Analysis Methods for Complex Sample Survey Data (SURVMETH 614)

Summer Institute in Survey Research Techniques, 2019

Instructor:

Yajuan Si
4014 ISR
(734) 764-6935
yajuan@umich.edu
www.umich.edu/~yajuan

Course: SURVMETH 614
Dates: June 3 - July 26, 2019
Lectures: Monday, Wednesday, 1:00 - 3:00 pm
Computer Lab: Friday, 1:00 - 3:00 pm
Locations: Lectures – 3207 LSA Building
Computer Labs – 2302 School of Education (SoE) Building

Course Description

Standard courses on statistical analysis assume that survey data arise from a simple random sample of the target population. Little attention is given to characteristics often associated with survey data, including unequal probabilities of selection, stratified multistage sample designs, and missing data. Most standard statistical procedures in software packages commonly used for data analysis (e.g. R, SAS®, SPSS®, and Stata®) do not allow the analysts to take these properties of survey data into account unless specialized survey procedures are used. Failure to do so can have an important impact on the results of all types of analyses, ranging from simple descriptive statistics to analytic estimates of parameters of multivariate models.

This course first provides an introduction to specialized software procedures that have been developed for the analysis of complex sample survey data. The course begins by considering the sampling designs of three specific surveys: the National Comorbidity Survey-Replication (NCS-R), the National Health and Nutrition Examination Surveys (NHANES), and the Health and Retirement Survey (HRS). Relevant design features of the NCS-R, NHANES and HRS include weights that take into account differences in probability of selection into the sample and differences in response rates, as well as the stratification and clustering in multistage sampling procedures used to identify the sampled households and individuals. Example designs from other surveys (e.g., the European Social Survey, or ESS) will be considered as well.

The course will then move on to the introduction of variance estimation techniques that have been developed to take into account stratification and cluster sampling, which are properties of the multistage sampling designs used by most major survey programs. These techniques will initially be discussed in terms of the estimation of sampling variances for descriptive statistics, sample means, proportions and quantiles of distributions. The course syllabus will then turn to software procedures for commonly used analyses, including testing for between-group differences in means and proportions, regression analysis, logistic regression and multilevel modeling. We will also consider the consequences of nonresponse and missing data on survey analysis and methods for dealing with missing data.

Specialized procedures for survey data analysis from the SAS® and Stata® systems for data management and analysis will be used in conjunction with the Survey Research Center’s own IVEware® system to develop course examples and exercises; illustrations will also be presented using procedures from the R and SUDAAN® systems.
software packages that have been specifically designed for the analysis of survey data. Data from the NCS-R, NHANES, HRS, and ESS will be used to illustrate the various analysis procedures covered during the course in lab sessions. Homework exercises will enable students to practice analyses using the same three data sets used in the lab sessions.

**Textbook and Class Reading**

The textbook for this course will be *Applied Survey Data Analysis, Second Edition* (ASDA, 2017; authors: Brady West, Steven Heeringa and Pat Berglund; publisher: Chapman Hall / CRC Press). Students can purchase the course text from Ulrich’s book store (https://www.bkstr.com/ulrichsstore/home), or from online retailers (e.g., Amazon.com, or crcpress.com). Assigned readings will generally consist of selected sections from the chapters in the course text. The instructor also recommends that students who have a strong interest in the theory of analysis of complex sample surveys consider purchasing a copy of *Analysis of Complex Sample Survey Data*, authored by Skinner, Holt and Smith (1989). This text is out of print but it may be possible to locate a copy on Amazon.com or through other book reselling services.

In addition to assigned readings from the course text (ASDA), the instructor has prepared a supplemental readings list that includes several review articles. These supplemental readings are provided in electronic format via the University of Michigan Canvas system. Some of the supplemental readings on Canvas will be assigned, and others will be recommended. **Students are required to have finished all assigned readings prior to the lecture or lab for which they have been assigned.**

**Prerequisites**

This course is taught at an intermediate level, emphasizing both the theory and practice of the analysis of complex sample survey data. The course does not require rigorous training in mathematics; however, proficiency in basic mathematics, including algebra and functions, is essential. Knowledge of calculus and linear algebra is useful but not required for the course. A first course in survey sampling methods and a basic understanding of sampling concepts such as stratification, cluster sampling and weighting is required. Many students enrolled in this course will have also taken SURVMETH 612, Methods of Survey Sampling. Students should also have familiarity with basic statistical concepts, including point estimates, sampling variance, confidence intervals, p-values, the maximum likelihood estimation method and simple linear and logistic regression models.

**Format**

Lectures on Mondays and Wednesdays will cover basic theory and methods for each topic, and discuss examples and homework exercises. Students are encouraged to ask questions about the lecture material or the assigned readings, and discuss the topics with other students. Lecture notes and examples will be presented using PowerPoint, and electronic copies of these materials will also be provided to each student on the course Canvas website, along with other relevant information and web links pertinent to the course.

The computer lab sessions on Fridays will allow students to analyze actual survey data using the methods learned during the lectures, with assistance from the instructor, and also perform exploratory analyses and generate hypotheses for their final class projects. Lab notes with detailed syntax for the computer packages will be made available on the Canvas site at the beginning of the course. **Students should review the scheduled lab exercise prior to the Friday lab sessions.**

**Grading**


The course grading will be based on two criteria:

1. Completion of 5 homework assignments (50%)
2. A final class project (50%)

The homework assignments will typically involve carrying out an analysis of a specified survey data set (see Course Description above) and providing interpretation of results. These analyses can be done on a student’s own PC or laptop, or on workstations that are available in the University’s on-campus computing facilities. The software packages that will be used include R, SAS®, Stata®, and SRC’s IVEware®. Students are encouraged to work in groups on the homework assignments. However, the work that is submitted must be completed individually by each student. The five homework assignments are due at the beginning of the class session on the specified due dates.

Final Class Project

The primary aims of this course are to provide class participants with instruction in the theory and practice in the application of the software and methods for the analysis of complex sample survey data. The ultimate goal of this course is to prepare students to apply appropriate methods and software in the analysis of survey data and to effectively communicate the results of their analysis in the form of papers, technical reports or others forms of scientific communication. To this end, the course will require each student to develop a final project paper based on an independent analysis of a survey data set. The survey data set may be identified by the student or chosen from a list of course data sets. Work on the final project paper will begin in weeks 1 and 2 with a topic search and investigation of potential data sets. Selection of a project survey data set and topic will be finalized in week 3, and basic descriptive analyses and multivariate analyses will be conducted and reviewed by the instructor in weeks 4-6. A preliminary draft of the final paper with the initial sections (background, literature review, data and methods) will be due Friday, July 12. Students are expected to finalize the analysis for their project and complete writing their final project paper during weeks 7 and 8 of the course, and lab time will be allotted for students to complete these tasks during the final week of the course. The final paper will be due to the instructor in electronic format at 5:00 pm on Friday, July 26. The instructor will be available throughout the course to assist students in each successive phase of the development of the final project paper.

Attendance

Students are expected to attend each lecture, attend each lab session, participate in class discussions, and complete all homework and lab assignments on time. If you must miss a lecture or lab, please let the Dr. Si know in advance.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please contact 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.

Academic Conduct

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.umi...
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 taking any exams 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.
Course Syllabus
SURVMETH 614: Analysis Methods for Complex Sample Survey Data
Summer, 2019

<table>
<thead>
<tr>
<th>Date</th>
<th>Day</th>
<th>Type</th>
<th>Topics / Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Homework #1 distributed. Course Projects Introduced.</td>
</tr>
<tr>
<td>June 7</td>
<td>Fri</td>
<td>Lecture 3 / Lab 1</td>
<td>Sampling error calculation models; ultimate clusters. Preparing for survey data analysis.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lab 1: Introduction to the course computing facilities, and introduction to the NCS-R, HRS and NHANES data sets.</td>
</tr>
<tr>
<td>June 14</td>
<td>Fri</td>
<td>Lab 2</td>
<td>Lab 2: Sampling error estimation for descriptive statistics.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Homework #1 Due. Homework #2 distributed.</td>
</tr>
<tr>
<td>June 21</td>
<td>Fri</td>
<td>Lab 3</td>
<td>Lab 3: Sampling errors for subpopulation estimates. Bivariate analysis (cross-tabulation). Hypothesis testing for contrasts of subpopulation estimates. (continued on next page…)</td>
</tr>
<tr>
<td>Date</td>
<td>Day</td>
<td>Type</td>
<td>Topics / Assignments</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>--------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>June 24</td>
<td>Mon</td>
<td>Lecture 8</td>
<td>Homework #2 due. Homework #3 distributed. Prospectus describing topic and data set for Final Course Project due.</td>
</tr>
<tr>
<td>June 26</td>
<td>Wed</td>
<td>Lecture 9</td>
<td>Linear Regression (Part 1).</td>
</tr>
<tr>
<td>June 28</td>
<td>Fri</td>
<td>Lab 4</td>
<td>Homework #3 due. Homework #4 distributed. Lab 4: Linear regression analysis computational exercise.</td>
</tr>
<tr>
<td>July 1</td>
<td>Mon</td>
<td>Lecture 10</td>
<td>Logistic Regression.</td>
</tr>
<tr>
<td>July 3</td>
<td>Wed</td>
<td>July 4th Holiday</td>
<td>No class!</td>
</tr>
<tr>
<td>July 5</td>
<td>Fri</td>
<td>Lab 5</td>
<td>Lab 5: Logistic Regression analysis computational exercise.</td>
</tr>
<tr>
<td>July 8</td>
<td>Mon</td>
<td>Lecture 11</td>
<td>Homework #4 due. Homework #5 distributed. Multinomial, ordinal logistic regression. Other generalized linear models (GLMs). Hypothesis testing.</td>
</tr>
<tr>
<td>July 10</td>
<td>Wed</td>
<td>Lecture 12</td>
<td>Poisson and negative binomial regression.</td>
</tr>
<tr>
<td>July 12</td>
<td>Fri</td>
<td>Lab 6</td>
<td>Lab 6: Multinomial and ordinal logistic regression models. Examples of interpreting estimated coefficients, testing hypotheses, and making inferences.</td>
</tr>
<tr>
<td>July 15</td>
<td>Mon</td>
<td>Lecture 13</td>
<td>Homework #5 due. Introduction and design/methods for course project due. Survival analysis and event history analysis.</td>
</tr>
<tr>
<td>July 17</td>
<td>Wed</td>
<td>Lecture 14</td>
<td>Imputation of item missing data. Multiple imputation inference for survey data.</td>
</tr>
<tr>
<td>July 19</td>
<td>Fri</td>
<td>Lab 7</td>
<td>Lab 7: Multiple Imputation computational exercise.</td>
</tr>
<tr>
<td>July 22</td>
<td>Mon</td>
<td>Lecture 15</td>
<td>Multi-level models for complex sample survey data.</td>
</tr>
<tr>
<td>July 24</td>
<td>Wed</td>
<td>Work Day</td>
<td>Open lab for work on final projects.</td>
</tr>
<tr>
<td>Date</td>
<td>Day</td>
<td>Type</td>
<td>Topics / Assignments</td>
</tr>
<tr>
<td>--------</td>
<td>-----</td>
<td>------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>July 26</td>
<td>Fri</td>
<td>Work Day</td>
<td>Open lab for work on final projects. Final project due, 5pm.</td>
</tr>
</tbody>
</table>
SurvMeth 614: Analysis of Complex Sample Survey Data  
Summer 2018 Reading Assignments

* Assigned Readings need to be completed prior to the onset of the indicated class.

<table>
<thead>
<tr>
<th>Class Date</th>
<th>Topic</th>
<th>Assigned Readings*</th>
</tr>
</thead>
</table>
| June 4     | Survey estimation and inference for complex sample designs (Part 1). Complex sample designs, survey estimation and inference. Multi-stage designs, stratification, clustering, weighting, item missing data, finite population corrections. | BOOK / OTHER:  
1. Syllabus  
2. Assigned Readings  
3. References  
4. Chapters 1 and 2, ASDA |
| June 6     | Survey estimation and inference for complex sample designs (Part 2). Models and assumptions for inference from complex sample data. Sampling distribution, confidence intervals. Design effects. Introduction of course data sets. | BOOK:  
1. Chapter 3 (3.1 to 3.5), ASDA  
CANVAS:  
| June 8 (lab)| Sampling error calculation models; ultimate clusters. Preparing for survey data analysis. | BOOK:  
1. Chapter 4, ASDA |
1. Chapter 3 (3.6.1 to 3.6.2), ASDA  
2. Appendix A, ASDA (browse)  
CANVAS:  
1. Rust (1985)  
2. Siller and Tompkins (2005) |
1. Chapter 3 (3.6.3 to 3.8), ASDA  
CANVAS:  
1. Kovar et al. (1988) |
| June 15 (lab)| Sampling error estimation for descriptive statistics. | BOOK:  
1. Chapter 5 (5.1 to 5.3), ASDA  
CANVAS:  
1. SAS, Stata, SPSS documents (browse)  
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Materials</th>
</tr>
</thead>
<tbody>
<tr>
<td>June 20</td>
<td>Methods for Categorical Data.</td>
<td>BOOK: 1. Chapter 6, ASDA</td>
</tr>
<tr>
<td>June 22 (lab)</td>
<td>Sampling errors for subpopulation estimates. Bivariate analysis (cross-tabulation). Hypothesis testing for contrasts of subpopulation estimates.</td>
<td>None (catch up on previous assigned readings).</td>
</tr>
<tr>
<td>June 25</td>
<td>Linear Regression (Part 1).</td>
<td>BOOK: 1. Chapter 7 (through 7.3.3), ASDA</td>
</tr>
<tr>
<td>June 27</td>
<td>Linear Regression (Part 2).</td>
<td>BOOK: 1. Chapter 7 (through end), ASDA</td>
</tr>
<tr>
<td>June 29 (lab)</td>
<td>Linear Regression Analysis Computational Exercise.</td>
<td>None (catch up on previous assigned readings).</td>
</tr>
<tr>
<td>July 2</td>
<td>Logistic Regression (Part 1).</td>
<td>BOOK: 1. Chapter 8 (through 8.5), ASDA</td>
</tr>
<tr>
<td>July 9</td>
<td>Multinomial, ordinal logistic regression. Other GLMs. Hypothesis testing.</td>
<td>BOOK: 1. Chapter 9 (9.1 to 9.3), ASDA</td>
</tr>
<tr>
<td>July 11</td>
<td>Poisson Regression. Examples of Poisson Regression Analysis from Lab Notes.</td>
<td>BOOK: 1. Chapter 9 (9.4), ASDA</td>
</tr>
<tr>
<td>July 13 (lab)</td>
<td>Multinomial and ordinal logistic regression models. Examples of interpreting estimated coefficients, testing hypotheses, and making inferences.</td>
<td>None (catch up on previous assigned readings).</td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
<td>Sources</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>July 16</td>
<td>Survival analysis and event history analysis.</td>
<td><strong>BOOK:</strong> 1. Chapter 10, ASDA</td>
</tr>
<tr>
<td>July 18</td>
<td>Imputation of item missing data. Multiple imputation inference for survey data.</td>
<td><strong>BOOK:</strong> 1. Chapter 12 (ASDA)       <strong>CANVAS:</strong> 1. Raghunathan et al. (2001)</td>
</tr>
<tr>
<td>July 20 (lab)</td>
<td>Multiple Imputation Analysis.</td>
<td><strong>CANVAS:</strong> 1. Carlin et al. (2008)</td>
</tr>
<tr>
<td>July 23</td>
<td>Multilevel models for complex sample survey data.</td>
<td><strong>BOOK:</strong> 1. Chapter 13 (ASDA)       <strong>CANVAS:</strong> 1. Rabe-Hesketh and Skrondal (2006)</td>
</tr>
<tr>
<td>July 25 (lab)</td>
<td>Work Day.</td>
<td>None.</td>
</tr>
</tbody>
</table>
Selected articles on Canvas (others may be added as the course proceeds):


9. **Required**: StataCorp (2007), Stata Survey Data Release 10 Manual, Stata Statistical Software: Release 10, College Station, TX: StataCorp LP.


