduhweb.rti.org/respweb/homepage.cfm)
- ◊ National Health Interview Survey (https://www.cdc.gov/nchs/nhis/nhis_questionnaires.htm)
- ◊ Current Population Survey (http://www.census.gov/cps/)
- ◊ American Community Survey (https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml); you can also extract ACS and Census data through Social Explorer (http://www.socialexplorer.com/)
- ◊ UNICEF MICS (http://mics.unicef.org/)
- ◊ World Bank Microdata Library (http://surveys.worldbank.org/tools/microdata-library)

Another option may be to use survey data collected by your advisor or another professor with whom you work. This is contingent on the professor's ability to share their data with you and willingness to work with you to help you get to know the data and study context.

**Class Participation** 10%
**Homework** 50%
**Extended Abstract** 5%
**Preliminary Results** 5%
**Final Paper** 30%
**TOTAL** 100%

## IX. Grading Scale

| | | |
|---|---|---|
| A = 100-94 | B = 85-83 | C = 75-70 |
| A- = 93-90 | B- = 82-80 | D = 60-69 |
| B+ = 89-86 | C+ = 79-76 | F = less than 60 |

## X. Course Schedule, Readings, and Assignments

| Week | Date | Topics | Readings | Assignments |
|---|---|---|---|---|
| 1 | Jan. 11 | Course overview<br>Review of basic statistical concepts<br>Basic features of Stata | Skim: Mehmetoglu & Jakobsen – Ch. 1 and Long & Freese – Ch. 2 | |
| 2 | Jan. 18 | Planning a quantitative research project with existing data<br>Locating data<br>Data extraction and documentation<br>Importing data into Stata, dataset and variable manipulation, and calculating basic descriptive statistics in Stata | Mehmetoglu & Jakobsen – Ch. 2 | |
| 3 | Jan. 25 | Cross-tabulations and elaboration<br>Sampling, estimation, and hypothesis testing<br>Bivariate regression | Mehmetoglu & Jakobsen – Ch. 3 | HW #1 due |
| 4 | Feb. 1 | Multivariate linear (OLS) regression | Mehmetoglu & Jakobsen – Ch. 4 & 5 | Paper idea due |
| 5 | Feb. 8 | Challenges in regression: confounders, mediators, omitted variable bias, etc. | Allison 1999 – Ch. 3<br>Smyth et al. 2018 | HW #2 due |
| 6 | Feb. 15 | Nonlinear relationships<br>Interactions among independent variables | Allison 1999 – Ch. 8<br>Ludwig-Dehm & Iceland 2017 | |
| 7 | Feb. 22 | Regression diagnostics, errors, and residuals | Mehmetoglu & Jakobsen – Ch. 7<br>Price & Bohon 2019 | HW #3 due |
| 8 | Mar. 1 | Missing data<br>Handling complex survey data (sample design, data clustering, and weighting) | Mehmetoglu & Jakobsen – Ch. 13 pp. 338-347<br>Berndt & Austin 2021 | Extended abstract due |
| 9 | Mar. 8 | NO CLASS – SPRING BREAK! | | |
| 10 | Mar. 15 | Binary logistic regression | Long & Freese – Ch. 5 | HW #4 due |
| 11 | Mar. 22 | Binary logistic regression | Long & Freese – Ch. 6<br>Díaz McConnell & Yellow Horse 2021 | |
| 12 | Mar. 29 | Multinomial logistic regression | Long & Freese – Ch. 8 p. 385-436<br>Akchurin 2020 | HW #5 due |
| 13 | Apr. 5 | Ordered logistic regression | Long & Freese – Ch. 7<br>Rodriguez-Lonebear 2021 | |
| 14 | Apr. 12 | Count models | Long & Freese – Ch. 9<br>Behrman 2017 | HW #6 due |
| 15 | Apr. 19 | How to present your findings<br>Catch up | Mehmetoglu & Jakobsen – Ch. 11<br>García, Gee, and Jones 2017 | Preliminary results due |
| 16 | April 26 | Work on final paper | | |
| FINALS | May 3 | FINAL PAPER DUE | | FINAL PAPER DUE |

## Articles and Book Chapters:


Akchurin, Maria. (2020). Mining and defensive mobilization: Explaining opposition to extractive industries in Chile. *Sociology of Development*, *6*(1), 1–29.

Allison, Paul (1999). Chapter 3: What can go wrong with multiple regression? In *Multiple Regression: A Primer*. London: Sage.

Allison, Paul (1999). Chapter 8: How can multiple regression handle nonlinear relationships? In *Multiple Regression: A Primer*. London: Sage.

Behrman, Julia. A. (2017). Women's land ownership and participation in decision-making about reproductive health in Malawi. *Population and Environment*, *38*(4), 327–344.

Berndt, Virginia K., & Austin, Kelly. F. (2021). Drought and disproportionate disease: An investigation of gendered vulnerabilities to HIV/AIDS in less-developed nations. *Population and Environment*, *42*(3), 379–405.

Díaz McConnell, Eileen, & Yellow Horse, Aggie J. (2021). Vulnerable and resilient: Legal status, sources of support, maternal knowledge, and the family routines of Mexican and Central American-origin mothers in Los Angeles. *International Migration Review*, *55*(2), 514–546.

García, Jennifer J., Gee, Gilbert. C., & Jones, Malia. (2016). A critical race theory analysis of public park features in Latino immigrant neighborhoods. *Du Bois Review*, *13*(2), 397–411.

Ludwig-Dehm, Sarah M., & Iceland, John. (2017). Hispanic concentrated poverty in traditional and new destinations, 2010–2014. *Population Research and Policy Review 36***,** 833–850.

Rodriguez-Lonebear, Desi. (2021). The blood line: Racialized boundary making and citizenship among Native nations. *Sociology of Race and Ethnicity*, *7*(4), 527–542.

Price, Carmel. E., & Bohon, Stephanie. A. (2019). Eco-moms and climate change: The moderating effects of fertility in explaining gender differences in concern. *Social Currents*, *6*(5), 422–439.

Smyth, Jolene D, Swendener, Alexis, and Kazyak, Emily. (2018). Women's work? The relationship between farmwork and gender self-perception. *Rural Sociology*. 83(3): 654-676.