Course Description and Syllabus

# SURV 617 (JPSM) / SURVMETH 687 (MPSM)
# Applications of Statistical Modeling

***Joint Program in Survey Methodology, University of Maryland***

***Program in Survey Methodology, University of Michigan***

***FALL, 2019***

**Abstract**: *Applications of Statistical Modeling*, designed and required for students on all three tracks of the two programs in survey methodology, will provide students with exposure to applications of more advanced statistical modeling tools for both substantive and methodological investigations that are not fully covered in other MPSM or JPSM courses. Modeling techniques to be covered include multilevel and marginal modeling techniques for clustered or longitudinal data (with applications to methodological studies of interviewer effects and modeling trends in the Health and Retirement Study), structural equation modeling (with an application of latent class models to methodological studies of measurement error), and classification trees (with an application to prediction of response propensity). Discussions and examples of each modeling technique will be supplemented with methods for appropriately handling complex sample designs when fitting the models. The class will focus on essential concepts, practical applications, and software, rather than extensive theoretical discussions.

## Instructor: Yajuan Si, Ph.D.

| | |
|---|---|
| **Course:** | **SURV 617 (JPSM) / SURVMETH 687 (MPSM)** |
| **Dates:** | **September 6 – December 13, 2019** |
| **Lectures:** | **Fridays, 10:00am – 12:30pm** |
| **Locations:** | **University of Michigan, ISR 1070** |
| | **University of Maryland, LeFrak Hall Room 2208** |

**Instructor Office, Phone Numbers, Email Addresse, and Web Pages:**

**Yajuan Si**
MPSM: 4014 ISR
JPSM: Occasionally Wandering the Hallways
Office: (734) 764-6935
Mobile: (845) 798-2013
Email: yajuan@umich.edu
Web: http://www.umich.edu/~yajuan

## Course Description

This course will introduce students to applications of more advanced statistical modeling tools for various substantive and methodological survey research investigations. The modeling techniques that will be covered in this course are typically given only brief, introductory overviews in other MPSM and JPSM courses, and the course is therefore designed to provide students on all three tracks with more in-depth exposure to practical applications of the various models. In terms of

learning outcomes, students should expect to gain practical knowledge with regard to real-world applications of these various models, in addition to a thorough understanding of more advanced methods for fitting, interpreting, and diagnosing complex statistical models.

The first general class of modeling tools covered in the course will be **multilevel and marginal models for clustered and longitudinal data**. Multilevel models are unique in that they include both *random effects* and *fixed effects*, enabling additional inferences about the variance in model coefficients between randomly sampled clusters or individuals at higher levels of a multilevel data hierarchy. They are also extremely flexible in their ability to effectively describe complex relationships in clustered or longitudinal data sets. In survey methodology, we are frequently interested in the effects that human interviewers have on the survey measurement process, and multilevel models provide natural tools for describing these effects, given that we can view the interviewers working on a given survey as being randomly sampled from a larger pool of hypothetical interviewers that could have worked on the project. After introducing general and essential statistical concepts related to multilevel modeling, we will then focus on a specific application of multilevel modeling to the study of interviewer effects in the European Social Survey, or ESS (focusing on a Belgian sample specifically). We will walk through a full example of specifying a model, using existing software to fit the model, and then interpreting and describing the results of that model in an academic report. Examples of various software procedures that can be used to fit different types of multilevel models will be covered as well, and this portion of the course will conclude with a discussion of how complex sample designs should be accounted for when fitting these types of multilevel models.

We will then turn to alternative models for longitudinal data. We will begin with a discussion of how multilevel models can be used to estimate trajectories in survey outcomes of interest in panel surveys, and estimate variance between subjects in terms of their trajectories. We will then consider alternative marginal modeling techniques for longitudinal data, including *marginal linear models with correlated errors*, and *generalized estimating equations (GEE)*. These alternative techniques will all be applied to longitudinal data from the Health and Retirement Study (HRS), enabling comparisons of results and the inferences that are possible when using the techniques. Techniques for accommodating complex sample designs when fitting models to longitudinal data (including correct handling of time-invariant and time-varying survey weights) will also be covered at the conclusion of this portion of the class.

The second general class of modeling tools covered in the course will be **structural equation models**. These models are generally defined by a combination of *measurement models*, which describe latent (unobserved) constructs that are not directly measured in a survey but rather indicated by multiple survey items, and *structural models*, which describe the causal relationships between latent constructs. *Path models* also fall under this general class of modeling tools, and describe causal relationships between variables that have been fully observed in a survey. Survey methodologists typically use structural equation models to handle problems with measurement error in survey items that supposedly measure the same construct. After introducing the general and essential concepts related to fitting structural equation models, we will consider an application of structural equation modeling to the problem of measurement error in survey items, using *latent class analysis*, and test hypotheses about causal relationships between variables of interest in a panel survey focusing on substance use behaviors (the National Epidemiological Survey on Alcohol and Related Conditions, or NESARC). Examples of various software procedures that can be used to fit these models will be covered, and methods for fitting these models when analyzing data from complex samples will be introduced as well.

The final general class of modeling tools covered in the course will be **classification trees**. These models are generally used as a type of *data mining* tool, and can be used to uncover complex interactions between predictor variables that could be used to classify cases in large data sets (e.g., "Big Data"). Classification trees can be used to mine these data and uncover cross-classes of the data that clearly delineate likely and unlikely buyers of the product. We will apply this modeling technique to the prediction of response propensity in the second wave of the NESARC (using Wave 1 covariates to build the classification tree). Response propensity modeling is often used for nonresponse adjustment purposes in large surveys, but logistic regression is frequently used to estimate these response propensities. Classification trees offer the advantage of identifying complex interactions between variables that strongly influence response propensity, and such interactions may be hard to identify using standard logistic regression modeling techniques. Methods for accounting for complex sample designs when building classification trees will also be covered briefly.

A variety of software tools will be illustrated throughout the class when fitting the models, including procedures from R, SAS, Stata, HLM, and Mplus; there will not be a focus on any one software product, although one package may be favored depending on a poll of the students. Examples of software code (or menu steps) that can be used to fit the models in a given software package will be provided regularly, in addition to annotated output from the software.

## Prerequisites

The prerequisites for SURV 617 / SURVMETH 687 include SURV 615 / SURVMETH 685 (Statistical Methods I), SURV 616 / SURVMETH 686 (Statistical Methods II), SURV 400 / SURVMETH 600 (Fundamentals of Survey Methodology) and SURV 625 / SURVMETH 625 (Methods of Survey Sampling / Applied Sampling), or equivalents of these four courses (e.g., Applied Statistics I and II, Introduction to Survey Methodology / Survey Research Methods, and Applied Sampling or Sampling Theory). Many Survey Methodology students will likely be taking SURV 701 / SURVMETH 701 (Analysis of Complex Sample Survey Data) concurrently. Permission of the instructor is also possible given adequate prior course work in graduate-level applied statistics and strong student interest in survey methodology. The course will be presented at a moderately advanced statistical level, and will assume that students are very familiar with commonly applied statistical methods (including multiple regression and logistic regression), maximum likelihood estimation principles, applied sampling methods, and hypothesis testing concepts.

## Course Format

Students in the class will meet on a weekly basis in classrooms in Ann Arbor (1070 ISR) and College Park (2208 Lefrak). These classrooms will be linked by an interactive video system that will allow the students at the two locations to see as well as hear the instructor and students in the other locations, and to view materials on an overhead display. The instructor will be in College Park for 2-3 class sessions, depending on travel schedules and in Ann Arbor for the remaining sessions.

Class time will be used for a combination of lectures and discussion of examples and analysis projects. Lecture notes and examples will be presented on PowerPoint slides, and copies of these materials will also be available to each student on the course web site (Canvas). Software packages

demonstrated in a "live" fashion during class session will include R, Stata, SAS, and Mplus. Other software (e.g., HLM) will be demonstrated when relevant. Questions are welcomed during lectures (see more below), and off-line discussions are encouraged.

Audio and video will be recorded for each class session. These recordings are to be used to review lectures, or on those rare occasions when a student must miss class, to watch the class at another time. The recordings will be available in the course website via Canvas. Students having problems accessing the recorded videos can email the instructor with questions.

If you can only join a class remotely (via BlueJeans), please let the instructor know in advance, and use this link https://bluejeans.com/563881251/6260. Participation will still be monitored carefully for those joining via BlueJeans.

## Class Readings and Participation

There is no single textbook for this course, given the variety of topics covered. Students will generally be assigned a handful of readings on a weekly basis, related to the modeling topics that will be covered in lecture in the following week. **There will be assigned readings for the first week of class.** Students with the aforementioned pre-requisites should be able to process the selected readings without much difficulty, and are welcome to email the instructor should any questions about the readings arise. The readings have been specifically selected to be more practical in nature, rather than theoretical, given the focus of the course.

**Students are required and strongly encouraged to finish all assigned readings prior to the lecture for which they have been assigned.** While this will not be checked specifically via quizzes or required email questions about the readings, class participation will be monitored carefully. **Frequent in-class participation noted by the instructor will be used to adjust the final grades for each analysis project in an upward direction, while frequent failure to participate will be used to adjust the final grades in a downward direction.** The instructor will be in constant communication about any perceived lack of participation, to ensure that this is not due to a failure to understand the course material.

## Grading

**EXAMINATIONS & FINAL GRADE**

There will be a two-hour, in-class, cumulative, open book / **open** notes midterm examination on **Friday, November 1**. (*If this examination time conflicts with other regularly scheduled examination times for University of Michigan or University of Maryland students, students must inform Dr. Si as soon as possible*.)

Final grades will be a weighted composite of homework (25%), the midterm examination (25%), and the final project report (50%). Additional details about the analysis project and the grading criteria that will be used can be found below. The final grades will not be based on any kind of curve, and will use a standard grading scale (99-100% = A+, 93-98% = A, 90-92% = A-, 88-89% = B+, 83-87% = B, 80-82% = B-, <79% = C).

## Homework

Three homework assignments are planned. Each are to be turned in by the beginning of the class session when due (see syllabus below), and each will receive equal weights for the final grades.

Homework assignments correspond with the course units and are designed to aid in skill development. Assignments will be graded check-plus (100 points), check (90), check-minus (80), and not submitted (0), and will ordinarily be marked and returned before the next class session. Students may request permission to submit homework late via email to the instructor, but the request must be no later than one hour before the homework is due. Permission is not guaranteed, although typically granted. If late submission is granted, there will be an agreed upon date and time when the late assignment must be submitted; such assignments will be graded using the specified marking system. If homework is submitted late without prior permission, scores will be check-plus (70), check (60), and check-minus (50).

# Analysis Projects

***Project Teams***. Students will form teams of two (2) for the analysis project. **Both students will be 100% responsible for careful proofreading and editing of the final report.** English grammar and clear writing will be an important part of the overall grade on each project (see the PDF file **styletips.pdf** on Canvas, courtesy of Dr. Rod Little, for importance advice). Each team will be able to decide whether they would like to either use the data set that has been discussed in class for a given topic or use a data set of their own choosing with the required variables present (e.g., interviewer / cluster ID codes). Students working on the same team are strongly encouraged to schedule a weekly 1-2 hour meeting time **outside of lecture** to make progress on the final project.

***Analysis Project Requirements***. Each team will be responsible for working on the following tasks while a given modeling topic is being covered (maximum points on the final project for each task are indicated in parentheses):

1. *Research Question (5 points):* Formulating a methodological or substantive research question that can be addressed using the modeling technique under discussion, including 1-2 (maximum) important pieces of literature motivating the research question;
2. *Data Set (5 points):* Briefly describing the data set and its structure, identifying variables in the data set that you will analyze (possibly including complex sample design features), and presenting simple (yet appropriate) descriptive statistics for each of the variables;
3. *Modeling Approach (10 points):* Specifying and clearly defining an appropriate statistical model for analyzing the data and answering the research question (**being extremely careful with notation**), and describing the general modeling approach;
4. *Software Coding (5 points):* Selecting software that you can use to fit the model and writing appropriate code that will compute the estimates of interest;
5. *Interpretation and Inference (20 points):* Correctly interpreting the modeling results, assessing model diagnostics / quality of fit, and making appropriate inferences with regard to your research question(s); and
6. *Analysis Report (50 points total):* Drafting a brief analysis report that describes Sections 1 through 5 above in clear detail (including model notation), **including overall conclusions** (*5 points*), interpretation of results, and software code used (**clearly commented in a required Appendix**).

**The final report for each topic should be submitted for grading through Canvas (only), no later than 5:00pm on the due date indicated in the syllabus.** The reports will be graded with respect to: 1) Parts 1 through 6 above; 2) the clarity of the material presented; 3) the quality of the writing; and 4) correct interpretation of the results. Figures and tables can certainly be used to enhance the report and describe the data being analyzed or the results of interest. Analysis reports should be **double-spaced, using 12-point font, and no more than 10 pages long**, including all equations, tables and figures (tables and figures will **not** be allowed in appendices). This makes concise writing extremely important. **The software code provided in the Appendix should be clearly commented and easy to follow.** An example of a high-quality analysis project from a previous year is provided on Canvas (under **Files**) as a reference point. **There will be a 5% reduction in the score for a given analysis project for each day after the due date that it takes for a team to submit the project via Canvas.**

Readings are assigned on a weekly basis. The team is expected to work on the project through the course, make regular progress on the analysis project and proofread the final report. **Making well-written, high-quality analysis reports is essential for earning high marks.** The instructor will be available to answer questions on a weekly basis with regard to the project and the analyses being conducted.

## Attendance Policy

Students are expected to attend each lecture, participate in class discussions (**again, participation may factor in to your final grade on each project**), and complete all analysis projects on time. If you must miss a lecture, please make sure to let the instructor know in advance. **Students are required to check on the timing of exams in other courses and let the instructor know if there will be any conflicts preventing their attendance on a given day**.

## Accommodations for Students with Disabilities

### University of Michigan

If you think that you need an accommodation for a disability, please contact the Services for Students with Disabilities (SSD) office to help us determine appropriate academic accommodations. SSD (734-763-3000; http://ssd.umich.edu) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. Any information you provide is private and confidential and will be treated as such.

### University of Maryland

For information about Accessibility and Disability Service (ADS) on campus, visit this web site: https://www.counseling.umd.edu/ads/.

Accessibility and Disability Service (ADS) Main Office:
Phone: 301.314.7682
0106 Shoemaker Building
Dissup@umd.edu

Office Hours: Monday - Friday (8:30am to 4:30pm)

ADS Testing Office (for ADS exams):
Phone: 301.314.7217
0118 Shoemaker Building
DSSTest@umd.edu
ADS Exam Hours: Monday - Friday (9am to 4pm)

Overview of ADS Services:
In order to receive services you must contact our office to register in person for services. Please call the office to set up an appointment to register with an ADS counselor. Contact the ADS office at 301.314.7682.

There are a number of FAQs on the following page that you may find very helpful:
http://www.counseling.umd.edu/DSS/add_questions.html

## Academic Conduct

**University of Michigan.** Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct, may be found at the Rackham web site for the University of Michigan:

http://www.rackham.umich.edu/policies/academic-policies/section11#112

Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or completing any projects in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

**University of Maryland.** Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct, may be found at the University of Maryland, Office of the President's website:

http://www.president.umd.edu/policies/docs/III-100A.pdf

# Course Schedule / Assigned Readings / Deadlines

All assigned readings need to be completed prior to the start of the indicated class, and topics from these readings will be discussed in class. **The order of the topics is subject to change.**

| Class Date | Topic / Deadline | Assigned Readings |
|---|---|---|
| September 6 | Course overview. Multilevel modeling background and concepts. Introduction of European Social Survey (ESS) data.<br><br>**HW 1 assigned: multilevel models** | 1. Syllabus<br>2. Gill and Womack (2013)<br>3. Merlo et al. (2005) |
| September 13 | Software for fitting multilevel models. | 1. Galecki and West (2013)<br>2. West and Galecki (2011) |
| September 20 | Interviewer effects: A review.<br><br>Multilevel modeling application: Interviewer effects in ESS. | 1. West et al. (2013)<br>2. O'Muircheartaigh and Campanelli (1998) |
| September 27 (Guest lecture) | Accounting for complex sample design features when fitting multilevel models.<br><br>****HW 1 DUE: multilevel models** | 1. Carle (2009)<br>2. Rabe-Hesketh and Skrondal (2006) |
| October 4 | Alternative statistical models for longitudinal data: An overview.<br><br>Software for fitting models to longitudinal data.<br><br>**HW 2 assigned: marginal models** | 1. Ballinger (2004)<br>2. Steele (2008) |
| October 11 | Application: Alternative approaches to fitting growth curve models to HRS data. | 1. Kreuter and Muthen (2008)<br>2. Hubbard et al. (2010)<br>3. Twisk (2004) |
| October 18 (JPSM onsite) | Accounting for complex sample designs when fitting models to longitudinal data. | 1. Heeringa et al. (2017)<br>2. Veiga et al. (2014)<br>3. Thompson (2015) |
| October 25 | Structural equation models (SEMs): Overview.<br><br>****HW 2 DUE: marginal models** | 1. Hox and Bechger (1998) |

| | | |
|---|---|---|
| | **HW 3 assigned: structural equation models** | |
| November 1 | **Mid-term exam** | |
| November 8 (JPSM onsite) | Multiple group Analysis (MGA) and latent class analysis (LCA).<br><br>Software for LCA. | 1. Lanza et al. (2007)<br>2. Kreuter, Yan, and Tourangeau (2008) |
| November 15 | Structural equation modeling application: Applying LCA to data from the National Epidemiologic Survey of Alcohol and Related Conditions (NESARC) | 1. McCabe and Cranford (2012)<br>2. Biemer and Wiesen (2002) |
| November 22 | Accounting for complex sample designs when fitting structural equation models.<br><br>****HW 3 DUE: structural equation models** | 1. Stapleton (2006)<br>2. Oberski (2014) |
| November 29 | **NO CLASS: HAPPY THANKSGIVING!** | 1. Pillow (2018) |
| December 6 | Classification trees: Overview and software | 1. Lemon et al. (2003)<br>2. Lewis (2000)<br>3. Ledolter (2013) |
| December 13 | Classification tree application: Prediction of response propensity in Wave 2 of the NESARC.<br><br>****Final report DUE** | 1. Wun et al. (2007)<br>2. Tollenaar and van der Heijden (2013) |