Stat 579: **Selected Topics in Statistics: Advanced Survey Sampling**

Instructor: Dr. Yan Lu, luyan@math.unm.edu, http://www.math.unm.edu/~luyan/

Time and Location: 12:30pm-1:45pm TR, MITCH 219

Office hours: SMLC 316, TR 2:00pm-3:30pm or by appointment

Prerequisites: Stat 472/572 (Sampling: Theory and Practice) or Stat 428/528 Advanced data analysis II or equivalent. Stat 461/561 (probability) is a helpful course but is not required.

Topics: In classical finite population sampling, probability sampling is used to select a sample from a directory or map of units called a sampling frame. Sample surveys are increasingly used to obtain information about the finite populations of interest in many areas, including education, public health, sociology, ecology, agriculture, genetics, quality improvement, marketing, and accounting.

This course will begin with a review of stat 472/572, sampling: Theory and Practice, including sampling design and basic analysis of survey data. Then we discuss topics including dual frame surveys, regression, random forests, mixed models and small area estimation. These topics will be introduced with examples from survey data. R is used to aid in designing and analyzing surveys. Occasionally, SAS will also be used.

After this course, you will have basic knowledge of survey design and analysis and the application of some popular techniques such as mixed models, random forests etc, in survey sampling area. All the methods can be applied to other statistical areas. After checking for the research articles from library, reading research articles and discussing them in class, you will also have basic preparation for your future research.

Grading: Homework Assignments will be given about every 2 or 3 weeks to make for about a total of about 6-7 assignments. There is no midterm and no final exam.

Tentative schedule

Review of survey sampling basics, including introduction, probability sampling, complex surveys, variance estimation, about 2 weeks

Dual frame surveys introduction, about 1 week

Regression in complex surveys, about 2 weeks

Random forests and applications in surveys, about 2-3 weeks

Mixed models and applications in surveys, about 2-3 weeks

Small area estimation, about 3 weeks

Missing data, other design issues, about 1 week

References:

Note Textbooks are not required, handouts and several research articles will be distributed in class.

Books

Lohr, S. (2010) Sampling: Design and Analysis 2nd edition. Pacific Grove, CA: Duxbury Press.

William Gemmell Cochran, Sampling Techniques, 3rd edition, 1977

Complex surveys, A guide to analysis using R, Thomas Lumley

Jimin Jiang, Linear and Generalized linear mixed models and their applications

C.J.Skinner, D. Holt and T.M.F Smith, Analysis of complex surveys

J.N.K.Rao, Small area estimation

Thomas Lumley, Complex Surveys: A guide to analysis using R

Eugene Demidenko, Mixed Models, Theory and applications

Roderick J. A. Little and Donald B. Rubin, Statistical Analysis with Missing Data

Papers

Hartley, H. O. (1962). Multiple frame surveys. In ASA Proceedings of the Social Statistics Section (pp. 203-206). American Statistical Association.

Lohr, S. L., & Rao, J. N. K. (2000). Inference from dual frame surveys. Journal of the American Statistical Association, 95, 271-280.

Lu, Y. and Lohr, S. Gross flow estimation in dual frame surveys, Survey Methodology, 36, 13-22

Breiman L (1996) Bagging predictors. Machine learning 24, 123-140

Breiman L (1999) Using adaptive bagging to debias regressions. Technical Report 547

Breiman L (2001) Random forests. Machine learning 45:5-32

Guillermo Mendez, Tree-based methods to model dependent data, Dissertation