STATISTICAL DESIGN AND ANALYSIS IN EVALUATION
LECTURE NOTES

DAY ONE: DESIGN
DAY TWO: ANALYSIS

Joseph George Caldwell, PhD
1432 N Camino Mateo, Tucson, AZ 85745-3311 USA
Tel. (001)(520)222-3446, E-mail jcaldwell9@yahoo.com

January 9, 2012
Updated January 23, 2012, November 9, 2016

Copyright © 2012-2016 Joseph George Caldwell. All rights reserved.

Contents
1. Introduction ........................................................................................................................................ 2
2. Course Schedule ................................................................................................................................. 6
3. Course Syllabus ................................................................................................................................. 8
4. Course Critique Form .........................................................................................................................11
5. Résumé of Course Developer ............................................................................................................13
6. Day 1: Causal Modeling ....................................................................................................................15
   6.1. Overview of First Day's Course Content ....................................................................................15
   6.2. Causal Modeling .........................................................................................................................15
   6.3. Impact Measures .........................................................................................................................24
   6.4. Basic Evaluation Designs; Models; Estimators .........................................................................30
   6.5. Sample Survey Design for Evaluation .....................................................................................38
7. Day 2: Causal Analysis .....................................................................................................................59
   7.1. Overview of Second Day's Course Content ...............................................................................59
1. Introduction

These notes are intended to accompany a lecture, using a board or projector to augment the oral presentation. They have been prepared so that the student may listen to the presentation without having to take notes.

The lecture may be accompanied by examples and handouts, which are not included in these notes.

The course may also include in-class student exercises.

The course may be covered in two six-hour days (three hours in morning, three hours in afternoon), or in four half days (three hours per day). The split-up sessions are intended to accommodate clients whose employees would find it inconvenient or impractical to allocate an entire day, or three days in sequence, to a course.

The course is intended for any class size, but a smaller class size (e.g., 10-30 students) is better for interactive discussion (responses to student questions, clarifications, additional examples).

The topics covered in the two-day course are:

Day 1: Statistical Design in Evaluation
Day 2: Statistical Analysis in Evaluation

Day 1 includes discussion of causal modeling, impact measures and estimators, basic evaluation designs, and principles and techniques of design and analysis for evaluation. Day 2 is concerned with impact analysis.

The level and scope of the course; managing expectations

This course is an introductory course on statistical design and analysis in evaluation. It assumes that the student has taken a prerequisite course in “college math,” and it requires a little prior knowledge of calculus (e.g., to understand the meaning of a continuous probability distribution and expectation). For students having knowledge of calculus (or some background in probability and statistics, say from an elementary course in statistics), some additional information may be presented. This additional material is marked with the notation optional. These optional sections (few in number) are omitted from the course presentation.

Attendees should be somewhat familiar with basic statistical concepts, such as probability, discrete and continuous probability distributions, expectation, the mean and variance of a
distribution, the normal distribution, the binomial distribution, estimation and hypothesis testing, confidence intervals, and regression and correlation. Needed material from these topics is reviewed, but this review is not sufficient (too much material, too quickly) for a person having no previous knowledge of probability or statistics. Ideally, a person attending this course would have previously taken an elementary course in statistics. A person with no previous training in statistics could follow much of the lecture, but it would be expecting a lot to absorb the basic concepts of statistics or econometrics “on the fly,” in addition to the material specific to sample survey.

This course would be useful for a person who is familiar with techniques of sample survey design for descriptive surveys (used to estimate overall characteristics of populations and subpopulations, such as totals, means and proportions) and wishes to learn the basic concepts of analytical surveys (used to estimate relationships and to test hypotheses, such as whether a program has a positive impact).

As background material for the course, the following reference texts are recommended:


This course builds on material in other courses and articles of the author, including the course in *Sample Survey Design and Analysis: A Three-Day Course*, and the article *Sample Survey Design for Evaluation*. These are posted at [http://www.foundationwebsite.org](http://www.foundationwebsite.org).

The sample survey course addresses mainly the design and analysis of descriptive surveys (used for monitoring), with some coverage of analytical surveys (used for impact assessment). The article includes discussion of concepts in design of analytical surveys. The present course presents this material and also includes material on analysis of analytical surveys. Examples of statistical power analysis are presented using the author’s Microsoft Excel computer program, *JGCSampleSizeProgram*, posted at [http://www.foundationwebsite.org](http://www.foundationwebsite.org).

This course is intended to cover a broad range of topics in sample survey and econometric design and analysis. To do so, it does not cover each topic in great detail. The concern is with known results and how and when to apply them, not in proving them.

The course is introductory and conceptual, but somewhat technical, relatively comprehensive, and certainly intensive. At the end of the course, a person with some mathematical ability should be able to recognize which basic type of evaluation and sample design is appropriate in a given situation, be able to estimate the sample size required to produce a specified level of precision or power, and be able to conduct standard analyses of the collected sample data (using the *Stata* statistical program package).

There is no way, however, that a two-day course will make an “instant survey statistician” or “econometrician” out of anyone. In a survey design situation that is complex or that will involve large amounts of time, effort, or money, the advice of an expert sample survey statistician or econometrician should be sought.
The course is basically conceptual and discusses technical terms, but no time is spent on working through detailed examples, such as numerical calculation of formulas. Someone wishing to construct an actual evaluation or survey design and analyze the survey data would likely want to consult a reference text to review detailed examples and gain expertise by working through exercises.

This course is an ideal introduction for a project director or government technical (project) officer who wishes to understand the basic concepts of sample survey and evaluation design in order to effectively manage or monitor a project involving sample survey and econometric design and analysis. With the background of this course, the project manager should be able to sense what type of survey design is appropriate in a given situation, and be able to converse meaningfully with a consulting survey statistician or econometrician on a project involving a sample survey.

While it is possible to present a course on sample survey or econometrics with virtually no reference to mathematical symbology, such a course would not be of use to a person who actually wanted to design a survey and analyze survey data. This course is not a “no-math” course. While it is introductory and does not require extensive knowledge or use of calculus, it does require some familiarity with mathematics at the “college math” level, and a some background in introductory calculus and statistics. Persons with little mathematical background could attend the course and understand much of the conceptual material, but it is unlikely that they would be able to implement the methodology properly.

Nobody likes unpleasant surprises. One of the purposes for publishing these notes on the Internet is so that prospective students may quickly peruse them and assess whether the material is too advanced for them, given their present background in mathematics.

The course covers a lot of material in a short time. These notes will enable the student to pay attention to the lecture without having to take notes. It is not expected that everything will “sink in” in a two-day course, and it is recommended that the student who wishes to apply the techniques in practice acquire a reference text for study, or attend a formal course in which many homework exercises will “fix” the concepts.

Each attendee to the course is asked to complete a course evaluation. One of the questions asked is whether the course should spend less time on many topics (as it does), or concentrate on a smaller number of topics.

The course is comprehensive, but it is certainly not exhaustive. It provides an introduction to the major aspects of sample survey design and analysis in evaluation. There are many specialized topics that it does not cover, and it does not address every possible combination of survey design elements.

These notes are available for review by anyone considering enrolling in the course. The notes do not contain all of the exercises, examples, and handouts that may be included in the presentation.

It is not expected for the student to memorize the various formulas presented, but it is expected that some of the major ones would be familiar and recognized, by the end of the course (e.g., the formula for a double-difference measure and a double-difference estimator, and a general linear statistical model). A certain amount of course material (e.g., examples, supplementary material, details) is included as “background,” to place the essential concepts in context. It is
not expected that the student remember all of the material presented, and the really important concepts will be identified and stressed.

In a usual academic course, the material covered here would be written out by the professor, over the course of 48 one-hour class sessions. If all of the material covered here were written out, it would not be possible to cover it in a two-day course. Hence, in addition to obviating the need for taking notes, the course notes enable much more material to be covered than would be possible in a usual course. It is recognized that there is a learning advantage to the student’s writing his own notes, but this benefit has been sacrificed in order to cover much material in a short time. The material presented in the notes is available in a variety of reference texts on sample survey, experimental design and evaluation. *The essential feature of the course is the lecture and in-class interaction, not the notes.* The notes are made available simply to enable the student to take full advantage of these aspects.

The course lasts 12 hours. This is about 1/4 of the class time of a “three-unit” college semester (three hours per week for 16 weeks, or 48 hours). The college course, however, would include substantial amounts of homework, which this course does not include.

Sample survey and econometric modeling and analysis involve a lot of formulas. There are a number of different designs and estimation techniques, and each of them involves its own formulas (or procedures, such as resampling) for calculating estimates and standard errors of estimates. These course notes include a number of formulas, for reference, but no class time is spent in working with the formulas. They are too many and too complicated to learn well in a two-day course. The class time is spent in discussing concepts, examples, and approaches, not with working through complicated estimation formulas. A few detailed numerical examples may be worked out in the early part of the course, so that the student may become familiar with the computational requirements of the estimation formulas. After that, formulas will be shown in order to illustrate concepts and general forms, but no further calculations will be made using them. Discussion of statistical computer software for performing the calculations is included.

*Note on course content.* If presented on an advertised basis (individual enrollments), the course follows these notes closely. If presented for a single client, the content may be modified somewhat to suit the client’s interests.

The pace of the course, the selection of topics, and the time spent on various topics may be adjusted a little by the instructor, in order to address specific concerns or interests of the students.

While these notes parallel the lecture, not every item included in the notes is necessarily included in the lecture, and not every item included in the lecture is included in the notes. The notes are intended to reduce the requirement for the student to take copious notes during the lecture. They are not intended to be a detailed recording of the lecture. For additional detail and examples, the student should consult reference textbooks on sample survey design or econometric analysis.

This course covers both *estimation* (point and interval estimation) and *hypothesis testing.* It differs from a standard course on sample survey design, which is concerned mainly with sampling from finite populations and estimation of overall features of the population and subpopulations, rather than with estimation of relationships and tests of hypotheses about a process or conceptually infinite population.
Course Pricing

At present, the course is not given on an advertised basis, but only “in-house” at a client’s facility. The price is negotiated. The estimated price for the course, conducted over a two-day period at a client’s facility, is USD10,000 plus travel and lodging expense for two people. If the course is conducted over more days, the price is estimated at USD 5,000 per day plus travel and lodging expenses.

This price is all-inclusive, subject to the following limitations. Half payment is requested in advance, and half payment upon completion. Travel and per diem (meals, lodging and incidental) expense for the course staff are charged in accordance with US Government maximum travel per diem allowances (or international-organization allowances) for the travel (from presenter’s home base to client’s location, time spent at the client’s location, and return to the presenter’s home base).

It is agreed that the client will download the course notes from the Internet website http://www.foundationwebsite.org, and print sufficient copies for all attendees. Note: The Internet version of the course notes does not include handouts. These supplementary items, if available, will be e-mailed (as computer files) to the client prior to the course. If the client does not print the course notes and/or the supplementary items, the course will be presented without course notes and/or supplementary items. This is not the intended format, or the format that has been used successfully in the past for similar courses. As discussed, much material is presented, and it is not possible to write out this material during a two-day course. At the same time, restricting the course to a lecture, without benefit of the notes, would lose much. The course is intended to be a lecture supplemented with the Course Notes.

The client is expected to provide a comfortable environment conducive to learning. If the client does not have suitable accommodations at its own facility, it is recommended that facilities be procured at a local commercial hotel, many of which have excellent facilities for seminars. It is requested that the client provide a computer (with a Microsoft Windows operating system), computer-driven projector and projection screen, for displaying the Course Notes. It is also requested that a medium be provided for ad-hoc classroom presentation by the lecturer. For small groups, this may be a wall board (with chalk or markers) or “flip-chart-and-easel” (with marking pen). For larger groups it is recommended that a “view-graph” projector be available (for displaying writing using markers on clear acetate sheets).

It is requested that the client provide snacks and drinks for the breaks. The client is encouraged to provide lunch to presenters and attendees for full-day sessions, but this is at the client’s discretion. (This was the practice when a similar course was presented on an advertised basis at a commercial hotel, and it works well (it keeps the class together, and avoids late returns to class after lunch).)

2. Course Schedule

Course Schedule

9:00 - 9:20  Introduction; Course Objectives and Outline; Overview of First Day's Course Content
9:20 -10:00  Causal Models
10:00 -10:30  Impact Measures and Impact Estimators
10:30 -10:45  Break
10:45 -12:00  Basic Evaluation Design / Models / Estimators
12:00 - 1:00  Lunch
1:00 - 2:30  Sample Survey Design for Evaluation (Principles of Experimental Design; Descriptive and Analytical Surveys; Sample Survey Design Techniques)
2:30 - 2:45  Break
2:45 - 4:00  Sample Survey Design for Evaluation (Matching; Estimation of Sample Size; Sample Selection Procedures); Survey of References; Outline of Topics for Second Day; Questions and Answers

Day 2: Impact Analysis

9:00 - 9:15  Overview of Second Day's Course Content
9:15 - 9:45  Parametric Estimation and Hypothesis Testing (General Linear Statistical Model (Regression Model); Logistic Regression Model)
9:45 - 10:30  Impact Analysis
10:30 -10:45  Break
10:45 -12:00  Impact Analysis
12:00 - 1:00  Lunch
1:00 - 2:30  Impact Analysis
2:30 - 2:45  Break
2:45 - 3:30  Analysis of Baseline Data
3:30 - 4:00  Computer Software for Analysis of Survey Data; Questions and Answers
3. Course Syllabus

**Course Syllabus**

**Day 1:** Causal Modeling, Impact Measures and Estimators, Basic Evaluation Designs, and Principles and Techniques of Design and Analysis for Evaluation

1. Introduction
   - Course Objectives and Outline
   - Overview of First Day's Course Content

2. Causal Modeling
   - Causal Modeling (Pearl Graphical Approach to Causal Modeling)
   - Causal Relationship; Causal Structure (Causal Graph, Path Diagram, Directed Acyclic Graph)
   - Equational Model (Structural Equation Model)
   - Causal Effect
   - Potential Outcomes, Counterfactuals, Neyman-Fisher-Cox-Rubin Causal Model

3. Impact Measures
   - Review of Impact Measures
   - Counterfactual Model
   - Matching and Regression Analysis

4. Basic Evaluation Designs / Models / Estimators
   - Designs Based on Randomization
   - Two-Group Designs
   - Interrupted Time-Series Design
   - Four-Group Designs
   - Quasi-Experimental Designs
   - Observational Data

5. Sample Survey Design for Evaluation
   - Principles of Experimental Design
   - Descriptive Surveys vs. Analytical Surveys
   - Sample Survey Design Techniques
   - Use of Matching to Improve Precision and Power
   - Estimation of Sample Size for Descriptive Surveys
   - Estimation of Sample Size for Analytical Surveys
   - Statistical Tests of Hypothesis
   - Sample Selection Procedures

6. Survey of References; Outline of Topics for Second Day; Questions and Answers

**Day 2:** Parametric Estimation, Analysis of Baseline Data, Impact Analysis, Use of Stata Statistical Program Package

1. Overview Of Second Day's Course Content
Parametric Estimation and Hypothesis Testing (General Linear Statistical Model, Logistic Regression Model; Two-Step M-Estimators)
Analysis of Baseline Data (Descriptive Surveys, Analytical Surveys)
Use of Stata Statistical Program Package (Summary Description and Examples)
Survey of References
Questions and Answers

2. Parametric Estimation and Hypothesis Testing
   - Properties of Estimators
   - Estimation Procedures
   - General Linear Statistical Model
   - Generalized Linear Statistical Model
   - Principle of Conditional Error

3. Impact Analysis
   - Design-Based Estimates
   - Model-Based Estimates
   - Model Specification
   - Use of Matching to Reduce Model Dependence
   - Estimators Based on Ignorability (Conditional Independence)
   - Selection vs. Outcome Models (Two-Step Models)
   - Estimation of Standard Errors
   - Estimation Based on Combining Regression and Propensity-Score Matching (Matching Estimators; Doubly Robust Matching/Regression Approach)
   - Instrumental-Variable Estimators
   - Local Average Treatment Effect (LATE), Intention-to-Treat Effect (ITT)
   - Regression-Discontinuity Design
   - Responses Discrete or in a Limited Range
   - Multivalued (Categorical) Treatment Variable
   - Multiple Treatment Variables
   - Panel Data (Fixed-Effects Model, Random-Effects Model)
   - Treatment of Missing Values
   - Use of Weights
   - Estimation of Standard Errors

4. Analysis of Baseline Data
   - Descriptive Surveys
   - Analytical Surveys
5. Use of *Stata* Statistical Program Package
   - Summary Description of *Stata*
   - Examples

6. Survey of References; Questions and Answers
4. Course Critique Form

Course Critique Form

Dear Participant:

We appreciate your attendance and are interested in your comments in order to improve our course. Please answer the following questions, adding additional comments as necessary, and send the form back in the attached envelope. Thank you.

Date of course ___________________ Location of course _________________________________

Course Content

1. How useful do you consider the information? ________________________________
2. Was the material presented in sufficient detail? ________________________________
3. Were there some topics you would have preferred more discussion on? Yes___ No___
   If so, which ones? ____________________________________________________________

Course Delivery

1. Were the presentations effective? ______________________________________________
2. Were the visual aids helpful? _________________________________________________
3. Were the Course Notes sufficiently detailed? ________________________________

Facilities

1. Was the seating arrangement satisfactory? _______________________________________
2. Were the meals satisfactory? _________________________________________________
3. Was parking adequate? ______________________________________________________
4. Is the location convenient? _________________________________________________

General

1. How did you find out about this course? _______________________________________
   Brochure in mail ________________
   Organizational channels __________
   Associate _______________________
   Internet _________________________
   Other (specify) __________________

2. Did you have sufficient registration time? _______________________________

3. Did you feel the course was as you expected it to be, from the flyer? ______________

4. Did you feel the course was as you expected it to be, from the Course Notes (if examined on the Internet)? ______________

5. If from out of town: Did you stay at the hotel where the course was presented? _____
6. This course was presented to provide a broad overview of Statistical Design and Analysis in Evaluation. Would you have preferred to concentrate on fewer topics?

7. Have you ever attended a course on Statistical Design and Analysis in Evaluation before?
   Yes____   No____

8. Would you prefer a more detailed course of 3-4 days,____
   or a less detailed course of 1 day?____

9. Would you prefer a more advanced course,____
   or a less advanced course?____

10. Compared to other short courses of which you are familiar, was the cost of this course:
    About right________
    Rather high________
    Lower than expected____

11. What additional seminars might you be interested in?
    Sample Survey Design and Analysis____
    Time Series Analysis, Forecasting and Control____
    Biostatistics________
    Experimental Design____
    Quality Control____
    Evaluation Research____
    Introduction to Statistics and Data Analysis________
    Simulation and Modeling________
    Optimization________
    Other (specify)________

Additional Comments:

Name (optional)______________________________________________________________

Organization (optional)________________________________________________________
5. Résumé of Course Developer

Joseph George Caldwell, Ph.D. (Statistics)

Professional Profile: Career in management consulting, system development, research, and teaching. Directed projects in strategic planning, policy analysis, program evaluation, economics, public finance, statistics, operations research / systems analysis, and information technology for US, state and foreign governments, and US and foreign organizations. Areas of expertise include health, education, vocational rehabilitation, welfare, public finance (tax policy analysis, Medicaid and AFDC financing), agriculture, civil rights, economic development, energy, environment, population, and defense (US Army, Navy, Air Force, Department of Defense). Considerable overseas experience.

2005- Semi-retired. Consultant in Statistics and Information Technology, Spartanburg, South Carolina, and Tucson, Arizona, USA. Consultant on research design for impact evaluation of development projects (assignments in Honduras, Ghana, Burkina Faso, Namibia and other African countries funded mainly by Millennium Challenge Corporation); also computer / communication system development work in Timor Leste funded by UNDP and in Guinea and Liberia funded by US Agency for International Development).


1964-1982 Consultant or employee to firms in South Carolina, North Carolina, Virginia, Maryland, District of Columbia, Haiti, Philippines.


Personal: Born March 23, 1942, in Kingston, Ontario, Canada. Nationality: United States of America, Canada. Author of articles and books on divers topics (e.g., population, environment, economics, politics, defense and music, including The Late Great United States (2008); Can America Survive? (1999); The Value-Added Tax: A New Tax System
for the United States (1987); How to Play the Guitar by Ear (for mathematicians and physicists) (2000). See Internet website http://www.foundationwebsite.org to view these and other articles.

Contact information: 1432 N Camino Mateo, Tucson, AZ 85745-3311USA, e-mail jcalwell9@yahoo.com
6. Day 1: Causal Modeling

**DAY 1: CAUSAL MODELING, IMPACT MEASURES AND ESTIMATORS, BASIC EVALUATION DESIGNS, AND PRINCIPLES AND TECHNIQUES OF DESIGN AND ANALYSIS FOR EVALUATION**

6.1. Overview of First Day's Course Content

INTRODUCTION; COURSE OBJECTIVES AND OUTLINE; OVERVIEW OF FIRST DAY'S COURSE CONTENT

- CAUSAL MODELING (PEARL GRAPHICAL APPROACH TO CAUSAL MODELING; POTENTIAL OUTCOMES, COUNTERFACTUALS, NEYMAN-FISHER-COX-RUBIN CAUSAL MODEL)
- IMPACT MEASURES
- BASIC EVALUATION DESIGNS / MODELS / ESTIMATORS
- SAMPLE SURVEY DESIGN FOR EVALUATION
- SURVEY OF REFERENCES; OUTLINE OF TOPICS FOR SECOND DAY; QUESTIONS AND ANSWERS

6.2. Causal Modeling

**CAUSAL MODELING**

GOAL: TO ASSESS THE CAUSAL EFFECT (CAUSAL IMPACT, IMPACT) OF A PROJECT OR PROGRAM

CAUSAL EFFECT: DEFINED WITH RESPECT TO A CAUSAL MODEL. REFERENCES:


Additional references on structural equation modeling:
WHY THE INTEREST IN CAUSAL MODELING? MODEL IDENTIFICATION. (IS THE MODEL “IDENTIFIED”?) THE CORRECT SPECIFICATION OF A MODEL (VARIABLES, RELATIONSHIPS, PARAMETER VALUES) CANNOT BE INFERRED FROM SAMPLE DATA. THE MODEL PARAMETERS CANNOT BE ESTIMATED UNLESS CERTAIN TECHNICAL ASSUMPTIONS HOLD, AND THOSE ASSUMPTIONS FLOW FROM THE CAUSAL MODEL. ESTIMABILITY OF MODEL PARAMETERS IS DETERMINED BY THE CAUSAL MODEL.

THE CAUSAL MODEL SHOWS, FOR EXAMPLE, WHICH VARIABLES OF A MODEL ARE ENDOGENOUS (AND REQUIRE SPECIAL TREATMENT IN REGRESSION ANALYSIS), AND WHICH VARIABLES MAY SERVE AS INSTRUMENTAL VARIABLES FOR THEM.

NOTE: SOME VARIATION IN TERMINOLOGY. STATISTICIANS OFTEN USE THE TERM “CORRECTLY SPECIFIED” OR “ESTIMABLE” INSTEAD OF “IDENTIFIED.”

FOR PROGRAM EVALUATION APPLICATIONS, WE ASK: WHAT IS EFFECT OF A CAUSE, NOT WHAT IS THE CAUSE OF AN EFFECT.

ASSOCIATIONAL RELATIONSHIP: A VARIABLE X HAS AN ASSOCIATIONAL RELATIONSHIP TO A VARIABLE Y IF THE PROBABILITY DISTRIBUTION OF Y CONDITIONAL ON X DIFFERS FOR DIFFERENT VALUES OF X.

A CAUSAL RELATION DIFFERS FROM AN ASSOCIATIONAL RELATION

“CORRELATION IS NOT CAUSATION.” “NO CAUSATION WITHOUT MANIPULATION (INTERVENTION).” TO ESTIMATE THE EFFECT OF CHANGES TO A VARIABLE IN A SYSTEM, ONE MUST MAKE FORCED CHANGES TO THE VARIABLE.

CAUSAL MODELING SUBSUMES (INCLUDES) STATISTICAL MODELING. STATISTICS TEXTS AVOID MENTION OF CAUSALITY (CAUSAL, CAUSAL EFFECT), PREFERRING “EFFECT.” CAUSALITY MAY BE INFERRED FROM EXPERIMENTAL DESIGNS HAVING RANDOMIZED ASSIGNMENT TO TREATMENT. IN OTHER CASES, NEED A CAUSAL MODEL AS A BASIS FOR MAKING CAUSAL INFERENCES. THE CAUSAL MODEL IS USED
TO GUIDE MODEL IDENTIFICATION, AND STATISTICAL INFERENCE IS USED TO ESTIMATE MODEL PARAMETERS FOR AN IDENTIFIED MODEL.

DIFFERENCES IN THE ROLES OF CAUSAL MODELING AND STATISTICAL ANALYSIS. CAUSAL MODELING ASSISTS MODEL IDENTIFICATION. STATISTICAL ANALYSIS IS USED TO ESTIMATE PARAMETERS IN IDENTIFIED MODELS.

DIFFERENCES IN DESCRIPTIVE SURVEY DESIGN AND ANALYSIS AND ANALYTICAL SURVEY DESIGN AND ANALYSIS

DESCRIPTIVE SURVEY APPROACH: ESTIMATE CHARACTERISTICS (TOTALS, MEANS, PROPORTIONS) OF AN EXISTING FINITE POPULATION AND SUBPOPULATIONS. HAS NOTHING TO DO WITH A CAUSAL MODEL, AND DOES NOT NEED A CAUSAL MODEL. NOT CONCERNED WITH INFERENCES ABOUT CAUSE-AND-EFFECT RELATIONSHIPS.

THE POPULATION IS REGARDED AS FIXED. THE RANDOM VARIABLE OF INTEREST IS THE PROBABILITY OF SELECTION OF UNITS FOR THE SAMPLE. CONCERNED WITH PRECISION OF ESTIMATES (STANDARD ERRORS AND CONFIDENCE INTERVALS). IF SAMPLE ENTIRE POPULATION, STANDARD ERRORS ARE ZERO (FINITE POPULATION CORRECTION, OR FPC). DESIGN-BASED ESTIMATES.

ANALYTICAL SURVEY APPROACH: ESTIMATE CHARACTERISTICS OF A PROCESS, OR CAUSAL EFFECT OF AN INTERVENTION; RELATIONSHIP OF OUTCOME VARIABLES TO INPUT VARIABLES; TESTS OF HYPOTHESIS ABOUT OVERALL IMPACT, OR RELATIONSHIP OF IMPACT TO EXPLANATORY VARIABLES. THE RANDOM VARIABLES OF INTEREST ARE ALL VARIABLES OF A MODEL DESCRIBING THE PROCESS OF INTEREST. CONCERNED WITH POWER OF TESTS OF HYPOTHESIS ABOUT PROCESS CHARACTERISTICS (E.G., IS EFFECT OF TREATMENT INTERVENTION POSITIVE). CONCEPTUALLY INFINITE POPULATION: EVEN IF SAMPLE ENTIRE PHYSICAL POPULATION, STILL HAVE SAMPLING VARIATION. MODEL-BASED OR MODEL-ASSISTED ESTIMATES. INCORPORATES ASPECTS OF EXPERIMENTAL DESIGN.

ANALYTICAL SURVEY DESIGN AND ANALYSIS IS SUBSTANTIALLY MORE DIFFICULT THAN DESCRIPTIVE SURVEY DESIGN AND ANALYSIS. IT REQUIRES CONSIDERATION OF CAUSAL MODELS, AND INCORPORATES ELEMENTS OF EXPERIMENTAL DESIGN (THE FIELD OF DESIGN OF EXPERIMENTS) AND ECONOMETRICS. THIS TWO-DAY COURSE IN STATISTICAL DESIGN AND ANALYSIS IN EVALUATION WILL DESCRIBE THE ESSENTIAL BASIC CONCEPTS AND PROVIDE SOME DETAIL FOR MOST-USED TECHNIQUES.

REFERENCES ON DESCRIPTIVE SURVEY DESIGN AND ANALYSIS:


REFERENCES ON ANALYTICAL SURVEY DESIGN AND ANALYSIS:


REFERENCES ON EXPERIMENTAL DESIGN:


REFERENCES ON CAUSAL MODELING


REFERENCE ON CAUSAL ANALYSIS (IN THE PARTICULAR FIELD OF ECONOMETRICS):


CAUSAL MODEL; CAUSAL EFFECT; POTENTIAL OUTCOMES, COUNTERFACTUALS

EARLY DEVELOPMENT: CHARLES SPEARMAN (FACTOR ANALYSIS, 1904), SEWELL WRIGHT (PATH ANALYSIS, 1918), O. D. DUNCAN (STRUCTURAL EQUATION MODELING, 1966).

TWO MAIN KINDS OF CAUSAL MODELS:

RECURSIVE (NO FEEDBACK LOOPS, CAUSAL EFFECTS UNIDIRECTIONAL, DISTURBANCES UNCORRELATED; IDENTIFIED; ORDINARY LEAST SQUARES)

NONRECURSIVE (FEEDBACK LOOPS, CORRELATED DISTURBANCES; SOME NONIDENTIFIED; INDIRECT LEAST SQUARES, TWO-STAGE LEAST SQUARES, FULL INFORMATION MAXIMUM LIKELIHOOD)
(THE TERM "RECURSIVE" IS MISLEADING – IT REFERS TO THE METHOD OF SOLVING STRUCTURAL EQUATIONS, NOT TO THE CAUSAL FLOW IN A PATH DIAGRAM.

(A RECURSIVE MODEL IS HIERARCHICAL: ALL ENDOGENOUS VARIABLES CAN BE ARRANGED IN A SEQUENCE $X_1, X_2, \ldots, X_m$ SUCH THAT FOR ANY $X_i$ AND $X_j$ WHERE $i < j$, $X_j$ IS NOT A CAUSE OF $X_i$. HENCE THE FIRST ENDOGENOUS VARIABLE IS INFLUENCED ONLY BY EXOGENOUS VARIABLES, THE SECOND IS INFLUENCED ONLY BY EXOGENOUS VARIABLES AND THE FIRST ENDOGENOUS VARIABLE, AND SO ON.)

SOME DESCRIPTION OF CAUSAL MODELING, IN THE RECURSIVE CASE (THE FOLLOWING IS DERIVED FROM PEARL OP. CIT.)

CAUSAL RELATIONSHIP: IF THE PROBABILITY DISTRIBUTION OF Y CHANGES WHEN A RANDOMLY-TIMED FORCED CHANGE IS MADE IN X, THEN WE SAY THAT THERE IS A CAUSAL RELATIONSHIP BETWEEN X AND Y (AND THAT X HAS A CAUSAL INFLUENCE ON Y).

NOTE THAT THE CAUSAL MODEL IS NOT INFERRED FROM PROBABILISTIC ASSOCIATIONS. CAUSAL RELATIONSHIPS ARE SPECIFIED IN A CAUSAL MODEL. THE STRENGTH OF THOSE RELATIONSHIPS MAY BE SPECIFIED BY PROBABILITY DISTRIBUTIONS, BUT THE CAUSAL RELATIONSHIPS ARE NOT DERIVED FROM ASSOCIATIONAL RELATIONSHIPS. THE CAUSAL MODEL PROVIDES INFORMATION ON HOW CAUSAL EFFECTS SHOULD BE ESTIMATED (I.E., ON ESTIMABILITY, OR IDENTIFIABILITY).

NOTATION: DENOTE RANDOM VARIABLES BY UPPER-CASE LETTERS (E.G., X, Y, U, V, W). DENOTE SPECIFIC REALIZATIONS OF RANDOM VARIABLES BY LOWER-CASE LETTERS (E.G., x, y, u, v, w). WILL USUALLY DENOTE VECTORS USING UNDERSCORES OR BOLD FACE (E.G., $x' = (x_1, x_2, \ldots, x_n)$ OR $x' = (x_1, x_2, \ldots, x_n)$, BUT OCCASIONALLY $x = (x_1, x_2, \ldots, x_n)$. WILL USUALLY DENOTE COLUMN VECTORS WITHOUT PRIMES AND ROW VECTORS WITH PRIMES.

CAUSAL STRUCTURE (CAUSAL GRAPH, PATH DIAGRAM): A CAUSAL STRUCTURE OF A SET OF VARIABLES IS A DIRECTED ACYCLIC GRAPH (DAG) IN WHICH EACH NODE REPRESENTS A VARIABLE AND EACH LINK REPRESENTS A CAUSAL RELATIONSHIP BETWEEN THE LINKED VARIABLES.

EXAMPLE OF DAG [SHOW GRAPH].

\[ z \rightarrow w \rightarrow y \]

where

$z =$ education on exercise

$w =$ exercise

$y =$ health

or
\( z = \) random assignment to treatment

\( w = \) treatment

\( y = \) outcome

\( z \) AFFECTS \( w \), BUT NOT \( y \) DIRECTLY. \( z \) IS CALLED AN INSTRUMENT FOR \( w \).

**CAUSAL MODEL:** THE PROBABILITY DISTRIBUTIONS ASSOCIATED WITH A CAUSAL STRUCTURE. THE PROBABILITY DISTRIBUTION FOR VARIABLE (NODE) \( x \) IS CONDITIONAL ONLY ON ITS PARENT VARIABLES (NODES), AND, CONDITIONAL ON THE PARENT VALUES, \( x \) IS INDEPENDENT OF ALL OTHER VARIABLES. (NOTE: PEARL DEFINES A DETERMINISTIC CAUSAL MODEL, AND THEN A PROBABILISTIC CAUSAL MODEL.)

**MORE SPECIFICALLY,** EQUATIONAL MODEL (STRUCTURAL EQUATION MODEL):

FOR EACH VARIABLE (NODE) \( x_i, x_i = f_i(p_{a_i}, u_i) \), \( i = 1, \ldots, n \), WHERE \( f_i \) DENOTES A FUNCTIONAL RELATIONSHIP, \( p_{a_i} \) DENOTES THE VALUES OF PARENT VARIABLES, \( u_i \) DENOTE BACKGROUND (EXOGENOUS) VARIABLES, AND \( P(u) \) DENOTES THE JOINT PROBABILITY DISTRIBUTION OF THE \( u \), WHICH ARE ASSUMED TO BE INDEPENDENT. WE DO NOT KNOW THE \( f_i \) OR \( P \). (\( u \) MAY DENOTE A SINGLE \( u_i \) OR A VECTOR OF ALL \( u_i \).)

**NOTATION:** U DENOTES THE SET OF ALL BACKGROUND (EXOGENOUS) VARIABLES, AND V DENOTES THE SET OF ALL VARIABLES DETERMINED BY VARIABLES IN THE MODEL (I.E., ENDOGENOUS VARIABLES). LOWER-CASE LETTERS ARE REALIZATIONS OF UPPER-CASE RANDOM VARIABLES.

**TERMINOLOGY:**

- **ENDOGENOUS VARIABLES:** VARIABLES THE CAUSES OF WHICH ARE EXPLICITLY REPRESENTED IN THE CAUSAL MODEL
- **PREDETERMINED VARIABLES:** VARIABLES THE CAUSES OF WHICH ARE NOT EXPLICITLY REPRESENTED IN THE MODEL
- **LAGGED ENDOGENOUS VARIABLES** HAVE VALUES EQUAL TO THE VALUES OF ENDOGENOUS VARIABLES AT PREVIOUS TIMES
- **EXOGENOUS VARIABLES** (BACKGROUND VARIABLES): VALUES DETERMINED OUTSIDE THE MODEL (NOT PRIOR VALUES OF ENDOGENOUS VARIABLES)

THE EQUALITY SIGNS IN THE EQUATIONAL MODEL IMPLY “IS DETERMINED BY.” (CAUSATION, NOT SIMPLE ASSOCIATION.)

**NOTE:** THE MODEL IS MARKOVIAN, I.E., THE DISTRIBUTION DEPENDS ONLY ON THE VALUES OF THE PARENT VARIABLES, AND NO PREVIOUS ANCESTOR NODES. FOR DIFFERENT POPULATION UNITS, THE RESPONSES ARE INDEPENDENT, GIVEN THE PARENT VARIABLES. THIS IS CALLED THE STABLE UNIT TREATMENT VALUE ASSUMPTION (SUTVA, THE NO-MACRO-EFFECTS ASSUMPTION, OR THE PARTIAL-EQUILIBRIUM ASSUMPTION). (FOR Example, individuals do not compete for resources.)
MORE COMPLICATED EXAMPLE OF STRUCTURAL EQUATIONS. PRICE AND DEMAND EQUATIONS:

\[ q = b_1 p + d_1 i + u_1 \]
\[ p = b_2 q + d_2 w + u_2 \]

where

- \( q \) = quantity of household demand for a product
- \( p \) = unit price of product
- \( i \) = household income
- \( w \) = wage rate for producing the product
- \( u_i \) = error terms.

CAUSAL MODELING = PATH ANALYSIS, LATENT-VARIABLE MODELING. STRUCTURAL EQUATION MODELING (SEM) (ALTERNATIVE REPRESENTATIONS OF CAUSAL INFORMATION – EQUATIONS RATHER THAN GRAPHS). IN SOME FIELDS (PROCESS CONTROL, BIOLOGY, PSYCHOLOGY), MODELS MAY BE COMPLEX.

IN THE FIELD OF EVALUATION IN INTERNATIONAL DEVELOPMENT, THE CAUSAL MODELS ARE USUALLY SIMPLE, INVOLVING VARIABLES THAT AFFECT SELECTION FOR TREATMENT, A TREATMENT VARIABLE (USUALLY A SINGLE BINARY VARIABLE), AND A NUMBER OF COVARIATES. CONSIDERATION OF THE CAUSAL MODEL STRUCTURE (DIAGRAM) INDICATES VALID STATISTICAL MODELS TO USE FOR PARAMETER ESTIMATION (E.G., INSTRUMENTAL VARIABLES).

EXAMPLE: DIRECT FEEDBACK LOOP (SHOW GRAPH).

EXAMPLE: PANEL MODEL FOR LONGITUDINAL DATA (SHOW GRAPH)

RECIPIROCAL CAUSATION REPRESENTED BY CROSS-LAG DIRECT EFFECTS BETWEEN RESPONSE AND EXPLANATORY VARIABLES MEASURED AT DIFFERENT TIMES. NO DIRECT FEEDBACK LOOP.

NOTATION: WILL GENERALLY (BUT NOT ALWAYS) USE UPPER-CASE LETTERS (X, Y, Z...) TO DENOTE RANDOM VARIABLES (FUNCTIONS DEFINED ON A SAMPLE SPACE), AND LOWER-CASE LETTERS TO DENOTE SPECIFIC REALIZATIONS (NUMERICAL VALUES, \( x, y, z... \)). WILL USE GREEK LETTERS TO DENOTE POPULATION VALUES (“TRUE” VALUES), AND CARETS OR BARS TO DENOTE SAMPLE ESTIMATES.

CAUSAL EFFECT: GIVEN TWO DISJOINT SETS OF VARIABLES, X AND Y, THE CAUSAL EFFECT OF X ON Y, DENOTED \( P(y|\xi) \) OR AS \( P(y|do(x)) \), IS A FUNCTION FROM X TO THE SPACE OF PROBABILITY DISTRIBUTIONS ON Y. FOR EACH REALIZATION \( x \) OF X, \( P(y|do(x)) \) GIVES THE PROBABILITY OF \( Y = y \) INDUCED BY DELETING FROM THE EQUATIONAL MODEL ALL EQUATIONS CORRESPONDING TO VARIABLES IN X AND SUBSTITUTING \( X = x \) IN THE REMAINING EQUATIONS. (ALTERNATIVELY (ROSENBAUM AND RUBIN), THE CAUSAL EFFECT IS THE DIFFERENCE \( \epsilon(Y|do(x_1)) - \epsilon(Y|do(x_2)) \), WHERE \( x_1 \) AND \( x_2 \) ARE TWO DISTINCT REALIZATIONS OF X.)
COMPARE STATISTICAL $P(y|see \ x)$ TO CAUSAL $P(y|do(\ x))$. CAUSAL MODEL IS ONTOLOGICAL (RELATES TO HOW THINGS ARE), STATISTICAL MODEL IS EPISTEMOLOGICAL (RELATES TO HOW THINGS APPEAR (BELIEF, KNOWLEDGE)).

EVERYTHING MAY HAVE A CAUSE, BUT NOT EVERYTHING CAN BE A CAUSE (ATTRIBUTES ARE NEVER CAUSES).

TO ESTABLISH CAUSALITY REQUIRES EXPERIMENTATION (RANDOM, FORCED CHANGES IN EXPLANATORY (“PARENT”) VARIABLES).

PHILOSOPHICALLY (METAPHYSICALLY), A TIME VARIABLE IS REQUIRED TO ESTABLISH CAUSALITY; CAUSAL MODEL ENABLES PREDICTION. CAUSALITY MAY BE INFERRED FROM EXPERIMENTS (E.G., USING AN EXPERIMENTAL DESIGN IN WHICH TREATMENT IS RANDOMLY ASSIGNED).

ALTERNATIVE MODELS: MODEL PREFERENCE:

CAUSAL INFLUENCE: A VARIABLE X IS SAID TO HAVE A CAUSAL INFLUENCE ON A VARIABLE Y IF A DIRECTED PATH FROM X TO Y EXISTS IN EVERY MINIMAL STRUCTURE CONSISTENT WITH THE DATA. (A MINIMAL THEORY IS ONE THAT IS SIMPLER THAN OTHERS IN SOME SENSE, AND CONSISTENT WITH THE DATA.) OCCAM’S RAZOR: PREFER SIMPLER EXPLANATIONS (POPPER: SIMPLER MODEL LESS EASY TO FALSIFY, FEWER OPPORTUNITIES FOR OVERFITTING THE DATA.)

COUNTERFACTUALS MODEL (POTENTIAL OUTCOMES MODEL; RUBIN CAUSAL MODEL; NEYMAN-RUBIN CAUSAL MODEL; NEYMAN-FISHER-COX-RUBIN CAUSAL MODEL)

A SUBMODEL, $M_x$, OF A CAUSAL MODEL, CORRESPONDING TO A PARTICULAR REALIZATION $x$ IN X, IS THE CAUSAL MODEL DETERMINED BY DELETING ALL FUNCTIONS $f_i$ CORRESPONDING TO MEMBERS OF THE SET X AND REPLACING THEM WITH THE SET OF CONSTANT FUNCTIONS $X = x$.

EFFECT OF ACTION (CAUSAL EFFECT): THE EFFECT OF ACTION $do(X = x)$ ON A CAUSAL MODEL M IS GIVEN BY THE SUBMODEL $M_x$.

POTENTIAL RESPONSE: LET X AND Y BE TWO SUBSETS OF VARIABLES IN V. THE POTENTIAL RESPONSE OF Y TO ACTION $do(X = x)$ IS THE SOLUTION, $Y_x(u)$ (OR $Y(x,u)$), FOR Y IN THE SET OF EQUATIONS $f_i$ IN THE SUBMODEL $M_x$.

COUNTERFACTUAL: LET X AND Y BE TWO SUBSETS OF VARIABLES IN V. THE SENTENCE “THE VALUE THAT Y WOULD HAVE OBTAINED, HAD X BEEN x” IS INTERPRETED AS DENOTING THE POTENTIAL RESPONSE $Y_x(u)$.

THE TERM COUNTERFACTUAL IS OFTEN MISUSED IN PLACE OF THE TERM POTENTIAL OUTCOME. POTENTIAL OUTCOMES ALWAYS EXIST IN CONCEPT. A COUNTERFACTUAL EXISTS ONLY AFTER ONE OF THE POTENTIAL OUTCOMES HAS BEEN PHYSICALLY REALIZED.

A CAUSAL MODEL IS A BASIS FOR PREDICTION. THE POTENTIAL OUTCOMES / COUNTERFACTUAL TERMINOLOGY IS USEFUL FOR EXPLANATION, BUT IS A DIFFICULT APPROACH TO CAUSAL MODEL SPECIFICATION. IT IS NOT AS COMPLETE A
FRAMEWORK AS THE PEARL APPROACH (I.E., MAY FAIL TO CERTIFY SOME VALID CONCLUSIONS). IT ENCOUNTERS DIFFICULTIES FROM A QUANTUM-THEORETIC VIEWPOINT, AND IS REJECTED BY SOME (E.G., DAWID). IN THE NEYMAN-RUBIN FRAMEWORK, \( Y_x(u) \) IS VIEWED AS A PRIMITIVE (UNDEFINED QUANTITY), WHEREAS IN THE PEARL FRAMEWORK, IT IS A DERIVED QUANTITY. A SIGNIFICANT ADVANTAGE OF THE PEARL FRAMEWORK IS THAT IT PROVIDES A BASIS FOR DETERMINING (BY EXAMINING THE DAG) WHICH INFERENCES ARE IDENTIFIABLE (ESTIMABLE) ("INTERVENTION CALCULUS," GRAPHICAL TESTS OF IDENTIFIABILITY, NOT ADDRESSED HERE). THE COUNTERFACTUALS APPROACH REQUIRES ASSUMPTIONS ABOUT THE CONDITIONAL INDEPENDENCE OF COUNTERFACTUAL OUTCOMES,

CONDITIONAL INDEPENDENCE: X AND Y ARE SAID TO BE CONDITIONALLY INDEPENDENT GIVEN Z IF \( P(x|y,z) = P(x|z) \) WHENEVER \( P(y,z)>0 \).

NOTATION (CONDITIONAL INDEPENDENCE, MUCH USED IN COUNTERFACTUALS APPROACH):

\[ y \perp w | x : y \text{ IS (CONDITIONALLY) INDEPENDENT OF } w \text{ GIVEN } x \]

\[ y \perp w | x: y \text{ IS UNCORRELATED WITH (ORTHOGONAL TO) } w \text{ GIVEN } x \]

ADJUSTMENT FOR OTHER VARIABLES: COVARIATES, CONCOMITANTS, AND CONFOUNDERS

X AND Y ARE CONFOUNDED WHEN THERE IS A THIRD VARIABLE Z THAT INFLUENCES BOTH X AND Y (Z IS CALLED A CONFOUNDER OF X AND Y).

TWO VARIABLES X AND Y ARE NOT CONFOUNDED IF AND ONLY IF EVERY VARIABLE Z THAT IS NOT AFFECTED BY X IS EITHER (1) UNASSOCIATED WITH X; OR (2) UNASSOCIATED WITH Y, CONDITIONAL ON X.

SIMPSON’S PARADOX: ANY STATISTICAL RELATIONSHIP BETWEEN TWO VARIABLES MAY BE REVERSED BY INCLUDING ADDITIONAL FACTORS IN THE MODEL. FOR EXAMPLE (FROM PEARL), “…WE MAY FIND THAT STUDENTS WHO SMOKE OBTAIN HIGHER GRADES THAN THOSE WHO DO NOT SMOKE BUT, ADJUSTING FOR AGE, SMOKERS OBTAIN LOWER GRADES IN EVERY AGE GROUP AND, FURTHER ADJUSTING FOR FAMILY INCOME, SMOKERS AGAIN OBTAIN HIGHER GRADES THAN NONSMOokers IN EVERY INCOME-AGE GROUP, AND SO ON.”

(TECHNICAL DEFINITION OF SIMPSON’S PARADOX:

\[ E(y|w=1) – E(y|w=0) > (\) 0, while \( E(y|x,w=1) – E(y|x,w=0) < (\) 0 for all x.)

THE ADJUSTMENT PROBLEM: WHAT COVARIATES SHOULD BE INCLUDED IN A MODEL? CANNOT BE DETERMINED FROM DATA – RESTS ON ASSUMPTIONS.

THE POTENTIAL-OUTCOMES APPROACH. AS PRESENTED BY ROSENBAUM AND RUBIN, HAS SOME PROBLEMS. THE COUNTERFACTUAL (THE VALUE THAT \( Y \) WOULD HAVE BEEN, HAD \( X \) BEEN \( x \)) IS TAKEN AS A PRIMITIVE, NOT A DERIVED CONSTRUCT. THAT APPROACH DOES NOT SPECIFY A MATHEMATICAL MODEL FROM WHICH CAUSAL CONSISTENCY MAY BE ESTABLISHED, OR FOR ADDRESSING THE ISSUE OF
Completeness (whether the relationships at hand are sufficient for determining the validity of all inferences).

Ignorability: “Z is an admissible set of covariates if, given Z, the value that Y would obtain had X been x is independent of X.” Counterfactuals are not observable, and assessment of conditional independence is difficult.

Pearl presents a method, based on analysis of causal graphs, for solving the adjustment problem.

Testing for identification. Pearl presents graphical tests of identifiability (of causal effects): Back-door criterion, Front-door criterion. Methodology includes analysis of primitive interventions (do(x)) and complex plans (“do x if you see z” or “do x with probability p if you see p”).

Other methods for testing for identification are discussed in Berry (Nonrecursive Causal Models). (Verify that can solve for simultaneous-equation model parameters, given values for the reduced equation.)

Summary and Conclusion: Pearl’s approach to causal modeling is a sound mathematical basis for estimating causal effects. Neyman-Rubin potential-outcomes model can be represented in this framework, and is useful for explanation. The models that arise in international development applications are relatively simple, and restricting discussion to the potential-outcomes approach is adequate.

6.3. Impact Measures

**Impact Measures**

Sequence of implementation of intervention: input-process-output-outcome-impact

Impact measures:

- Observed treatment effect (OTE): Difference in mean outcomes for treated and untreated sample units
- Average treatment effect (ATE): Average effect of treatment on a randomly selected member of the eligible (target) population
- Average treatment effect on the treated (ATT, TOT): Average effect of treatment for treated individuals
- Intention-to-treat effect (ITT): Average effect of treatment for those selected (by randomization) for treatment
- Local average treatment effect (LATE): Effect of treatment on “compliers” (individuals who take treatment if offered and do not take treatment if not offered)
PARTIAL TREATMENT EFFECT (PTE): CHANGE IN ATE PER UNIT CHANGE IN A TREATMENT VARIABLE
MARGINAL TREATMENT EFFECT (MTE): AVERAGE EFFECT OF TREATMENT FOR THOSE "JUST ELIGIBLE" FOR A PROGRAM
QUANTILE TREATMENT EFFECT (QTE): AVERAGE EFFECT OF TREATMENT AT A QUANTILE OF A DISTRIBUTION

WILL DESCRIBE MOST, BUT NOT ALL, OF THE PRECEDING IMPACT MEASURES, AND DISCUSS A FEW IN DETAIL.

WILL FIRST DEVELOP THEORY FOR THE CASE OF A SINGLE BINARY TREATMENT VARIABLE, IN THE CASE OF A SINGLE CROSS-SECTIONAL DESIGN (TWO SAMPLES, ONE TREATED AND THE OTHER NOT TREATED; TREATMENT SAMPLE AND COMPARISON SAMPLE (USE THE TERM "CONTROL" MAINLY IN EXPERIMENTS WITH RANDOMIZED ASSIGNMENT TO TREATMENT)).

THE TERM “GROUP” MAY REFER EITHER TO A SAMPLE OR A POPULATION (E.G., “TREATMENT GROUP” MAY REFER TO THE ENTIRE TREATED POPULATION, OR TO A SAMPLE FROM THAT POPULATION).

POTENTIAL OUTCOMES MODEL (SOMETIMES CALLED THE COUNTERFACTUAL MODEL):
FOR EACH SAMPLE UNIT (INDEX i), THERE ARE TWO POTENTIAL OUTCOMES:

\[ y_i = y_{0i} \text{ if } i\text{-th unit is untreated} \]
\[ = y_{1i} \text{ if } i\text{-th unit is treated} \]

IMPACT OF TREATMENT ON THE i-TH UNIT IS \( y_{1i} - y_{0i} \)

IF DEFINE TREATMENT INDICATOR VARIABLE

\[ w_i = 0 \text{ if } i\text{-th unit is untreated} \]
\[ = 1 \text{ if } i\text{-th unit is treated} \]

THEN

\[ y_i = y_{0i} + (y_{1i} - y_{0i})w_i. \]

FOR SIMPLICITY, DROP INDEX i (USED IN FORMULAS), AND WRITE:

\[ y = y_0 + (y_1 - y_0)w. \]

PRIMARY GOAL OF EVALUATION IS TO ESTIMATE THE AVERAGE TREATMENT EFFECT (ATE), OR EXPECTED TREATMENT EFFECT:

\[ ATE = E(y_1 - y_0), \]

OR TO ESTIMATE THE AVERAGE TREATMENT EFFECT ON THE TREATED (ATT)
\[ ATT = E(y_1 - y_0 | w=1) = E(y_1 | w=1) - E(y_0 | w=1). \]

Expressions such as \( E(y_1 | w=1) \) are conditional expectations.

Note that
\[ E(y | w=1) = E(y_0 + (y_1 - y_0)w | w=1) = E(y_1 | w=1) \]
and
\[ E(y | w=0) = E(y_0 + (y_1 - y_0)w | w=0) = E(y_0 | w=0). \]

Example:

Effect of an education or training program on earnings (e.g., Farmer Training and Development Assistance (FTDA) program in Honduras).

ATE is the average increase in earnings for a randomly selected member of the population. (This is what is of primary interest to a government policy maker.)

ATT is the average increase in earnings for program graduates.

ATT is probably higher than ATE, since graduates are likely to be more motivated, ambitious and able than the general population; or because the program implementer may select participants who are more likely to succeed.

As a result, the OTE is probably higher than the ATE. OTE would likely be a biased-high estimate of ATE.

Formulas for ATE and ATT:

\[
ATE = E(y_1 - y_0) = E(y_1) - E(y_0)
\]
\[
= E(y_1|w=1) \ P(w=1) + E(y_1|w=0) \ P(w=0) - E(y_0|w=1) \ P(w=1) - E(y_0|w=0) \ P(w=0).
\]

From this expression, can obtain:

\[
OTE = E(y|w=1) - E(y|w=0)
\]
\[
= E(y_1|w=1) - E(y_0|w=0)
\]
\[
= E(y_1 - y_0) + [E(y_0|w=1) - E(y_0|w=0)] + (1 - P(w=1)) [E(y_1 - y_0 | w=1) - E(y_1 - y_0|w=0)]
\]
\[
= ATE + \text{Selection Bias} + \text{Differential Treatment Effect}.
\]

We also have:

OTE = E(y|w=1) - E(y|w=0)
\[ E(y_1|w=1) - E(y_0|w=0) = [E(y_0|w=1) - E(y_0|w=1)] + [E(y_0|w=1) - E(y_0|w=0)] \]

\[ = \text{ATT + Selection Bias.} \]

**NOTE THAT ATE AND ATT DIFFER ONLY BY THE DIFFERENTIAL TREATMENT EFFECT.**

**IF WE ASSIGN TREATMENT BY RANDOMIZATION, THE POTENTIAL OUTCOMES \( y_0 \) and \( y_1 \) ARE INDEPENDENT OF TREATMENT, \( w \). IN THIS CASE:**

\[ E(y_0|w=1) = E(y_0|w=0), \]

SO THE SELECTION BIAS IS ZERO, SO THAT ATE = ATT; AND

\[ E(y_1 - y_0|w=1) = E(y_1 - y_0|w=0) \]

SO THE DIFFERENTIAL TREATMENT EFFECT IS ALSO ZERO, SO THAT ATE=ATT (=OTE).

**AN ALTERNATIVE WAY OF SHOWING THE LATTER RESULT IS TO WRITE**

\[ y_0 = \mu_0 + v_0 \]

and

\[ y_1 = \mu_1 + v_1 \]

where

\[ \mu_i = E(y_i), i = 0, 1 \]

and

\[ E(v_i) = 0 \]

(THAT IS, SPLIT \( y_i \) INTO ITS MEAN PLUS A RANDOM COMPONENT). IN THIS NOTATION,

\[ y_1 - y_0 = \mu_1 - \mu_0 + v_1 - v_0 = \text{ATE} + v_1 - v_0. \]

**TAKING EXPECTATIONS ON BOTH SIDES OF THE EQUATION, CONDITIONAL ON \( w = 1 \), YIELDS**

\[ E(y_1 - y_0|w=1) = \text{ATE} + E(v_1 - v_0|w=1) \]

or

\[ \text{ATT} = \text{ATT} + E(v_1 - v_0|w=1). \]

 THAT IS, ATT AND ATE DIFFER BY THE DIFFERENTIAL TREATMENT EFFECT.
IN SUMMARY, UNDER THE ASSUMPTION OF INDEPENDENCE OF \((y_0, y_1)\) and \(w\), ATT EQUALS ATE, AND BOTH MAY BE ESTIMATED AS THE DIFFERENCE IN MEANS BETWEEN THE TREATED AND UNTREATED POPULATIONS.

THIS RESULT UNDERSCORES THE DESIRABILITY OF USING RANDOMIZED ASSIGNMENT OF TREATMENT – THE ANALYSIS IS STRAIGHTFORWARD, AND THE ESTIMATES ARE UNBIASED.

FOR MANY APPLICATIONS IN INTERNATIONAL DEVELOPMENT, IT IS NOT FEASIBLE TO RANDOMLY ASSIGN TREATMENT. MUCH OF THE SUBJECT OF STATISTICAL DESIGN AND ANALYSIS ADDRESSES THE ISSUE OF WHAT TO DO IN THAT CASE.

NOTE: THE PRECEDING DISCUSSED OTE, ATE AND ATT. THE OTHER IMPACT ESTIMATORS LISTED EARLIER WILL BE ADDRESSED LATER.

ALTERNATIVES TO RANDOMIZED ASSIGNMENT OF TREATMENT: CONDITIONAL INDEPENDENCE ("IGNORABILITY OF TREATMENT")

TWO TECHNIQUES: MATCHING AND REGRESSION ANALYSIS (REGRESSION ADJUSTMENT, COVARIATE ADJUSTMENT).

MATCHING: SELECT COMPARISON SAMPLE TO BE SIMILAR TO THE TREATMENT SAMPLE WITH RESPECT TO OBSERVABLE VARIABLES (THAT ARE BELIEVED TO HAVE AN EFFECT ON OUTCOME).

REGRESSION ADJUSTMENT: DEVELOP A STATISTICAL MODEL THAT ESTIMATES PARAMETERS OF THE CAUSAL MODEL, AND USE THIS AS A BASIS FOR ESTIMATING IMPACT (AVERAGE TREATMENT EFFECT OR PARTIAL TREATMENT EFFECT).

KEY RESULT: MATCHING AND REGRESSION CAN BE COMBINED TO PRODUCE CONSISTENT ESTIMATES IF EITHER THE REGRESSION MODEL OR THE MATCHING MODEL IS CORRECTLY SPECIFIED. THIS APPROACH IS SAID TO BE DOUBLY ROBUST.

REFERENCES ON MATCHING:


IDEAL SITUATION: WANT INDEPENDENCE OF POTENTIAL OUTCOMES \((y_0, y_1)\) AND TREATMENT \(w\) (E.G., AS IN AN EXPERIMENTAL DESIGN (RANDOMIZED ASSIGNMENT TO TREATMENT)).

ACHIEVABLE GOAL: CONDITIONAL INDEPENDENCE OF \((y_0, y_1)\) AND TREATMENT \(w\), CONDITIONAL ON OBSERVED COVARIATES \(x\). (BOLD LETTERS WILL NOW BE USED TO DENOTE VECTORS.) THIS CONDITION IS CALLED CONDITIONAL INDEPENDENCE, IGNORABILITY OR UNCONFOUNDEDNESS. IF, IN ADDITION, \(0 < P(w=1|x) < 1\), THE
CONDITION IS CALLED STRONG IGNORABILITY (FOR ANY SETTING OF \( x \) IN THE POPULATION, THERE IS A CHANCE OF SEEING UNITS IN THE TREATMENT AND COMPARISON GROUPS). THE TERM “IGNORABLE” IS USED SINCE, CONDITIONAL ON \( x \), THE POTENTIAL OUTCOMES FOR A UNIT DO NOT DEPEND ON TREATMENT (\( w \)), I.E., TREATMENT STATUS OR MECHANISM MAY BE “IGNORED” (WHEN COMPARING THE POTENTIAL OUTCOMES).

A BALANCING SCORE IS ANY FUNCTION OF THE COVARIATES, \( b(x) \), SUCH THAT THE CONDITIONAL DISTRIBUTION OF \( x \) GIVEN \( b(x) \) IS INDEPENDENT OF TREATMENT (I.E., THE SAME FOR TREATMENT AND COMPARISON UNITS). ROSENBAUM AND RUBIN SHOWED THAT THE “COARSEST” BALANCING SCORE IS THE PROPENSITY SCORE, WHICH IS THE PROBABILITY OF BELONGING TO THE TREATMENT GROUP, \( e(x) = P(w=1|x) \).


ROSENBAUM AND RUBIN’S RESULT IS FOR THE TRUE PROPENSITY SCORE. IN PRACTICE, WE CAN MATCH ONLY ON OBSERVABLE COVARIATES (\( x \)), AND IT IS NECESSARY TO ASSUME CONDITIONAL INDEPENDENCE, GIVEN MATCHING ON OBSERVABLES.

REFERENCES ON IMPACT ESTIMATORS IN THE POTENTIAL-OUTCOMES FRAMEWORK:


PRECISION, BIAS AND ACCURACY

PRECISION (RELIABILITY) REFERS TO VARIATION OF A PARAMETER ESTIMATOR AROUND ITS EXPECTED VALUE. IT IS MEASURED BY THE STANDARD ERROR (STANDARD DEVIATION) OF THE ESTIMATOR.

BIAS (VALIDITY) IS THE DIFFERENCE BETWEEN THE EXPECTED VALUE OF AN ESTIMATOR AND THE (TRUE) VALUE OF THE PARAMETER BEING ESTIMATED.

ACCURACY IS A COMBINATION MEASURE OF PRECISION AND BIAS. IT REFERS TO VARIATION OF A PARAMETER ESTIMATOR AROUND THE (TRUE) VALUE OF THE PARAMETER BEING ESTIMATED. IT IS MEASURED BY THE ROOT MEAN SQUARED
ERROR, MSE, EQUAL TO THE SQUARE ROOT OF THE VARIANCE OF THE ESTIMATOR PLUS THE SQUARE OF THE BIAS.

UNBIASEDNESS AND CONSISTENCY

AN ESTIMATOR IS UNBIASED IF ITS EXPECTED VALUE IS EQUAL TO THE (TRUE) VALUE OF THE PARAMETER BEING ESTIMATED.

AN ESTIMATOR IS CONSISTENT IF ITS EXPECTED VALUE CONVERGES TO THE (TRUE) VALUE OF THE PARAMETER BEING ESTIMATED, AS THE SAMPLE SIZE BECOMES VERY LARGE.

IN EVALUATION (ECONOMETRICS, ANALYTICAL SURVEYS), EXPLANATORY VARIABLES ARE USUALLY VIEWED AS RANDOM VARIABLES, NOT AS FIXED QUANTITIES, AND ATTENTION FOCUSES ON ASYMPTOTIC PROPERTIES OF ESTIMATORS (I.E., ON CONSISTENCY).

INTERNAL VALIDITY AND EXTERNAL VALIDITY

INTERNAL VALIDITY OF AN ESTIMATOR REFERS TO ITS ACCURACY RELATIVE TO THE POPULATION UNDER STUDY. (USE A CONSISTENT ESTIMATOR.)

EXTERNAL VALIDITY OF AN ESTIMATOR REFERS TO ITS ACCURACY RELATIVE TO A LARGER POPULATION OF INTEREST. (SELECT A PROBABILITY SAMPLE FROM THE TARGET POPULATION.)

6.4. Basic Evaluation Designs; Models; Estimators

BASIC EVALUATION DESIGNS / MODELS / ESTIMATORS

EXAMPLE DESIGNS TO BE DISCUSSED:

DESIGNS BASED ON RANDOMIZATION

TWO-GROUP DESIGNS
  TREATMENT/CONTROL DESIGN
  PRETEST/POSTTEST DESIGN
  INTERRUPTED TIME-SERIES DESIGN
  FOUR-GROUP DESIGN

QUASI-EXPERIMENTAL DESIGNS

OBSERVATIONAL DATA

DESIGNS BASED ON RANDOMIZATION

NOTE: RANDOMIZATION SHOULD OCCUR AT THE LEVEL OF THE ULTIMATE SAMPLING UNIT, E.G., AT THE HOUSEHOLD LEVEL. IN MANY EVALUATION STUDIES, IT OCCURS AT A HIGHER ADMINISTRATIVE LEVEL, SUCH AS A VILLAGE, CENSUS ENUMERATION
AREA, OR DISTRICT. THIS TYPE OF RANDOMIZATION (USUALLY REFERRED TO AS “CLUSTER RANDOMIZATION”) PRESENTS PROBLEMS, WHICH WILL BE DISCUSSED LATER. (SUCH DESIGNS SHOULD BE CLASSIFIED AS “QUASI-EXPERIMENTAL” DESIGNS.)

**TWO-GROUP DESIGNS**

**TREATMENT/CONTROL DESIGN**

RANDOMLY SELECT A UNIT FROM THE POPULATION OF INTEREST (TARGET POPULATION, ELIGIBLE POPULATION), AND RANDOMLY ASSIGN EACH UNIT TO TREATMENT OR CONTROL (NON-TREATMENT). APPLY THE TREATMENT TO THOSE SELECTED FOR TREATMENT, AND MAKE OBSERVATION.

DESIGN USEFUL IF RESULTS OCCUR QUICKLY.

NOTE TWO TYPES OF RANDOMIZATION:

- RANDOM SELECTION FROM THE POPULATION (FOR EXTERNAL VALIDITY)
- RANDOM ASSIGNMENT TO TREATMENT (FOR INTERNAL VALIDITY)

**STATISTICAL MODEL:**

\[ y_i = \alpha + \beta w_i + e_i \]

or

\[ y = \alpha + \beta w + e \]

where

- \( y \) = outcome variable (explained variable, dependent variable, response variable, regressand)
- \( w \) = treatment variable (explanatory variable, independent variable, regressor); = 0 if untreated, 1 if treated
- \( e \) = error term (mean zero, independent of each other and independent of \( w \)).

NOTE: UNDER THE POTENTIAL-OUTCOMES CONCEPTUAL FRAMEWORK, THE PRECEDING MODEL APPLIES TO EACH UNIT, ALTHOUGH ONLY ONE OUTCOME IS OBSERVED FOR EACH UNIT (I.E., FOR \( w = 0 \) OR FOR \( w = 1 \), BUT NOT BOTH).

**IMPACT MEASURE:**

\[ \text{ATE} = \text{DIFFERENCE IN MEANS OF POTENTIAL OUTCOMES} \]

\[ = E(y_1 - y_0) = E(y|w=1) - E(y|w=0) = (\alpha + \beta) - (\alpha) = \beta. \]

NOTE: ARE AVERAGING OVER THE POPULATION OF UNITS.
IMPACT ESTIMATOR:

DIFFERENCE IN MEANS OF TREATMENT AND CONTROL GROUPS

\[ E\left(\sum_{i=1}^{n_T} y_{Ti}/n_T\right) - E\left(\sum_{i=1}^{n_C} y_{Ci}/n_C\right) = (\alpha + \beta) - (\alpha) = \beta. \]

NOTE: ARE AVERAGING OVER THE SAMPLE UNITS.

BECAUSE OF RANDOMIZED ASSIGNMENT TO TREATMENT, THE DISTRIBUTION OF EVERY VARIABLEexcept treatment is the same for the treatment and control samples.

PRETEST/POSTTEST DESIGN ("BEFORE/AFTER," "REFLEXIVE" DESIGN)

RANDOMLY SELECT A UNIT FROM THE POPULATION OF INTEREST (TARGET POPULATION, ELIGIBLE POPULATION). MAKE OBSERVATIONS AT TWO POINTS IN TIME (PRETEST/POSTTEST, BEFORE/AFTER).

NOTE: JUST ONE TYPE OF RANDOMIZATION: RANDOM SELECTION FROM THE POPULATION (FOR EXTERNAL VALIDITY), BUT NO RANDOM ASSIGNMENT TO TREATMENT (FOR INTERNAL VALIDITY).

MODEL, IMPACT MEASURE AND IMPACT ESTIMATOR SAME AS ABOVE.

DESIGN MAY BE APPROPRIATE FOR ASSESSING AN IMPACT THAT MANIFESTS QUICKLY.

IF THE TIME BETWEEN SURVEYS IS LONG, MANY VARIABLES THAT MAY AFFECT OUTCOME MAY CHANGE OVER TIME, AND THE POSTULATED MODEL MAY BE INVALID, IN WHICH CASE THE DIFFERENCE IN THE TWO SAMPL MEANS (BEFORE AND AFTER) WOULD NOT BE AN UNBIASED ESTIMATE OF THE IMPACT MEASURE.

INTERRUPTED TIME-SERIES DESIGN

MAY BE USEFUL IN FINANCIAL INTERVENTIONS, WHERE MUCH MONTHLY DATA IS AVAILABLE:

\[ y_1, y_2, y_3, \ldots y_t, \text{ intervention, } y_{t+1}, y_{t+2}, \ldots y_n \]

INTERVENTION AT A RANDOM TIME.

MAY HAVE MANY TIME SERIES (E.G., MANY BANKS).

STATISTICAL MODEL: AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODEL OR AUTOREGRESSIVE CONDITIONAL HETEROSKEDASTICITY (ARCH) MODEL, E.G., ARMA(1,1)

\[ y_t = x_1\beta + \rho(y_{t-1} - x_{t-1}\beta) + \theta\epsilon_{t-1} + \epsilon_t. \]

THE IMPACT IS ONE OF THE \( \beta \) COEFFICIENTS.
FOUR-GROUP DESIGN (TREATMENT BEFORE, TREATMENT AFTER, CONTROL BEFORE, CONTROL AFTER)

POSSIBLE MODEL SPECIFICATION (REPRESENTATION):

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>w=0</td>
<td>w=0</td>
</tr>
<tr>
<td></td>
<td>t=0</td>
<td>t=1</td>
</tr>
<tr>
<td>Treatment</td>
<td>w=0</td>
<td>w=1</td>
</tr>
</tbody>
</table>

RANDOMIZED ASSIGNMENT OF TREATMENT.

DESIGN USEFUL IF RESULTS REQUIRE SOME TIME TO OCCUR, AND IF VARIABLES OTHER THAN TREATMENT THAT AFFECT OUTCOME HAVE THE SAME VALUES FOR THE TREATMENT AND CONTROL SAMPLES.

STATISTICAL MODEL:

\[ y = \alpha + \theta t + \phi w + e. \]

EFFECT (IMPACT) IS COEFFICIENT, \( \phi \), OF TREATMENT VARIABLE, \( w \).

ALTERNATIVE MODEL SPECIFICATION (MORE USEFUL LATER ON, TO ACCOUNT FOR SELECTION EFFECTS):

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>w=0</td>
<td>w=0</td>
</tr>
<tr>
<td></td>
<td>t=0</td>
<td>t=1</td>
</tr>
<tr>
<td>Treatment</td>
<td>w=1</td>
<td>w=1</td>
</tr>
</tbody>
</table>

STATISTICAL MODEL:

\[ y = \alpha + \theta t + \phi w + \delta tw + e. \]

EFFECT IS COEFFICIENT, \( \delta \), OF THE INTERACTION EFFECT OF TREATMENT AND TIME.

BECAUSE OF RANDOMIZED ASSIGNMENT TO TREATMENT, \( \phi=0 \) (NO DIFFERENCE BETWEEN TREATMENT AND CONTROL GROUPS AT TIME 0 (THE BASELINE)). LATER, THIS ASSUMPTION WILL BE DROPPED.

IMPACT MEASURE:

DOUBLE-DIFFERENCE MEASURE:

\[ DD = (\mu_{T1} - \mu_{T0}) - (\mu_{C1} - \mu_{C0}) \]
WHERE THE FIRST SUBSCRIPT SPECIFIES TREATMENT OR CONTROL GROUP (T OR C,
CORRESPONDING (IN THE ALTERNATIVE MODEL REPRESENTATION) TO w = 1 OR w =
0) AND THE SECOND SUBSCRIPT SPECIFIES TIME (t = 0 OR t = 1).

DOUBLE-DIFFERENCE ESTIMATOR:

\[ \hat{D} = (\bar{y}_{T1} - \bar{y}_{T0}) - (\bar{y}_{C1} - \bar{y}_{C0}) \]

THE TREMENDOUS ADVANTAGE OF THE RANDOMIZED-ASSIGNMENT, IF ALL
COVARIATES HAVE THE SAME VALUES (OR TIME-VARY THE SAME WAY) FOR THE
TREATMENT AND CONTROL GROUPS, IS THAT THE ESTIMATOR DOES NOT HAVE TO
TAKE ANY COVARIATES INTO ACCOUNT. THAT IS, THE “RAW” DOUBLE-DIFFERENCE
ESTIMATOR (OTE) IS AN UNBIASED ESTIMATE OF IMPACT.

THE FACT IS, OVER TIME, MANY VARIABLES AFFECTING OUTCOME, OTHER THAN
TREATMENT, MAY VARY DIFFERENTLY FOR THE TREATMENT AND CONTROL GROUPS,
AND TAKING THIS INTO ACCOUNT WILL BE NECESSARY EVEN FOR A RANDOMIZED
FOUR-GROUP (BEFORE AND AFTER) DESIGN.

QUASI-EXPERIMENTAL DESIGNS

QUASI-EXPERIMENTAL DESIGNS HAVE STRUCTURE SIMILAR TO RANDOMIZED
DESIGNS, BUT LACK PROPER RANDOMIZED ASSIGNMENT TO TREATMENT. THE
COMPARISON SAMPLE IS NOT FORMED BY RANDOMIZED ASSIGNMENT TO
TREATMENT, BUT BY MATCHING.

A NUMBER OF MATCHING PROCEDURES IS AVAILABLE, INCLUDING JUDGMENT
MATCHING AND VARIOUS NUMERICAL-ALGORITHM PROCEDURES. MATCHING IS
DONE USING AVAILABLE DATA ON VARIABLES THAT MAY AFFECT OUTCOME (OR
SELECTION). SINCE DATA ARE AVAILABLE GENERALLY FOR ADMINISTRATIVE UNITS
SUCH AS VILLAGES, CENSUS ENUMERATION AREAS, AND DISTRICTS, MATCHING IS
GENERALLY DONE AT THOSE LEVELS, NOT AT THE LEVEL OF THE HOUSEHOLD
(ULTIMATE SAMPLE UNIT).

NOTE THAT MATCHING IS DONE ON INDEPENDENT VARIABLES, NOT ON DEPENDENT
VARIABLES. MATCHING ON DEPENDENT VARIABLES CAN INTRODUCE REGRESSION-
EFFECT BIAS (AS IN THE WESTINGHOUSE – OHIO STATE EVALUATION OF THE HEAD
START PROGRAM).

MATCHING ADJUSTS FOR OVERT BIAS (IMBALANCE IN A COVARIATE), BUT NOT
HIDDEN BIAS.

REGRESSION-EFFECT BIAS (REGRESSION TO THE MEAN):

FRANCIS GALTON: CHILDREN’S (ADULT) HEIGHT IS A WEIGHTED AVERAGE OF
PARENTS’ HEIGHT AND THE POPULATION AVERAGE HEIGHT, I.E., IT “REGRESSES” TO
THE POPULATION MEAN.

EXPLANATION (EXAMPLE FROM ANGRIST, OP. CIT. P. 108):

HEIGHT IS STATIONARY (DOES NOT CHANGE MUCH OVER TIME).
BIVARIATE PLOT SHOWS CLEAR LINEAR RELATIONSHIP [SHOW GRAPH].

\[ y = \text{child height} \]
\[ x = \text{parent height} \]
\[ y = \alpha + \beta x + e \]

Since height is not changing over time (i.e., across generations), the mean and variance of \( x \) and \( y \) are the same. Hence

\[ \beta = \frac{\text{cov}(y,x)}{\text{var}(x)} = \frac{\text{cov}(y,x)}{\sqrt{\text{var}(y) \text{var}(x)}} = \rho_{xy} \]

and

\[ \alpha = E(y) - \beta E(x) = \mu(1 - \beta) = \mu(1 - \rho_{xy}) \]

where \( \rho_{xy} \) is the intergenerational correlation coefficient in height and \( \mu = E(y) = E(x) \) is the population mean height.

Hence the conditional expectation of \( y \) given \( x \) is

\[ E(y|x) = \alpha + \beta x = \mu(1 - \rho_{xy}) + \rho_{xy} x \]

i.e., the height of an adult child is a weighted average of his parents’ height and the population mean.

THE PHENOMENON OF REGRESSION TO THE MEAN OCCURS WHENEVER MATCHING IS DONE ON AN IMPRECISE PRE-MEASURE OF THE DEPENDENT VARIABLE (E.G., SELECTING A CONTROL GROUP FROM A GENERAL POPULATION TO MATCH TO A LESS-ABLE POPULATION, USING TEST SCORES).

TYPES OF MATCHING

MATCHING PRIOR TO RANDOMIZED ASSIGNMENT TO TREATMENT. THIS IS NOT THE TYPE OF MATCHING DISCUSSED HERE, FOR CONSTRUCTION OF CONTROL GROUPS WHEN RANDOMIZED ASSIGNMENT HAS NOT BEEN DONE. THAT TYPE OF MATCHING (“BLOCKING” FOR LOCAL CONTROL) IS USED FOR “MATCHED PAIRS” DESIGNS, TO IMPROVE PRECISION. IT HAS NOTHING TO DO WITH THE PRESENT TOPIC.

EX-ANTE MATCHING. MATCHING PRIOR TO CONDUCT OF THE SURVEY, DURING SURVEY DESIGN. USUALLY CONDUCTED AT AN AGGREGATE LEVEL, SUCH AS VILLAGE, CENSUS ENUMERATION AREA OR DISTRICT (FOR WHICH DATA ARE AVAILABLE FOR USE IN SURVEY DESIGN).

EX-POST MATCHING. MATCHING DONE ON THE FULL DATA SET (E.G., AT THE HOUSEHOLD LEVEL), DURING ANALYSIS (TO REDUCE MODEL DEPENDENCE ON THE DATA). ALSO CALLED TRIMMING, PRUNING OR CULLING OF THE SAMPLE. “DISCARDS,” “THROWS AWAY,” OR “LOSES” DATA. OF LIMITED USE IN INTERNATIONAL DEVELOPMENT APPLICATIONS, WHICH TYPICALLY INVOLVE TWO-STAGE SAMPLE
DESIGNS WITH SMALL FIRST-STAGE SAMPLE SIZES. CANNOT AFFORD TO LOSE DATA. USE PRIMARILY REGRESSION ADJUSTMENT, NOT EX-POST MATCHING, FOR DATA ANALYSIS IN THIS FIELD.

MANY MATCHING TECHNIQUES AVAILABLE, E.G.:

- EXACT MATCHING
- DISTRIBUTION MATCHING
- MATCHING ON A SCORE (MEASURE OF CLOSENESS)
  - NEAREST-NEIGHBOR(S) MATCHING
  - MAHALANOBISS-DISTANCE MATCHING
  - PROPENSITY-SCORE MATCHING
    - NEAREST-NEIGHBOR
    - RADIUS (CALIPER)
    - STRATIFIED
    - KERNEL (WEIGHTING)
  - IMPORTANCE-FUNCTION MATCHING
  - SELECTION-CRITERION-SCORE MATCHING FOR REGRESSION-
    DISCONTINUITY DESIGN
  - OPTIMAL MATCHING

SEE HO ET AL. ARTICLE FOR DETAILS ON MANY OF THE PRECEDING. SEE “MATCHIT” SOFTWARE FOR IMPLEMENTATION.

“JUDGMENT MATCHING” – ALL MATCHING METHODS ARE “JUDGMENT MATCHING.” REMEMBER: CAN MATCH “AT WILL” ON EXPLANATORY VARIABLES.

ALL MATCHING TECHNIQUES ARE BASED ON OBSERVABLES (AVAILABLE EX ANTE OR EX POST).

WITH EX ANTE MATCHING (IN DESIGN), WILL ALMOST INEVITABLY USE REGRESSION ADJUSTMENT AFTER THE FULL DATA SET IS AVAILABLE. USE OF MATCHING ALONE IS INADEQUATE MEANS OF REDUCING BIAS.

WHAT VARIABLES TO INCLUDE AS MATCH VARIABLES? THOSE THAT ARE BELIEVED TO HAVE AN EFFECT ON OUTCOME, BASED ON A CAUSAL MODEL. WITH RESPECT TO PRECISION, MATCHING ON VARIABLES RELATED TO SELECTION IS OF VALUE ONLY TO THE EXTENT THAT THE VARIABLES AFFECTING SELECTION ALSO AFFECT OUTCOME. WITH RESPECT TO BIAS, MATCHING ON VARIABLES RELATED TO SELECTION IS USEFUL (E.G., PROPENSITY-SCORE MATCHING), AS LONG AS THOSE VARIABLES ALSO AFFECT OUTCOME.

EXACT MATCHING:

MATCH EACH TREATMENT UNIT TO ALL POSSIBLE COMPARISON UNITS HAVING EXACTLY THE SAME VALUE ON ALL (MATCH) COVARIATES. FOR EACH TREATMENT UNIT, SELECT A MATCHING COMPARISON UNIT RANDOMLY.
ADVANTAGES: SIMPLICITY, PROBABILITY DISTRIBUTIONS MATCH EXACTLY FOR ALL MATCH VARIABLES. JOINT PROBABILITY DISTRIBUTION MATCHES.
DISADVANTAGES: IN MANY APPLICATIONS, CANNOT FIND EXACT MATCHES (UNLESS CLASSIFICATION CATEGORIES ARE VERY “COURSE”) “CURSE OF DIMENSIONALITY.”
DISTRIBUTION MATCHING:

DISCARD UNITS OUTSIDE OF COMMON SUPPORT. SUBSAMPLE.
DISADVANTAGE: IF MATCH MARGINAL DISTRIBUTIONS, THEN JOINT PROBABILITY DISTRIBUTION DOESN'T MATCH (UNLESS INDEPENDENT).

MATCHING ON A SCORE: DEFINE A MEASURE OF “CLOSENESS” OF UNITS. MATCH CLOSE UNITS (TYPICALLY ONE TO ONE OR ONE TO MANY). ADDRESSES THE “CURSE OF DIMENSIONALITY” (SINCE MATCH ON A SCALAR SCORE).

MAHALANOBIS-DISTANCE MATCHING: IF V IS THE VARIANCE-COVARIANCE MATRIX OF THE MATCH VARIABLES, THE DISTANCE BETWEEN TWO UNITS WITH MATCH VALUES \( x_1 \) AND \( x_2 \) IS \( D = (x_1 - x_2) V^{-1} (x_1 - x_2) \). MATCH NEAREST NEIGHBORS.
DISADVANTAGES: LOW FACE VALIDITY. DOES NOT TAKE INTO ACCOUNT IMPORTANCE OF VARIABLES IN AFFECTING OUTCOME. NO THEORETICAL ADVANTAGES. MOST APPROPRIATE FOR APPLICATIONS INVOLVING THE NORMAL DISTRIBUTION, WHICH IS CHARACTERIZED BY THE MEAN VECTOR AND VARIANCE-COVARIANCE MATRIX.

PROPENSITY-SCORE MATCHING:

ESTIMATE THE PROPENSITY SCORE, WHICH IS THE PROBABILITY OF SELECTION FOR TREATMENT = \( \Pr(w=1|x) \) (LOGISTIC REGRESSION MODEL). VARIOUS MATCHING PROCEDURES (NEAREST-NEIGHBOR, RADIUS, STRATIFIED, KERNEL).
ADVANTAGE: ROSENBAUM AND RUBIN (1983): UNDER THE ASSUMPTION OF CONDITIONAL INDEPENDENCE (“IGNORABILITY”) OF OUTCOMES \((y_0, y_1)\) AND \( w \) GIVEN \( x \), DIFFERENCE OF MEANS OF TREATMENT AND COMPARISON SAMPLES WILL BE UNBIASED ESTIMATE OF IMPACT (ATE = \( E(y_1 - y_0) \)).
DISADVANTAGES: UNITS THAT MATCH ON PROPENSITY SCORE MAY BE VERY DIFFERENT WITH RESPECT TO VARIABLES THAT HAVE AN IMPORTANT EFFECT ON OUTCOME, SO THAT THE PRECISION AND POWER OF THE MATCHED SAMPLES MAY BE VERY POOR. IN MANY CASES, THE ASSUMPTION CONDITIONAL INDEPENDENCE IS TENUOUS. SACRIFICING PRECISION AND POWER ON A TENUOUS HOPE THAT THE BIAS MAY BE REDUCED IS UNREASONABLE.

IMPORTANCE-SCORE MATCHING:

ASSIGN WEIGHTS REFLECTING THE IMPORTANCE OF EACH MATCH VARIABLE IN AFFECTING OUTCOMES OF INTEREST. FOR EACH UNIT, CALCULATE THE WEIGHTED SUM OF MATCH VARIABLES. DISTANCE BETWEEN UNITS IS THE DIFFERENCE IN THE WEIGHTED SUM. MATCH NEAREST NEIGHBORS.
ADVANTAGES: SIMPLE. BASED ON CAUSAL MODEL. RESULTS IN MATCHES WITH HIGH PRECISION AND POWER, AND LOW BIAS.

SELECTION-CRITERION-SCORE MATCHING FOR REGRESSION-DISCONTINUITY DESIGN:

IN SOME PROGRAMS, THE TREATMENT POPULATION IS DETERMINED BY USE OF A SCORE, SUCH AS A “PROXY MEANS TEST” SCORE IN A CONDITIONAL-CASH-TRANSFER (CCT) PROGRAM. TREATMENT AND COMPARISON SAMPLES MAY BE
OBTAINED BY SELECTING UNITS JUST BELOW AND JUST ABOVE THE CUT-OFF SCORE.

ADVANTAGES: SIMPLE TO IMPLEMENT.
DISADVANTAGES: LIMITS SCOPE OF INFERENCE (EXTERNAL VALIDITY), BECAUSE IT IS RESTRICTED TO THE POPULATION CLOSE TO THE CUT-OFF THRESHOLD. CUT-OFF SCORE FOCUSES ON NEED FOR SERVICE, AND IGNORES THE IMPORTANCE OF VARIABLES THAT AFFECT OUTCOMES OF INTEREST (LOW PRECISION AND POWER).

OPTIMAL MATCHING:

NEAREST-NEIGHBOR MATCHING DEPENDS ON THE ORDER (SEQUENCE) OF THE MATCH, AND, FOR SMALL TREATMENT AND CONTROL POPULATIONS, THE MATCH MAY BE POOR FOR THE LAST MATCHES. OPTIMAL MATCHING SEEKS TO MATCH TO FIND THE MINIMAL VALUE FOR THE AVERAGE DISTANCE.

ADVANTAGE: MAY IMPROVE PRECISION AND POWER OVER NEAREST-NEIGHBOR APPROACH.
DISADVANTAGE: MORE COMPLEX.

REMEMBER: IN ALL INSTANCES, WILL STILL HAVE TO USE COVARIATE ADJUSTMENT (EVEN FOR PROPENSITY-SCORE MATCHING, BECAUSE THE CONDITIONAL INDEPENDENCE ASSUMPTION MAY NOT BE JUSTIFIED (AND THIS MAY BE ADDRESSED WHEN THE FULL DATA SET IS AVAILABLE)). IN PARTICULAR, IF DO MATCHING ON FIRST-STAGE SAMPLE UNITS, WILL HAVE TO COVARIATE-ADJUST WHEN FULL DATA SET IS AVAILABLE.

OBSERVATIONAL DATA

EXPERIMENT: AN EMPIRICAL INVESTIGATION IN WHICH ASSIGNMENT OF TREATMENT TO UNITS (SUBJECTS) IS CONTROLLED BY THE EXPERIMENTOR (INVESTIGATOR, RESEARCHER).

OBSERVATIONAL DATA: NO CONTROL OF ASSIGNMENT OF TREATMENT TO UNITS. (GOAL IS STILL THE SAME: ASSESS CAUSE-AND-EFFECT RELATIONSHIPS.)

IN THE ANALYSIS OF OBSERVATIONAL DATA, THERE IS LITTLE OR NO DESIGN STRUCTURE, AND NO CONTROL OVER THE ASSIGNMENT OF TREATMENT TO UNITS. THERE MAY BE ONE OR MORE CONTROL GROUPS AVAILABLE, BUT THEY ARE NOT CONSTRUCTED BY THE INVESTIGATOR. LOW EXTERNAL AND INTERNAL VALIDITY.

MATCHING AND COVARIATE ADJUSTMENT (OR CONTINGENCY TABLES) ARE DONE IN THE ANALYSIS, NOT IN THE DESIGN.

6.5. Sample Survey Design for Evaluation

SAMPLE SURVEY DESIGN FOR EVALUATION

SAMPLE SURVEY DESIGN FOR EVALUATION IS ANALYTICAL SURVEY DESIGN, WHICH COMBINES ELEMENTS OF DESCRIPTIVE SAMPLE SURVEY DESIGN AND EXPERIMENTAL DESIGN (DESIGN OF EXPERIMENTS).
PRINCIPLES OF EXPERIMENTAL DESIGN

RANDOMIZATION (TO ESTABLISH CAUSALITY)
REPLICATION (FOR ESTIMATION OF EXPERIMENTAL ERROR)
SYMMETRY (FOR EASE OF COMPUTATION AND TO ACHIEVE UNAMBIGUOUS RESULTS (ORTHOGONALITY))
LOCAL CONTROL (TO INCREASE PRECISION AND POWER)

FOR MULTIPLE TREATMENT VARIABLES AND LEVELS, EXPERIMENTAL DESIGNS ARE ELABORATE, E.G., FRACTIONAL FACTORIAL DESIGNS, BALANCED INCOMPLETE BLOCK DESIGNS, LATIN SQUARE DESIGNS. FOR DEVELOPMENT APPLICATIONS, OFTEN HAVE A SINGLE BIVARIATE TREATMENT VARIABLE (TREATED VS. UNTREATED). DESIGNS ARE SIMPLE, BUT THE PRINCIPLES OF EXPERIMENTAL DESIGN STILL APPLY, AND PLAY AN IMPORTANT ROLE.

ALSO VERY IMPORTANT ARE ASPECTS OF DESCRIPTIVE SAMPLE SURVEY DESIGN (STRATIFICATION, CLUSTER SAMPLING, MULTISTAGE SAMPLING), BUT WITH SOME MODIFICATION TO INCORPORATE DESIGN CONSIDERATIONS (MARGINAL STRATIFICATION; MATCHING; INTRODUCTION OF CORRELATIONS (FOR DIFFERENCE ESTIMATES); BALANCE AND SPREAD; ORTHOGONALIZATION OF COVARIATES); POWER ANALYSIS VS. PRECISION ANALYSIS); RANDOM EFFECTS; NO FPC.

DATA ANALYSIS (CONSIDERED IN DAY 2) IS RADICALLY DIFFERENT FOR ANALYTICAL SURVEY DESIGNS COMPARED TO DESCRIPTIVE SURVEY DESIGNS.

DESCRIPTIVE SURVEYS VS. ANALYTICAL SURVEYS

DESCRIPTIVE SURVEY:

ESTIMATE THE CHARACTERISTICS (TOTALS, MEANS, PROPORTIONS) FOR A POPULATION OR SUBPOPULATIONS. FIXED, FINITE POPULATION. DESIGN-BASED ESTIMATES. CLOSED-FORM FORMULAS FOR ESTIMATES AND STANDARD ERRORS. FINITE POPULATION CORRECTION (FPC) (PRECISION BECOMES PERFECT WHEN SAMPLE SIZE APPROACHES FINITE POPULATION SIZE). THE RANDOM VARIABLES OF INTEREST ARE THE EVENTS OF INCLUSION IN THE SAMPLE, NOT THE VARIABLES BEING MEASURED ON THE SAMPLE UNITS. FOCUS ON PRECISION OF POINT ESTIMATES AND (CONFIDENCE INTERVALS), NOT ON TESTS OF HYPOTHESIS. SAMPLE SIZE ESTIMATION BASED ON PRECISION ANALYSIS.

ANALYTICAL SURVEY:

ESTIMATE THE CHARACTERISTICS OF A PROCESS THAT AFFECTS OR GENERATES THE FINITE POPULATION BEING SURVEYED. IMPACT OF AN INTERVENTION, OR RELATIONSHIP OF IMPACT TO EXPLANATORY VARIABLES. MODEL-BASED (MODEL-ASSISTED) ESTIMATES. FINITE POPULATION BEING SURVEYED IS SIMPLY A REALIZATION OF AN INFINITE COLLECTION OF SAMPLES THAT THE PROCESS COULD PRODUCE. FPC DOES NOT APPLY. THE RANDOM VARIABLES OF INTEREST ARE THE VARIABLES BEING MEASURED ON THE SAMPLE UNITS. FOCUS ON POWER OF TESTS OF HYPOTHESIS ABOUT IMPACT. SAMPLE SIZE ESTIMATION BASED ON POWER ANALYSIS.
SAMPLE SURVEY DESIGN TECHNIQUES

CLUSTER SAMPLING
MULTISTAGE SAMPLING
STRATIFIED SAMPLING
   ORDINARY STRATIFICATION
   MARGINAL STRATIFICATION
TWO-PHASE SAMPLING
USE OF MATCHING TO IMPROVE PRECISION AND POWER
ESTIMATION OF SAMPLE SIZE FOR DESCRIPTIVE SURVEYS
ESTIMATION OF SAMPLE SIZE FOR ANALYTICAL SURVEYS
SAMPLE SELECTION PROCEDURES

CLUSTER SAMPLING

EFFICIENCY OF A DESIGN REFERS TO THE RATIO OF THE PRECISION OR POWER OF AN ESTIMATE TO SURVEY COST (MORE EFFICIENT DESIGN HAS HIGHER PRECISION (LOWER STANDARD ERRORS OF ESTIMATES) AND POWER).

DESIGN EFFECT. THE STANDARD MEASURE OF THE EFFICIENCY OF A DESIGN IS THE RATIO OF THE VARIANCE (OF AN ESTIMATOR) FOR THE DESIGN TO THE VARIANCE USING A SIMPLE RANDOM SAMPLE OF THE SAME SIZE. THIS IS CALLED KISH'S DESIGN EFFECT, OR “DEFF.” IT IS CALLED THE “MOULTON FACTOR” BY SOME ECONOMISTS.

CLUSTER SAMPLING. THE COST OF COLLECTING DATA FROM A SIMPLE RANDOM SAMPLE OF HOUSEHOLDS DISTRIBUTED OVER A LARGE GEOGRAPHIC AREA MAY BE VERY HIGH. IT MAY BE MUCH MORE EFFICIENT TO SELECT A RANDOM SAMPLE OF SMALL AREAS, AND SURVEY THE HOUSEHOLDS IN THE SELECTED AREAS. THE AREAS ARE CALLED “CLUSTERS,” AND THIS TYPE OF SAMPLING IS CALLED “CLUSTER SAMPLING.” THE CLUSTERS MAY BE DEFINED BY THE INVESTIGATOR, BUT MORE LIKELY ARE EXISTING ADMINISTRATIVE GROUPINGS, SUCH AS VILLAGES, CENSUS ENUMERATION AREAS, OR DISTRICTS.

THE EFFICIENCY OF CLUSTER SAMPLING DEPENDS ON THE RELATIVE COST OF SAMPLING CLUSTERS VS. UNITS WITHIN CLUSTERS, AND ON THE INTERNAL HOMOGENEITY OF THE CLUSTERS RELATIVE TO A VARIABLE OF INTEREST.

THE PRECISION OF CLUSTER SAMPLING FOR ESTIMATING THE MEAN OF A VARIABLE OF INTEREST DEPENDS ON THE INTERNAL HOMOGENEITY OF THE CLUSTERS RELATIVE TO THE GENERAL POPULATION.

THE INTERNAL HOMOGENEITY OF A CLUSTER IS MEASURED BY THE INTRA-CLUSTER CORRELATION COEFFICIENT (DENOTED icc OR ρ). IT VARIES FOR EACH VARIABLE OF INTEREST. IT IS EQUAL TO 1 IF ALL OF THE UNITS WITHIN A CLUSTER ARE IDENTICAL (WITH RESPECT TO THE VARIABLE OF INTEREST), AND EQUAL TO 0 IF THE UNITS WITHIN A CLUSTER HAVE THE SAME VARIABILITY AS THE GENERAL POPULATION. FOR MOST VARIABLES, IT IS POSITIVE (I.E., PRECISION IS DECREASED BY CLUSTER SAMPLING), BUT NOT ALWAYS (E.G., THE GENDER OF RESIDENTS IN TWO-MEMBER HOUSEHOLDS IS NEGATIVELY CORRELATED).

THE VALUE OF ρ TENDS TO DECREASE AS THE CLUSTER SIZE (NUMBER OF UNITS IN THE CLUSTER) INCREASES.

IF $s^2$ DENOTES THE VARIANCE OF UNITS IN THE POPULATION AND $s_w^2$ DENOTES THE WITHIN-CLUSTER VARIANCE OF UNITS, THEN (APPROXIMATELY)

$$\rho = 1 - \frac{s_w^2}{s^2},$$

AND THE DESIGN EFFECT IS GIVEN BY

$$deff = 1 + (M-1)\rho,$$


MULTISTAGE SAMPLING

CLUSTER SAMPLING IS NOT USED A LOT, BECAUSE IT IS USUALLY INEFFICIENT TO MEASURE (INTERVIEW) ALL UNITS WITHIN A CLUSTER. AT SOME POINT, IT BECOMES MORE EFFICIENT TO SAMPLE MORE CLUSTERS THAN TO COLLECT ADDITIONAL INFORMATION ABOUT THE CLUSTERS ALREADY IN THE SAMPLE (SINCE THEY OFTEN TEND TO BE SIMILAR). INSTEAD, IT IS MORE EFFICIENT TO SELECT A SAMPLE OF UNITS WITHIN EACH CLUSTER, THAN TO SELECT ALL OF THEM. THIS PROCEDURE IS CALLED MULTISTAGE SAMPLING. IN THIS CASE, THE CLUSTERS ARE REFERRED TO AS FIRST-STAGE SAMPLE UNITS OR PRIMARY SAMPLING UNITS (PSUs), AND THE UNITS WITHIN THE FIRST-STAGE SAMPLE UNITS ARE CALLED SECOND-STAGE SAMPLE UNITS OR “ULTIMATE SAMPLE UNITS.”

THE WITHIN-PSU SAMPLE SIZE MAY BE FIXED (DENOTED m) OR VARIABLE (DENOTED $m_i$). IN MANY APPLICATIONS, IT IS EFFICIENT (ADMINISTRATIVELY CONVENIENT) TO USE A FIXED WITHIN-PSU SAMPLE SIZE, m. FOR LARGE SAMPLE SIZES OF SIMILAR PSUs, RESULTS SIMILAR TO THOSE PRESENTED ABOVE HOLD:

$$\rho = 1 - \frac{s_w^2}{s^2},$$

AND

$$deff = 1 + (m-1)\rho.$$
IT IS OF INTEREST TO DETERMINE THE OPTIMAL VALUE OF $m$, GIVEN THE VALUES OF $\rho$ AND THE COSTS OF SAMPLING FIRST- AND SECOND-STAGE SAMPLE UNITS, $c_1$ and $c_2$, WHERE THE MARGINAL SURVEY COST IS GIVEN BY

$$C = c_1 n + c_2 nm$$

WHERE $n$ IS THE NUMBER OF FIRST-STAGE SAMPLE UNITS AND $m$ IS THE NUMBER OF SECOND-STAGE SAMPLE UNITS:

$$m_{\text{opt}} = \sqrt{\left(\frac{c_1}{c_2}\right) \left(\frac{1-\rho}{\rho}\right)}.$$ 

IF $C$ IS SPECIFIED, THEN THE FIRST-STAGE SAMPLE SIZE, $n$, IS DETERMINED BY SOLVING THE COST EQUATION.

THE INTERESTING FEATURE OF THE PRECEDING RESULT IS THAT THE OPTIMAL VALUE OF $m$ DOES NOT DEPEND ON $n$.

IN DESCRIPTIVE SURVEYS, IT IS USUALLY EFFICIENT TO HAVE THE PROBABILITY OF SELECTION OF THE ULTIMATE SAMPLE UNITS UNIFORM. THIS IS ACCOMPLISHED EITHER BY SELECTING THE PSUs WITH FIXED PROBABILITIES AND SETTING THE WITHIN-PSU SAMPLE SIZE PROPORTIONAL TO THE PSU SIZE, OR BY SELECTING THE PSUs WITH PROBABILITIES PROPORTIONAL TO SIZE AND SETTING THE WITHIN-PSU SAMPLE SIZE EQUAL TO A CONSTANT (I.E., TO $m$). (THE LATTER METHOD IS THE ONE USUALLY EMPLOYED, FOR REASONS OF EFFICIENCY AND CONVENIENCE IN FIELD WORK.)

STRATIFIED SAMPLING

FOR DESCRIPTIVE SURVEYS: ORDINARY STRATIFICATION

DIVIDE THE POPULATION INTO SUBPOPULATIONS (CATEGORIES, “STRATA”), AND SELECT A SAMPLE FROM EACH.

IN DEVELOPMENT APPLICATIONS, ARE USUALLY STRATIFYING FIRST-STAGE UNITS, SINCE DATA ARE AVAILABLE FOR THAT LEVEL, FOR USE IN SURVEY DESIGN (UNLESS DO A LISTING SURVEY).

REASONS FOR USING STRATIFICATION IN DESCRIPTIVE SURVEYS:

- WANT ESTIMATES FOR SUBPOPULATIONS
- ADMINISTRATIVE CONVENIENCE
- DIFFERENT SAMPLING APPROACHES
- GAIN IN PRECISION IF STRATA ARE INTERNALLY HOMOGENEOUS

(EQUIVALENT TO MULTISTAGE SAMPLING, WHERE A SAMPLE IS SELECTED FROM EVERY FIRST-STAGE SAMPLE UNIT IN THE POPULATION.)

ESTIMATE POPULATION MEAN AS A WEIGHTED AVERAGE OF STRATUM MEANS:

[FORMULAS]
NEED TO DETERMINE THE NUMBER OF STRATUM CATEGORIES, THE STRATUM BOUNDARIES, AND THE ALLOCATION OF THE SAMPLE TO THE STRATA. **OPTIMAL (NEYMAN) ALLOCATION, DALENIUS BOUNDARIES.**

ORDINARY STRATIFICATION USUALLY DONE WITH ONE OR TWO VARIABLES OF STRATIFICATION. WITH MANY VARIABLES OF STRATIFICATION, THE NUMBER OF STRATA IN A CROSS-STRATIFICATION BECOMES VERY LARGE, EASILY EXCEEDING THE SAMPLE SIZE. MANY STRATA CONTAIN 0 OR 1 UNIT (UNLESS COLLAPSE STRATA). DIFFICULT TO CONTROL SELECTION PROBABILITIES. LOW PRECISION, NO ADVANTAGE.

OTHER METHODS: TWO-WAY STRATIFICATION, CONTROLLED SELECTION. USEFUL FOR A FEW VARIABLES OF STRATIFICATION AND SMALL SAMPLE SIZES (E.G., STRATIFICATION OF PSUs).

FOR ANALYTICAL SURVEYS: MARGINAL STRATIFICATION

SPECIFY VARIABLE PROBABILITIES OF SELECTION, SO THAT THE EXPECTED NUMBER OF UNITS IN EACH MARGINAL STRATUM CATEGORY ("CELL") IS AS DESIRED.

REASONS FOR MARGINAL STRATIFICATION IN ANALYTICAL SURVEYS:

WORKS FOR ANY NUMBER OF VARIABLES, ANY NUMBER OF LEVELS PER VARIABLE
WANT TO CONTROL VARIATION (SPREAD, BALANCE, ORTHOGONALITY) FOR A LARGE NUMBER OF VARIABLES

APPROACH:

IDENTIFY VARIABLES (HAVING PRE-SURVEY DATA AVAILABLE) THAT AFFECT OUTCOME (BASED ON CAUSAL MODEL); ELIMINATE OR COMBINE CAUSALLY-REDUNDANT VARIABLES.
SET A SMALL NUMBER (<10) OF STRATA FOR EACH VARIABLE (NATURAL OR OTHERWISE)
SPECIFY A DESIRED ALLOCATION FOR THE MARGINAL DISTRIBUTION OF EACH VARIABLE OF STRATIFICATION. (THIS DETERMINES DESIRED SAMPLING FRACTIONS.)
USE AN ALGORITHM TO DETERMINE VARIABLE SELECTION PROBABILITIES TO ACHIEVE DESIRED ALLOCATION AS CLOSELY AS POSSIBLE, TAKING INTO ACCOUNT THE RELATIVE IMPORTANCE OF THE VARIABLES IN AFFECTING OUTCOME.

TWO-PHASE SAMPLING

A PROBABILITY SAMPLE IS ONE IN WHICH THE PROBABILITIES OF SELECTION OF THE ULTIMATE SAMPLE UNITS IS KNOWN, OR EQUAL TO A CONSTANT.

IF PROBABILITY SAMPLING IS USED, THE THEORY OF STATISTICS MAY BE APPLIED TO MAKE INFERENCES FROM THE SAMPLE DATA.
IN ORDER TO SELECT A PROBABILITY SAMPLE, IT IS DESIRABLE TO HAVE A SAMPLE FRAME (OR FRAME OR SAMPLING FRAME). A SAMPLE FRAME IS A COMPLETE, ACCURATE, UP-TO-DATE LIST OF THE PSUs AND A COMPLETE, ACCURATE UP-TO-DATE LIST OF THE UNITS WITHIN EACH SAMPLED PSU.

IF A SAMPLE FRAME IS NOT AVAILABLE, IT MAY BE ADVANTAGEOUS TO ALLOCATE PROJECT RESOURCES TO CONSTRUCTING ONE. ADVANTAGES: BETTER KNOWLEDGE OF POPULATION SIZES AND HENCE, SELECTION PROBABILITIES (REDUCE BIAS). ABLE TO STRATIFY (USE FIRST SURVEY TO ESTIMATE STRATUM WEIGHTS), AND THEREBY IMPROVE PRECISION.

GIVEN COST OF CLASSIFICATION PER UNIT AND COST OF MEASURING A UNIT IN EACH STRATUM, CAN DETERMINE OPTIMAL CHOICES FOR THE FIRST-PHASE SAMPLE SIZE (TO MEASURE AUXILIARY VARIATES) AND SECOND-PHASE SAMPLE SIZE.

USE OF MATCHING TO IMPROVE PRECISION AND POWER

TWO SITUATIONS FOR MATCHING:

FOR RANDOMIZED DESIGNS (MATCHING PRIOR TO RANDOMIZED ASSIGNMENT)

FOR QUASI-EXPERIMENTAL DESIGNS (TO FORM COMPARISON GROUPS; MATCHING AFTER SELECTION FOR TREATMENT)

FOR RANDOMIZED DESIGNS (MATCHED-PAIRS DESIGNS)

DOUBLE-MATCHED-PAIRS DESIGN

INTERVIEW OF SAME HOUSEHOLDS IN BOTH SURVEY ROUNDS MATCHING PRIOR TO RANDOMIZED ASSIGNMENT TO TREATMENT

REASON WHY MATCHING IS IMPORTANT: IT INTRODUCES A CORRELATION INTO VARIABLES IN MATCHED PAIRS, AND REDUCES VARIANCE OF ESTIMATES OF DIFFERENCES. THIS IMPROVES PRECISION OF DIFFERENCE ESTIMATES AND POWER OF TESTS OF HYPOTHESIS ABOUT DIFFERENCE MEASURES. (PAIR MATCHING INTRODUCES NO BIAS AS LONG AS THE MATCHING CRITERIA (MATCH VARIABLES) ARE UNRELATED TO THE DEPENDENT VARIABLES.)

TWO SAMPLES OF SIZE \( n_1 \) AND \( n_2 \). ASSUME \( n_1 = n_2 = n \). ASSUME EQUAL VARIANCE, \( \sigma^2 \).

IF INDEPENDENT:

\[
\text{var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = 2\sigma^2/n .
\]

IF MATCHED: LET \( \rho \) DENOTE THE CORRELATION BETWEEN MATCHED UNITS. THAT IS, \( \text{cov}(x_1, x_2) = \rho \sigma^2 \). SAMPLE OF \( n \) MATCHED PAIRS OF \( x_1 - x_2 \). \( \text{var}(x_1 - x_2) = \sigma_1^2 + \sigma_2^2 - 2 \text{cov}(x_1, x_2) = \sigma_1^2 + \sigma_2^2 - 2\rho \sigma^2 = (1-\rho)2\sigma^2 \). IT FOLLOWS THAT

\[
\text{var}(\bar{x}_1 - \bar{x}_2) = (1-\rho)2\sigma^2/n .
\]
FOUR SAMPLES (TREATMENT BEFORE, TREATMENT AFTER, CONTROL BEFORE, CONTROL AFTER):

IF INDEPENDENT:

$$\text{var}(\bar{x}_1 - \bar{x}_2 - \bar{x}_3 + \bar{x}_4) = 4\sigma^2/n.$$ 

IF MATCH BEFORE AND AFTER (REINTERVIEW OF SAME HOUSEHOLDS): LET $\rho_{12}$ DENOTE THE CORRELATION BETWEEN MATCHED UNITS. THEN

$$\text{var}(\bar{x}_1 - \bar{x}_2 - \bar{x}_3 + \bar{x}_4) = (1 - \rho_{12})4\sigma^2/n.$$ 

IF MATCH TREATMENTS AND CONTROLS: LET $\rho_{13}$ DENOTE THE CORRELATION BETWEEN MATCHED UNITS. THEN

$$\text{var}(\bar{x}_1 - \bar{x}_2 - \bar{x}_3 + \bar{x}_4) = (1 - \rho_{13})4\sigma^2/n.$$ 


$$\text{var}(\bar{x}_1 - \bar{x}_2 - \bar{x}_3 + \bar{x}_4) = (1 - \rho_{12} - \rho_{13} + \rho_{14} + \rho_{23} - \rho_{24} - \rho_{34})4\sigma^2/n.$$ 

INTRODUCTION OF CORRELATIONS (USE OF MATCHING) BETWEEN THE GROUPS ALLOWS FOR A SUBSTANTIAL REDUCTION IN THE VARIANCE OF DIFFERENCE ESTIMATORS.

HOW TO MATCH?

MATCH ON VARIABLES THAT HAVE AN EFFECT ON OUTCOME. IN SAMPLE SURVEY, CANNOT MATCH "AT WILL." MUST WORK WITH EXISTING FINITE POPULATION. MATCHING WILL INVOLVE TRADE-OFFS – GOOD MATCHES ON SOME VARIABLES WILL RESULT IN POOR MATCHES ON OTHERS. TAKE INTO ACCOUNT THE RELATIVE IMPORTANCE OF THE VARIABLES IN AFFECTING OUTCOME, ACCORDING TO CAUSAL MODEL.

DO NOT MATCH ON THE PROPENSITY SCORE. UNITS HAVING SIMILAR PROPENSITY SCORES MAY HAVE VERY DIFFERENT VALUES FOR VARIABLES THAT HAVE AN IMPORTANT EFFECT ON OUTCOME. EXAMPLE: ABILITY TO JUMP. MATCH ON WHETHER IN THE ARMED SERVICES. (ALSO, IN THE DESIGN STAGE OF A SURVEY THE ESTIMATED PROPENSITY SCORE MAY BE VERY UNRELIABLE.)

IN DEVELOPMENT APPLICATIONS, USUALLY MATCH FIRST-STAGE SAMPLE UNITS (PSUs, CLUSTERS): "CLUSTER-MATCH," "CLUSTER-RANDOMIZED" DESIGN.

IMPORTANT NOTE: RANDOMIZED ASSIGNMENT OF PSUs DOES NOT CONSTITUTE RANDOMIZED ASSIGNMENT OF THE ULTIMATE SAMPLE UNITS.

THREE APPROACHES:
RANDOMIZE ASSIGNMENT OF SAMPLE UNITS WITHIN TREATMENT AND
CONTROL UNITS
SELECT TREATMENT UNITS WITHIN TREATMENT AND CONTROL UNITS IN
EXACTLY THE SAME WAY
STRATIFY EACH PSU INTO A CERTAINTY STRATUM OF TREATMENT UNITS AND
A STRATUM OF NON-TREATMENT UNITS

IN LAST TWO CASES, MUST ADJUST FOR SELECTION EFFECTS BY REGRESSION
ANALYSIS.

RECOMMENDED APPROACH: MATCH ON A “DISTANCE” MEASURE (SCORE) THAT
TAKES INTO ACCOUNT THE RELATIVE IMPORTANCE OF EACH MATCH VARIABLE ON
OUTCOMES OF INTEREST. MATCH ALL UNITS OF THE POPULATION (FORM INTO
“MATCH SETS” (MATCHED PAIRS, TRIPLETS, QUADRUPLETS OR QUINTUPLETS, AS
DATA AVAILABILITY PERMITS – EXTRAS ARE INCLUDED FOR REPLACEMENTS)).
(BLOCKING.) ASSIGN ONE UNIT OF EACH PAIR TO TREATMENT AND ONE TO
CONTROL. APPLY MARGINAL STRATIFICATION TO THE TREATMENT UNITS TO SELECT
THE SAMPLE. PROBABILITIES OF INCLUSION OF EACH UNIT CAN BE DETERMINED
FROM THE PROBABILITIES OF SELECTION OF EACH UNIT.

[DESCRIBE IN GREATER DETAIL. SHOW FORMULA.]

AFTER FORM MATCH SETS (PAIRS OR GROUPS), RANDOMLY ASSIGN ONE TO
TREATMENT AND ONE TO CONTROL.

MATCHING FOR QUASI-EXPERIMENTAL DESIGNS

TREATMENT SAMPLE IS ALREADY SELECTED.

DO MATCHING IN THE USUAL WAY.

ESTIMATION OF SAMPLE SIZE

TWO MAIN AREAS OF STATISTICAL INFERENCE: ESTIMATION OF PARAMETERS AND
TESTING OF HYPOTHESES. DESCRIPTIVE SURVEYS ARE CONCERNED MORE WITH
PARAMETER ESTIMATION, AND ANALYTICAL SURVEYS ARE CONCERNED MORE WITH
TESTS OF HYPOTHESES.

APPROACH TO SAMPLE SIZE DETERMINATION:
SPECIFY FORMULA FOR PRECISION (DESCRIPTIVE SURVEYS) OR POWER
(ANALYTICAL SURVEYS) IN TERMS OF SAMPLE SIZE, AND SOLVE FORMULA FOR
SAMPLE SIZE

USUAL APPROACH (E.G., WILLIAM T. GRANT FOUNDATION “OPTIMAL DESIGN” (OD)
SOFTWARE): SPECIFY MODEL SOLELY IN TERMS OF SEVERAL VARIANCES OR FIXED-EFFECT SUMS OF SQUARES. DIFFICULT FOR COMPLEX MODELS (E.G., OD DOES NOT
ADDRESS DESIGNS INVOLVING MATCHING OR PANEL SAMPLING; USES “REPEATED
MEASURES WITH POLYNOMIAL TRAJECTORY”).

PROBLEM: NOT VERY INTUITIVE, AND DIFFICULT TO SPECIFY VALUES OF VARIANCES
OR FIXED-EFFECT SUMS OF SQUARES.
RECOMMENDED APPROACH: TO EXTENT POSSIBLE, SPECIFY MODEL IN TERMS OF CORRELATIONS. SIMPLER FORMULAS, MORE INTUITIVE.

NOTE: FOR THIS ANALYSIS, AND ALL TO FOLLOW, IT IS ASSUMED THAT SAMPLE SIZES ARE SUFFICIENTLY LARGE (E.G., >30) SO THAT THE LAW OF LARGE NUMBERS AND CENTRAL LIMIT THEOREMS APPLY. ALSO, THE FPC IS IGNORED.

PROPERTIES OF ESTIMATORS: BIAS (UNBIASEDNESS), PRECISION (MINIMUM VARIANCE), ACCURACY, CONSISTENCY.

POINT ESTIMATES (E.G., MEAN), INTERVAL ESTIMATES (E.G., CONFIDENCE INTERVAL).

SOME NOTATION:

$E(x) = \mu$ (or $\mu_x$)

$\text{var}(x) = E((x-\mu)^2) = \sigma^2$ (or $\sigma^2_x$)

$sd(x) = \sqrt{\text{var}(x)} = \sigma$

$\text{cov}(x,y) = E((x-\mu_x)(y-\mu_y))$

$\text{corr}(x,y) = \frac{\text{cov}(x,y)}{sd(x)sd(y)} = \rho$ or $\rho_{xy}$.

WILL MAKE USE OF

$E(x + y) = E(x) + E(y)$

$E(cx) = cE(x)$

$\text{var}(cx) = c^2 \text{var}(x)$

$\text{var}(x+y) = \text{var}(x) + \text{var}(y) + 2\rho_{xy} sd(x) sd(y) = \sigma_x^2 + \sigma_y^2 + 2\rho_{xy} \sigma_x \sigma_y$

(Check: If $x=y$, then $\rho_{xy}=1$, and $\text{var}(x+x) = \text{var}(2x) = 4 \text{var}(x)$ and $\text{var}(x-x) = \text{var}(0) = 0$.)

If $x$ and $y$ are independent, then $\rho_{xy} = 0$, so $\text{var}(x+y) = \text{var}(x) + \text{var}(y)$ and $\text{var}(x-y) = \text{var}(x) + \text{var}(y)$.

Define sample mean $= \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

The standard deviation of an estimate (e.g., the estimated mean) is called the standard error of the estimate.

$se(\bar{x}) = \sqrt{\text{var}(\bar{x})}$

For simple random sampling (SRS), can show $E(\bar{x}) = \mu$, so sample mean is an unbiased estimate of the population mean.
For simple random sampling (SRS), \( \text{var}(\bar{x}) = \sigma^2/n \) and \( \text{se}(\bar{x}) = \sigma/\sqrt{n} \).

For any design, \( \text{deff} \equiv (\text{variance for the design}) / (\text{variance for SRS}) \), so \( \text{var}(\bar{x}) = \text{deff} \sigma^2/n \) or \( \text{se}(\bar{x}) = \text{deff} \sigma/\sqrt{n} \).

E.g., for multistage sampling, \( \text{deff} = (1 + (m-1)\rho) \), so \( \text{var}(\bar{x}) = \text{deff} \sigma^2/n = (1 + (m-1)\rho) \sigma^2/n \).

If \( x \) is a binomial random variate (values 0 or 1, \( \text{Pr}(x=1) = p \); “sampling for proportions”), then \( E(x) = p \), \( \text{var}(x) = p(1-p) = pq \), and \( \text{var}(\bar{x}) = \text{var}(\hat{p}) = pq/n \).

**ESTIMATION OF SAMPLE SIZE FOR DESCRIPTIVE SURVEYS**

**STATISTICAL PRECISION ANALYSIS:** INVESTIGATION OF THE RELATIONSHIP BETWEEN SAMPLE SIZE AND THE PRECISION OF ESTIMATES.

**MOTIVATION FOR PRECISION ANALYSIS:** TO ASSURE THAT THE SAMPLE ESTIMATES WILL BE OF USEFUL PRECISION.

**TWO APPROACHES:** SPECIFY LEVEL OF PRECISION AND DERIVE CORRESPONDING SAMPLE SIZE; OR SPECIFY SAMPLE SIZE AND DERIVE CORRESPONDING PRECISION.

**NOTE:** “SAMPLE SIZE” REFERS BOTH TO PSU SAMPLE SIZE AND ULTIMATE-UNIT SAMPLE SIZE.

**CONCEPTS:** SAMPLING DISTRIBUTION OF AN ESTIMATOR; LAW OF LARGE NUMBERS; CENTRAL LIMIT THEOREM; POINT ESTIMATE; INTERVAL ESTIMATE; CONFIDENCE INTERVAL. THESE CONCEPTS ARE DISCUSSED IN ANY BOOK ON INTRODUCTION TO THE THEORY OF STATISTICS (E.G., *Introduction to the Theory of Statistics* by Alexander M. Mood, Franklin A. Graybill and Duane C Boes 3rd edition, McGraw Hill, 1974).

[GRAPH OF NORMAL DISTRIBUTION (DENSITY FUNCTION), SHOWING PERCENTILE POINTS \( \alpha/2 \) AND \( 1-\alpha/2 \)]

[GRAPH OF SAMPLING DISTRIBUTION (DENSITY FUNCTION) OF \( \bar{x} \), SHOWING A NUMBER OF DENSITIES CORRESPONDING TO DIFFERENT SAMPLE SIZES.]

**CONFIDENCE INTERVAL:**

Random sample \( x' = (x_1, \ldots, x_n) \) from density \( f(x; \theta) \).
Statistics \( T_1 = t_1(x) \) and \( T_2 = t_2(x) \) such that \( T_1 \leq T_2 \) for which \( P_{\theta}[T_1 < T(\theta) < T_2] = \gamma \), where \( \gamma \) does not depend on \( \theta \). (In the figure, \( \gamma = 1-\alpha \).)
The random interval \( (T_1, T_2) \) is called a 100\( \gamma \) percent confidence interval for \( T(\theta) \).
\( \gamma \) is called the **confidence coefficient**.
\( T_1 \) and \( T_2 \) (or realized values \( t_1 \) and \( t_2 \)) are called lower and upper confidence limits, respectively, for \( T(\theta) \).

**EXAMPLE:**

Sample mean \( \bar{x} \), sample standard deviation \( \hat{\sigma} \), \( \text{se}(\bar{x}) = \hat{\sigma}/\sqrt{n} \)
The statistic $z = \frac{\bar{x} - \mu}{se}$ has (approximately) a normal distribution, $N(\mu, \sigma) = N(0, 1)$.

Hence $P[-1.96 < Z < 1.96] = .95$

or $P[-1.96 < \frac{\bar{x} - \mu}{se} < 1.96] = .95$

or $P[-1.96se < \bar{x} - \mu < 1.96se] = .95$

or $P[\bar{x} - 1.96se < \mu < \bar{x} + 1.96se] = .95$

or $P[\bar{x} - 1.96se < \mu < \bar{x} + 1.96se] = .95$

So $(\bar{x} - 1.96se, \bar{x} + 1.96se)$ is a 95% confidence interval for $\mu$.

CASE 1: SINGLE SAMPLE: ESTIMATION OF POPULATION MEAN

PRECISION ANALYSIS: SPECIFY LEVEL OF PRECISION, AND DERIVE CORRESPONDING SAMPLE SIZE.

ESTIMATION OF SAMPLE SIZE FOR ESTIMATING A POPULATION MEAN, BASED ON A SINGLE SAMPLE.

POINT ESTIMATION: SPECIFY THE DESIRED STANDARD ERROR, $se$.

KNOW $se(\bar{x}) = \sqrt{deff \sigma / \sqrt{n}}$.

SET $se = \sqrt{deff \sigma / \sqrt{n}}$ AND SOLVE FOR $n$:

\[ n = \text{deff} \sigma^2 / se^2. \]

EXAMPLE 1.1. POINT ESTIMATION: DETERMINE SAMPLE SIZE REQUIRED TO ESTIMATE MEAN INCOME WITH A $10 STANDARD ERROR.

NEED TO KNOW STANDARD DEVIATION OF INCOME. ASSUME THAT IT IS $100$.

NEED TO KNOW TYPE OF SAMPLING (SAMPLE DESIGN). ASSUME SRS. HENCE deff=1.

\[ n = \text{deff} \sigma^2 / se^2 = 100^2 / 10^2 = 100. \]

EXAMPLE 1.2. POINT ESTIMATION: DETERMINE SAMPLE SIZE REQUIRED TO ESTIMATE MEAN INCOME WITH A RELATIVE STANDARD ERROR OF 5%. (NOTE: RELATIVE STANDARD ERROR = STANDARD ERROR / MEAN. RELATIVE STANDARD DEVIATION = $\sigma/\mu$ IS CALLED THE COEFFICIENT OF VARIATION. $\sigma^2/\mu^2$ IS CALLED THE RELATIVE VARIANCE, OR REL VARIANCE.)

NEED TO KNOW THE RELATIVE STANDARD ERROR OF INCOME. OFTEN IN THE RANGE 0.5 TO 2. ASSUME 1.0. ASSUME SRS, AS BEFORE.
\[ n = \text{deff} \sigma^2 / \text{se}^2 = \text{deff} (\sigma/\mu)^2 / (\text{se}/\mu)^2 = 1 \ 1^2 / .05^2 = 400. \]

**EXAMPLE 1.3. INTERVAL ESTIMATION:** DETERMINE SAMPLE SIZE SUCH THAT A 95% CONFIDENCE INTERVAL FOR THE MEAN IS OF SIZE (WIDTH) 10. HALF THE WIDTH OF THE 95% CONFIDENCE INTERVAL IS OFTEN CALLED THE ERROR BOUND, DENOTED BY E. SO IN THIS EXAMPLE, THE ERROR BOUND IS E = 10/2 = 5. OTHERWISE SAME AS EXAMPLE 1 (deff = 1, \( \sigma = 100 \)).

95% confidence interval: \( \bar{x} \pm 1.96 \text{se}(\bar{x}) = \bar{x} \pm E = \bar{x} \pm 5 \)

SO SET \( 1.96 \text{se}(\bar{x}) = E \), or \( 1.96 \sqrt{\text{deff} \sigma / \sqrt{n}} = E \), WHERE E = 5.

SOLVING FOR n:

\[ n = \text{deff} (1.96 \sigma / E)^2 = 1 (1.96 100 / 5)^2 = 1536.64 \approx 1,537. \]

**EXAMPLE 1.4. INTERVAL ESTIMATION:** 95% CONFIDENCE INTERVAL FOR PROPORTION. DETERMINE SAMPLE SIZE SUCH THAT A 95% CONFIDENCE INTERVAL FOR AN ESTIMATED PROPORTION IS +/- 3% (E.G., TELEVISION POLLS).

TO KNOW \( \sigma \), MUST SPECIFY \( p \). “WORST CASE” IS \( p = .5 \), FOR WHICH \( \sigma = \sqrt{p(1-p)} = .5 \).

PROCEEDING AS ABOVE,

95% confidence interval: \( \bar{x} \pm 1.96 \text{se}(\bar{x}) = \bar{x} \pm .03 \)

SO SET \( 1.96 \text{se}(\bar{x}) = .03 \), or \( 1.96 \sqrt{\text{deff} \sigma / \sqrt{n}} = .03 \).

SOLVING FOR n:

\[ n = \text{deff} (1.96 \sigma / .03)^2 = 1 (1.96 .5 / .03)^2 = 1,067.11 \approx 1,067. \]

MOST TELEVISION POLLS SET “SAMPLING ERROR” OR “ERROR BOUND” AT +/- 3%, AND USE A SAMPLE SIZE OF 1000-1200.

PRECEDING EXAMPLES ASSUMED SIMPLE RANDOM SAMPLING (SRS), WITH \( \text{deff} = 1 \).

IF USE CLUSTER SAMPLING OR MULTISTAGE SAMPLING, WILL HAVE A DIFFERENT VALUE FOR deff.

**EXAMPLE 1.5.** FOR EXAMPLE, SUPPOSE, IN EXAMPLE 1.3, THAT WE ARE USING A TWO-STAGE SAMPLE DESIGN – SAMPLES OF VILLAGES WITH SUBSAMPLES OF \( m = 10 \) HOUSEHOLDS PER VILLAGE. MUST SPECIFY A VALUE FOR THE INTRA-CLUSTER CORRELATION COEFFICIENT, \( icc \) (or \( \rho \)) AND THE WITHIN-PSU SAMPLE SIZE, \( m \), IN ORDER TO CALCULATE THE DESIGN EFFECT.

\[ \text{deff} = 1 + (m-1)icc. \]
THE VALUE OF icc DEPENDS ON THE VARIABLE BEING MEASURED. FOR MANY HOUSEHOLD VARIABLES, icc IS IN THE RANGE .05-.15. (ASSUME VALUE BASED ON EXPERIENCE, OR FROM ANALYSIS OF EXISTING DATA.)

\[
deff = 1 + (m-1)icc = 1 + (10 - 1).1 = 1.9
\]

\[
n = \text{deff} \left( \frac{1.96 \sigma}{5} \right)^2 = 1.9 \left( \frac{1.96 \ 100}{5} \right)^2 = 2919.16 \approx 2,919.
\]

THE SAMPLE SIZE VARIES AS \( \text{deff} \). THE "CLUSTER EFFECT" ON SAMPLE SIZE IS SUBSTANTIAL, EVEN FOR SMALL VALUES OF icc.

CASE 2: TWO SAMPLES; ESTIMATION OF DIFFERENCE IN GROUP MEANS ("SINGLE DIFFERENCE")

GROUPS 1 AND 2, GROUP MEANS \( \mu_1 \) AND \( \mu_2 \), GROUP STANDARD DEVIATIONS \( \sigma_1 \) AND \( \sigma_2 \), SAMPLE MEANS \( \bar{x}_1 \) and \( \bar{x}_2 \), INTERGROUP CORRELATION \( \rho_{12} \). (USE icc, INSTEAD OF \( \rho \), FOR INTRA-CLUSTER CORRELATION COEFFICIENT TO AVOID CONFUSION.)

ESTIMATE \( \bar{SD} = \bar{x}_1 - \bar{x}_2 \).

THEORY PROCEEDS SAME AS ABOVE (CASE 1), WHERE REPLACE \( \bar{x} \) by \( \bar{SD} \).

\[
\text{se}(\bar{SD}) = \sqrt{ \text{deff} \left( \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2\rho_{12}\sigma_1\sigma_2\sqrt{n_1n_2} \right) }.
\]

IF INDEPENDENT SAMPLES, THEN \( \rho_{12} = 0 \).

NOTE: FOR ESTIMATING DIFFERENCES, IT IS MOST EFFICIENT FOR THE GROUPS SAMPLE SIZES TO BE THE SAME (I.E., A "BALANCED" DESIGN). WILL ASSUME THIS IN THE EXAMPLES THAT FOLLOW. FOR SIMPLICITY, WILL ASSUME \( \sigma_1 = \sigma_2 = \sigma \). DENOTE \( n_1 = n_2 = n \). IN THIS CASE,

\[
\text{se}(\bar{SD}) = \sqrt{ \text{deff} \left( 2 \left( 1 - \rho_{12} \right) \sigma^2/n \right) }.
\]

NOTE THAT IF WE ASSUME THE SAME TOTAL SAMPLE SIZE AS IN CASE 1, THEN THE VALUE OF \( n \) IN THIS EXPRESSION (I.E., THE PER-GROUP SAMPLE SIZE) IS \( n/2 \) OF THE CASE 1 VALUE. IF \( \rho_{12} = 0 \) (INDEPENDENT SAMPLES), THE VARIANCE IS FOUR TIMES AS LARGE AS IN CASE 1, SO THE SAMPLE SIZE WILL BE FOUR TIMES AS LARGE.

EXAMPLE 2.1. (SIMILAR TO EXAMPLE 1.3.) INTERVAL ESTIMATION: DETERMINE SAMPLE SIZE SUCH THAT A 95% CONFIDENCE INTERVAL FOR THE DIFFERENCE IN MEANS IS OF SIZE (WIDTH) 10, I.E., ERROR BOUND \( E = 5 \). \( \text{deff} = 1 \). \( \sigma = 100 \). ASSUME INDEPENDENT SAMPLES, I.E., \( \rho_{12} = 0 \).

\[
95\% \text{ confidence interval: } \bar{SD} \pm 1.96 \text{se}(\bar{SD}) = \bar{SD} \pm E = \bar{SD} \pm 5
\]

SO SET \( 1.96 \text{se}(\bar{SD}) = E \), or \( 1.96 \sqrt{ \text{deff} \left( 2 \left( 1 - \rho_{12} \right) \sigma^2/n \right) } = E \), WHERE \( E = 5 \).

SOLVING FOR \( n \):

\[
n = \text{deff} \left( 2 \left( 1 - \rho_{12} \right) \left( 1.96 \sigma / E \right)^2 \right) = 12 \left( 1 - 0 \right) \left( 1.96 \ 100 / 5 \right)^2 = 3073.28 \approx 3,073.
\]
THIS IS THE \textit{per-group} sample size, i.e., the sample is \textit{four times} as large as in Example 1.3.

(Note: This result may seem to differ from the examples presented earlier, when examining the variance of matched samples. For the previous examples, the sample size was \(n\) for every group. Here it is \(n/2\).)

The sample sizes for estimating differences are very large, if use independent samples. Variance of the estimated difference is reduced by introducing correlations between the groups, e.g., by interviewing the same households in both survey rounds (in a \textit{before-after} design), or by matching treatment and control units (in a \textit{treatment-control} design).

**Example 2.2.** Same as Example 2.1, but \(\rho_{12} = .5\).

The per-group sample size decreases from 3,073 to half that value, 1,537.

Guidance for specifying \(\rho_{12}\).

For \textit{before-after} design, if interview same households in both survey rounds, set a high value for \(\rho_{12}\), such as .5.

For \textit{treatment-control} with matching of PSUs, set modest value, e.g., \(\rho_{12} = .3\).

**Case 3:** Four samples (treatment before, treatment after, comparison before, comparison after): estimation of double difference of group means.

Groups:

1: Treatment before
2: Treatment after
3: Comparison before
4: Comparison after

Group means \(\mu_1, \mu_2, \mu_3\) and \(\mu_4\), group standard deviations \(\sigma_1, \sigma_2, \sigma_3\) and \(\sigma_4\), sample means \(\bar{x}_1, \bar{x}_2, \bar{x}_3\) and \(\bar{x}_4\).

Intergroup correlations:

\begin{align*}
\text{Group 1 (TRT BEFORE)} & \quad \text{---} \rho_{12} \text{---} & \text{Group 2 (TRT AFTER)} \\
\rho_{13} & \quad \rho_{23} & \rho_{24} \\
\rho_{14} & \quad \rho_{24} & \text{---} \rho_{34} \text{---} \\
\text{Group 3 (COMP BEFORE)} & \quad \text{---} \rho_{34} \text{---} & \text{Group 4 (COMP AFTER)}
\end{align*}

Correlations \(\rho_{12}, \rho_{13}, \text{and } \rho_{34}\) are “real,” \(\rho_{23}, \rho_{24}, \text{and } \rho_{34}\) are “artifactual.”

Estimate \(\overline{DD} = (\bar{x}_2 - \bar{x}_1) - (\bar{x}_4 - \bar{x}_3)\).
THEORY PROCEEDS SAME AS ABOVE (CASE 1), WHERE REPLACE \( \bar{x} \) by \( \overline{DD} \).

\[
se(\overline{DD}) = \sqrt{ \text{deff} \left[ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + \frac{\sigma_3^2}{n_3} + \frac{\sigma_4^2}{n_4} - 2\rho_{12}\sigma_1\sigma_2\sqrt{\frac{1}{n_1 n_2}} - 2\rho_{13}\sigma_1\sigma_3\sqrt{\frac{1}{n_1 n_3}} + 2\rho_{14}\sigma_1\sigma_4\sqrt{\frac{1}{n_1 n_4}} + 2\rho_{23}\sigma_2\sigma_3\sqrt{\frac{1}{n_2 n_3}} - 2\rho_{24}\sigma_2\sigma_4\sqrt{\frac{1}{n_2 n_4}} - 2\rho_{34}\sigma_3\sigma_4\sqrt{\frac{1}{n_3 n_4}} \right] }.
\]

IF ALL FOUR SAMPLES ARE INDEPENDENT, THEN ALL OF THE \( \rho \)'s are zero.

IF DON'T INTERVIEW SAME HOUSEHOLDS IN BOTH ROUNDS, THEN \( \rho_{12} = 0 \).

IF DON'T MATCH TREATMENT AND COMPARISON UNITS, THEN \( \rho_{13} = 0 \).

GUIDANCE ON SETTING VALUES OF \( \rho \)'s: AS ABOVE (CASE 2).

NOTE THAT THE COVARIANCE MATRIX MUST BE POSITIVE DEFINITE.

NOTE: FOR ESTIMATING DIFFERENCES, IT IS MOST EFFICIENT FOR THE GROUPS SAMPLE SIZES TO BE THE SAME (I.E., A "BALANCED" DESIGN). WILL ASSUME THIS IN THE EXAMPLES THAT FOLLOW. FOR SIMPLICITY, WILL_ASSUME \( \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma \).

DENOTE \( n_1 = n_2 = n_3 = n_4 = n \). FOR SIMPLICITY, WILL (UNREALISTICALLY) ASSUME \( \rho_{12} = \rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = \rho_{34} = \rho \). IN THIS CASE,

\[
se(\overline{DD}) = \sqrt{ \text{deff} \left[ 4 (1 - \rho) \frac{\sigma^2}{n} \right] }.
\]

NOTE THAT IF WE ASSUME THE SAME TOTAL SAMPLE SIZE AS IN CASE 1, THEN THE VALUE OF \( n \) IN THIS EXPRESSION (I.E., THE PER-GROUP SAMPLE SIZE) IS \( n/4 \) OF THE CASE 1 VALUE. IF ALL \( \rho \)'s = 0 (E.G., INDEPENDENT SAMPLES), THE VARIANCE IS SIXTEEN TIMES AS LARGE AS IN CASE 1, SO THE SAMPLE SIZE WILL BE SIXTEEN TIMES AS LARGE.

EXAMPLE 3.1. (SIMILAR TO EXAMPLE 1.3.) INTERVAL ESTIMATION: DETERMINE SAMPLE SIZE SUCH THAT A 95% CONFIDENCE INTERVAL FOR THE DIFFERENCE IN MEANS IS OF SIZE (WIDTH) 10, I.E., ERROR BOUND \( E = 5 \). \( \text{deff} = 1 \). \( \sigma = 100 \). ASSUME INDEPENDENT SAMPLES, I.E., ALL \( \rho \)'s =0.

95% confidence interval: \( \overline{DD} \pm 1.96 se(\overline{DD}) = \overline{DD} \pm E = \overline{DD} \pm 5 \)

SO SET \( 1.96 se(\overline{DD}) = E \), or \( 1.96 \sqrt{ \text{deff} \left[ 4 (1 - \rho) \frac{\sigma^2}{n} \right] } = E \), WHERE \( E = 5 \).

SOLVING FOR \( n \):

\[
n = \text{deff} 4 (1 - \rho) \left( \frac{1.96 \sigma}{E} \right)^2 = 1 2 (1 - 0) \left( \frac{1.96 100 / 5}{1.96 100 / 5} \right)^2 = 6146.56 \approx 6,147.
\]

THIS IS THE PER-GROUP SAMPLE SIZE, I.E., THE SAMPLE IS SIXTEEN TIMES AS LARGE AS IN EXAMPLE 1.3.

THE SAMPLE SIZES FOR ESTIMATING DOUBLE DIFFERENCES ARE VERY LARGE, IF USE INDEPENDENT SAMPLES. VARIANCE OF THE ESTIMATED DOUBLE DIFFERENCE IS REDUCED BY INTRODUCING CORRELATIONS BETWEEN THE GROUPS, E.G., BY INTERVIEWING THE SAME HOUSEHOLDS IN BOTH SURVEY ROUNDS (IN A BEFORE-
AFTER DESIGN), OR BY MATCHING TREATMENT AND CONTROL UNITS (IN A
TREATMENT-CONTROL DESIGN).

EXAMPLE 3.2. SAME AS EXAMPLE 3.1, BUT \( p = .5 \).

THE PER-GROUP SAMPLE SIZE DECREASES FROM 6,147 TO HALF THAT VALUE, 3,073.

[HANDBOUTS: COMPUTER-PROGRAM PRINTOUTS]

ESTIMATION OF SAMPLE SIZE FOR ANALYTICAL SURVEYS

STATISTICAL POWER ANALYSIS: INVESTIGATION OF THE RELATIONSHIP BETWEEN
SAMPLE SIZE AND THE PROBABILITY OF DETECTING AN EFFECT OF A SPECIFIED
SIZE.

MOTIVATION FOR POWER ANALYSIS: LEHMANN: “THERE IS LITTLE POINT IN CARRYING
OUT AN EXPERIMENT WHICH HAS ONLY A SMALL CHANCE OF DETECTING THE
EFFECT BEING SOUGHT WHEN IT EXISTS.”

TWO APPROACHES: SPECIFY LEVEL OF POWER AND DERIVE CORRESPONDING
SAMPLE SIZE; OR SPECIFY SAMPLE SIZE AND DERIVE CORRESPONDING POWER.

NOTE: “SAMPLE SIZE” REFERS BOTH TO PSU SAMPLE SIZE AND TO ULTIMATE-UNIT
SAMPLE SIZE.

CONCEPTS (ADDITIONAL TO THOSE ADDRESSED ABOVE, IN SECTION ON
DESCRIPTIVE SURVEYS): STATISTICAL TESTS OF HYPOTHESIS.

[ONE-SAMPLE CASE: GRAPH OF SAMPLING DISTRIBUTION (DENSITY FUNCTION) OF \( \bar{X} \)
DISPLACED TO THE RIGHT OF THE ORIGIN, SHOWING A NUMBER OF DENSITIES
CORRESPONDING TO DIFFERENT SAMPLE SIZES.]

[TWO-SAMPLE CASE. SHOW TWO DENSITY FUNCTIONS. CAN REDUCE TO ONE-
SAMPLE CASE.]

STATISTICAL TESTS OF HYPOTHESIS:

Statistical hypothesis: A statement (assertion or conjecture) about the distribution of one
or more random variables. If the hypothesis completely specifies the distribution, it is a
simple hypothesis; otherwise it is a composite hypothesis. Denote an hypothesis by \( H \).

Test of a statistical hypothesis: a rule or procedure for deciding whether to reject a
statistical hypothesis. Denote a test by \( T \).

Nonrandomized test; critical region; critical function. Sample \( x' = (x_1, \ldots, x_n) \) in sample
space \( \chi \). Subset (region) \( C \epsilon \chi \). If reject \( H \) if and only if \( x \epsilon C \), then \( C \) is called the critical
region of the test. If define the indicator function \( \psi(x) = 1 \) if \( x \epsilon C \), 0 otherwise, then \( \psi \) is
called the critical function.
Null hypothesis; alternative hypothesis. Customary to refer to H as the null hypothesis, and denote it by $H_0$. The complementary hypothesis not-$H$ is referred to as the alternative hypothesis, denoted by $H_1$. If $H_0$ is not rejected, we say that it is accepted.

Types of error. Observe a sample and make a decision (accept or reject $H_0$). If reject $H_0$ when it is true, have made a Type I error. If accept $H_0$ when it is false, have made a Type II error. The probability of making a Type I error is called the size of the Type I error, and the probability of making a Type II error is called the size of the Type II error. If $H_0$ is a simple hypothesis, the size of the Type I error is called the size of the critical region, or the size of the test. Some writers refer to the size of the test as the significance level of the test.

Notation:

\[ \Pr[\text{Type I error}] = \alpha \text{ (significance level)} \]
\[ \Pr[\text{Type II error}] = \beta. \]

1 - $\beta$ is called the power of the test

Power function: Suppose that the distribution function of $x$ is a function of a parameter $\theta$ (denote $f(x; \theta)$ or $f_\theta(x)$). The power function of a test, $T$, denoted by $\pi(\theta)$, is the probability that $H_0$ is rejected when the sampling distribution is parameterized by $\theta$.

Simple likelihood-ratio test (simple null and alternative); most powerful test; Neyman-Pearson lemma.

Example, for a test of the hypothesis that the mean of a distribution is zero vs. the alternative that the mean is $D > 0$. (A “one-sided” test.)

[GRAPHS OF DENSITY FOR $H_0$ AND DENSITY FOR $H_1$.]

Case 1: Specify power and sample size, and determine minimum detectable effect, $D$.

Test statistic: sample mean, $\bar{x}$. Standard error $se(\bar{x})$. Population mean $\mu$. Use test of size (significance level) $\alpha$ and power $\beta$.

Want to determine the value, $D$, that the population mean must exceed. $D$ is called the “minimum detectable effect.”

\[ Pr\left[ \frac{\bar{x}}{se(\bar{x})} > z_{1-\alpha} | \mu = D \right] = 1 - \beta \]
\[ = Pr\left[ \frac{\bar{x} - D}{se(\bar{x})} > z_{1-\alpha} - \frac{D}{se(\bar{x})} | \mu = D \right] = 1 - \beta \]
\[ Pr\left[ z > z_{1-\alpha} - \frac{D}{se(\bar{x})} \right] = 1 - \beta \]
\[ \rightarrow z_{1-\alpha} - \frac{D}{se(\bar{x})} = z_\beta \]
\[ D = se(\bar{x})(z_{1-\alpha} - z_\beta) = se(\bar{x})(z_{1-\alpha} + z_{1-\beta}) \text{ since } z_{1-\beta} = -z_\beta. \]

STANDARD VALUES FOR $\alpha$ AND $\beta$ ARE $\alpha=.05$ AND $\beta=.1$ (POWER $1 - \beta = .90$), IN WHICH CASE $z_{1-\alpha} = 1.645$ AND $z_{1-\beta} = 1.282$. 

55
NOTE THAT THE PRECEDING IS A ONE-SIDED TEST. ONE-SIDED TESTS ARE USUAL IN EVALUATION, WHERE THE DIRECTION OF THE IMPACT IS RELATIVELY CERTAIN. (FOR TWO-SIDED TESTS, WOULD USE, FOR \( \alpha = .05 \), IN WHICH CASE \( z_{1-\alpha/2} = 1.96 \).

THE PRECEDING RESULT IS VERY GENERAL. IT MAY BE USED FOR ONE-GROUP, TWO-GROUP, AND FOUR-GROUP DESIGNS, WITH ANY TYPE OF SAMPLING (SRS, MULTISTAGE), AND ANY ESTIMATOR, \( \bar{x} \) (MEAN, SINGLE DIFFERENCE, DOUBLE DIFFERENCE), SIMPLY BY SUBSTITUTING THE APPROPRIATE VALUE FOR \( \text{se}(\bar{x}) \).

FOR EXAMPLE, FOR SIMPLE RANDOM SAMPLING WITH A ONE-GROUP DESIGN,

\[
\text{se}(\bar{x}) = \sigma / \sqrt{n}.
\]

FOR ANY OTHER TYPE OF SAMPLING (FOR A ONE-GROUP) DESIGN,

\[
\text{se}(\bar{x}) = \sqrt{\text{deff}} \sigma / \sqrt{n}.
\]

FOR EXAMPLE, WITH MULTISTAGE SAMPLING, \( \text{deff} = 1 + (m-1)\text{icc} \).

**Case 2.** One-group design. Specify \( \alpha, \beta, \) and \( D \), and determine the required sample size.

Use the same equation as above,

\[
D = \text{se}(\bar{x})(z_{1-\alpha} + z_{1-\beta})
\]

but substitute the appropriate formula for \( \text{se}(\bar{x}) \). That is,

\[
D = \text{se}(\bar{x})(z_{1-\alpha} + z_{1-\beta}) = (z_{1-\alpha} + z_{1-\beta}) \sqrt{\text{deff}} \sigma / \sqrt{n}.
\]

Solving for \( n \):

\[
n = \left( z_{1-\alpha} + z_{1-\beta} \right)^2 \text{deff} \sigma^2 / D^2.
\]

For example, suppose that we are using two-stage sampling with subsample size \( m=10 \) in each PSU. For a variable for which the intra-PSU correlation coefficient is \( \text{icc} = .1 \), the formula for the design effect is \( \text{deff} = 1 + (m-1)\text{icc} = 1 + (10 - 1).1 = 1.9 \). If \( \alpha = .05 \) and \( \beta = .1 \), then \( z_{1-\alpha} = 1.645 \) and \( z_{1-\beta} = 1.282 \). Let us suppose that the coefficient of variation of a variable of interest is 1.0, and that we wish to detect a minimum detectable effect equal to 10% of this value. Then \( \sigma/D = 1/1.1 = 10 \). Substituting these values yields

\[
n = (1.645 + 1.282)^2 1.9 10^2 = 1628.
\]

**SUMMARY:** TO ESTIMATE SAMPLE SIZE FOR AN ANALYTICAL SURVEY, THE FOLLOWING MUST BE SPECIFIED:

- THE IMPACT ESTIMATOR TO BE USED
- THE TEST PARAMETERS (POWER LEVEL, \( 1-\beta \), AND SIGNIFICANCE LEVEL, \( \alpha \))
- THE MINIMUM DETECTABLE EFFECT

56
CHARACTERISTICS OF THE SAMPLED POPULATION (MEANS, STANDARD DEVIATIONS, INTRA-PSU CORRELATION COEFFICIENT (IF MULTISTAGE SAMPLING IS USED)

THE SAMPLE DESIGN TO BE USED FOR THE SAMPLE SURVEY TO COLLECT QUANTITATIVE DATA

ALTERNATIVELY (RATHER THAN SPECIFYING THE POWER AND DETERMINING THE SAMPLE SIZE), IT MAY BE USEFUL TO SPECIFY THE SAMPLE SIZE AND DETERMINE THE POWER (OR THE POWER FUNCTION (I.E., THE POWER, AS A FUNCTION OF THE MINIMUM DETECTABLE EFFECT)).

THE PRECEDING EXAMPLES HAVE SHOWN FORMULAS FOR A ONE-GROUP DESIGN AND A MEAN. THE FORMULAS FOR TWO-GROUP DESIGN AND A SINGLE-DIFFERENCE ESTIMATOR, OR FOR A FOUR-GROUP DESIGN AND A DOUBLE-DIFFERENCE ESTIMATOR, ARE SIMILAR. ALL THAT CHANGES IS THE EXPRESSION FOR THE STANDARD ERROR OF THE ESTIMATOR.

FOR EXAMPLE, FOR A FOUR-GROUP DESIGN AND A DOUBLE-DIFFERENCE ESTIMATOR OF IMPACT, THE FORMULA FOR SAMPLE SIZE AS A FUNCTION OF THE VARIOUS PARAMETERS IS:

\[ n = \frac{(z_\alpha + z_\beta)^2 \text{varf}}{D^2} \]

where

- \( n \) = sample size
- \( D \) = minimum detectable effect
- \( z_\alpha \) = standard normal deviate having probability \( \alpha \) to the right, where \( \alpha \) denotes the significance level of the one-sided test of hypothesis that \( D \) exceeds zero (i.e., the probability of making a Type I error of deciding that \( D \) exceeds zero when in fact it does not)
- \( z_\beta \) = standard normal deviate having probability \( \beta \) to the right, where \( 1 - \beta \) denotes the power of the test (i.e., the probability of deciding that \( D \) exceeds zero when it does). (\( \beta \) denotes the probability of making a Type II error of deciding that \( D \) does not exceed zero when in fact it does.)
- \( \text{varf} \) = variance factor for impact estimator.

The value of var is given by

\[ \text{varf} = \text{deff} \left(1 - R^2\right) \left[ \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 - 2p_{12}\sigma_1\sigma_2 - 2p_{13}\sigma_1\sigma_3 + 2p_{14}\sigma_1\sigma_4 + 2p_{23}\sigma_2\sigma_3 - 2p_{24}\sigma_2\sigma_4 - 2p_{34}\sigma_3\sigma_4 \right] \]

where

- the four design groups are designated by the indices 1 (treatment before), 2 (treatment after), 3 (comparison before) and 4 (comparison after)
- \( \sigma_i^2 \) = variance for group i
- \( p_{ij} \) = coefficient of correlation between groups i and j
- \( \text{deff} \) = Kish’s design effect (to reflect the effect of two-stage sampling ("clustering"), if used) (\( \text{deff} \) is the ratio of the variance of the estimator under the design to the variance using a simple random sample of the same size)
- \( R^2 \) = multiple correlation coefficient (from regression model, if used) (this factor represents the effect of marginal stratification to achieve adequate variation in explanatory variables).

[EXAMPLES: HANDOUTS]
THE POWER ANALYSIS IS DONE TO ASSIST DETERMINATION OF SAMPLE DESIGN AND SAMPLE SIZE. ONCE THE SAMPLE HAS BEEN SELECTED, THE ROLE OF POWER ANALYSIS IS FINISHED.

SAMPLE SELECTION PROCEDURES

USUALLY SELECTION WITHOUT REPLACEMENT, WITH VARIABLE PROBABILITIES (TO EFFECT MARGINAL STRATIFICATION).

USUALLY MATCHING OF TREATMENT AND CONTROL PSUs.

USE A RANDOMIZED LIST OF THE ENTIRE POPULATION, TO ALLOW FOR SELECTION OF REPLACEMENTS.

REPLACEMENTS

WITH RANDOM LIST, SIMPLY GO FURTHER DOWN THE LIST FOR PSUS, SELECT REPLACEMENT UNITS FROM THE SAME PSU (TO REDUCE BIAS).

TO AVOID REPLACEMENT OF A MATCHED PSU PAIR WHEN ONE FAILS TO RESPOND, SELECT MATCH SETS (TRIPLETS, QUADRUPLETS).

WITHIN-PSU SELECTION

SYSTEMATIC SAMPLING WITH TWO RANDOM STARTS.

FROM LIST OR USING A STRUCTURED “RANDOM WALK.”

KEEP TRACK OF REPLACEMENTS, TO ENABLE CORRECT CALCULATION OF SAMPLE WEIGHTS.

REVIEW OF A TYPICAL SAMPLE DESIGN (HANDOUT).

SURVEY OF REFERENCES

QUESTIONS AND ANSWERS FOR DAY 1

OUTLINE OF TOPICS FOR SECOND DAY

END OF DAY 1
7. Day 2: Causal Analysis

**DAY 2: PARAMETRIC ESTIMATION, ANALYSIS OF BASELINE DATA, IMPACT ANALYSIS, USE OF STATA STATISTICAL PROGRAM PACKAGE**

7.1. Overview of Second Day’s Course Content

OVERVIEW OF SECOND DAY’S COURSE CONTENT

- **PARAMETRIC ESTIMATION** (GENERAL LINEAR STATISTICAL MODEL; LOGISTIC REGRESSION MODEL; TWO-STEP M-ESTIMATORS)
- **IMPACT ANALYSIS** (DESIGN-BASED ESTIMATES, MODEL-BASED ESTIMATES, MODEL SPECIFICATION, USE OF MATCHING TO REDUCE MODEL DEPENDENCE, ESTIMATORS BASED ON IGNORABILITY (CONDITIONAL INDEPENDENCE), SELECTION VS. OUTCOME MODELS (TWO-STEP MODELS), ESTIMATION OF STANDARD ERRORS, ESTIMATION BASED ON COMBINING REGRESSION AND PROPENSITY-SCORE MATCHING (MATCHING ESTIMATORS, DOUBLY-ROBUST MATCHING/REGRESSION APPROACH), INSTRUMENTAL-VARIABLE ESTIMATORS, LOCAL AVERAGE TREATMENT EFFECT (LATE), INTENTION-TO-TREAT EFFECT (ITT), REGRESSION DISCONTINUITY DESIGN, RESPONSES DISCRETE OR IN A LIMITED RANGE, MULTIVALUED (CATEGORICAL) TREATMENT VARIABLE, MULTIPLE TREATMENT VARIABLES, PANEL DATA (FIXED-EFFECTS MODEL, RANDOM-EFFECTS MODEL), TREATMENT OF MISSING VALUES, USE OF WEIGHTS, ESTIMATION OF STANDARD ERRORS)
- **ANALYSIS OF BASELINE DATA** (DESCRIPTIVE SURVEYS, ANALYTICAL SURVEYS)
- **USE OF STATA STATISTICAL PROGRAM PACKAGE** (SUMMARY DESCRIPTION AND EXAMPLES)
- **SURVEY OF REFERENCES**
- **QUESTIONS AND ANSWERS**

7.2. Parametric Estimation and Hypothesis Testing

**PARAMETRIC ESTIMATION AND HYPOTHESIS TESTING**

WILL FOCUS ON PARAMETER ESTIMATION RATHER THAN HYPOTHESIS TESTING

MANY PROCEDURES FOR ANALYSIS OF RELATIONSHIPS

- NONPARAMETRIC METHODS: CATEGORICAL DATA ANALYSIS (CROSS-CLASSIFIED TABLES, CONTINGENCY TABLES)
- PARAMETRIC METHODS: GENERAL LINEAR STATISTICAL MODEL (GLM, LEAST-SQUARES); MAXIMUM LIKELIHOOD (ML); M-ESTIMATORS; BAYESIAN ESTIMATES (MINIMIZE AVERAGE RISK/LOSS); MINIMAX ESTIMATES (MINIMIZE MAXIMUM RISK/LOSS).

WILL FOCUS ON PARAMETRIC METHODS (GLM, ML, M):
GENERAL LINEAR STATISTICAL MODEL (GLM): “MULTIPLE REGRESSION” MODEL (MR); ANALYSIS OF VARIANCE (ANOVA)
GENERALIZED LINEAR STATISTICAL MODEL (GLM): LOGISTIC REGRESSION MODEL (LR)
LIKELIHOOD RATIO: MAKE ASSUMPTIONS ABOUT THE UNDERLYING DISTRIBUTION
M-ESTIMATORS: USE A MORE GENERAL OPTIMIZATION CRITERION (USEFUL FOR NONLINEAR MODELS)

PROPERTIES OF ESTIMATORS:

UNBIASEDNESS: EXPECTED VALUE OF ESTIMATOR EQUALS VALUE OF PARAMETER BEING ESTIMATED
MINIMUM-VARIANCE: PRECISION IS NO LESS THAN OTHER ESTIMATORS BASED ON THE SAME SAMPLE
SUFFICIENCY: CONTAINS ALL OF THE INFORMATION IN THE SAMPLE THAT IS RELEVANT TO ESTIMATION OF THE PARAMETER
CONSISTENCY: CONVERGES TO THE DESIRED VALUE AS THE SAMPLE SIZE BECOMES LARGE (ASYMPTOTIC (LARGE-SAMPLE) PROPERTY, VS. SMALL-SAMPLE PROPERTY)
ASYMPTOTIC NORMALITY

NOTE THAT IN MANY INSTANCES, THE “SAMPLE ANALOG” OF A FUNCTION IS A REASONABLE ESTIMATOR FOR A PARAMETER, E.G., THE SAMPLE MEAN IS A REASONABLE ESTIMATOR FOR THE POPULATION MEAN. (COUNTEREXAMPLE: SAMPLE SPECTRUM IS NOT A CONSISTENT ESTIMATOR FOR THE TRUE SPECTRUM.) ALSO, SAMPLE MOMENTS ARE CONSISTENT ESTIMATORS OF POPULATION MOMENTS.

ESTIMATION PROCEDURES:

LEAST-SQUARES
NONLINEAR LEAST SQUARES
M-ESTIMATOR
TWO-STEP M-ESTIMATOR (FIRST-STEP SELECTION MODEL + SECOND-STEP OUTCOME MODEL)
MAXIMUM LIKELIHOOD

IN ECONOMIC APPLICATIONS, FOCUS MAINLY ON THE REDUCED MODEL, AND ON ESTIMATION OF A SINGLE EQUATION AT A TIME (LIMITED INFORMATION LEAST-SQUARES ESTIMATIONS VS. FULL INFORMATION MAXIMUM LIKELIHOOD (MULTIVARIATE ANALYSIS)).

ESTIMATION OF STANDARD ERRORS:

FOR SIMPLE MODELS, CLOSED-FORM EXPRESSIONS ARE AVAILABLE FOR COMPLEX MODELS, USE RESAMPLING METHODS (JACKKNIFE, BOOTSTRAP)
TESTS OF HYPOTHESIS (MORE LATER):

DIFFICULT, BECAUSE OF CONFOUNDING
PRINCIPLE OF CONDITIONAL ERROR

REFERENCES:


Technical references:


GENERAL LINEAR STATISTICAL MODEL (GENERAL LINEAR MODEL (GLM), LINEAR MODEL)

\[ y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_{kj} + e_j \]

where

\( j = \text{observation index}, \ j = 1, \ldots, n \)

\( n = \text{sample size} \)

\( y = \text{response variable (explained variable, dependent variable, regressand)} \)

\( \beta_i = \text{model parameter (regression coefficient)}, \ i = 1, \ldots, k \)
k = number of model parameters

$x_{ij} = \text{explanatory variable (independent variable, regressor), } i = 1,\ldots,k; j = 1,\ldots,n$

e_i = \text{model error term (or residuals), independently distributed and independent of the x's, with mean zero and variance } \sigma^2 \text{ (i.e., } E(e_i) = 0, \text{ var}(e_i) = \sigma^2, e_1 \text{╨ } e_2 \text{ and } e_1 \text{╨ } x_2 \text{ for all } j_1, j_2)\).

THE ASSUMPTION THAT THE ERRORS ARE INDEPENDENT CORRESPONDS TO THE STABLE UNIT TREATMENT VALUE ASSUMPTION (SUTVA) (NO-MACRO-EFFECTS ASSUMPTION, PARTIAL-EQUILIBRIUM ASSUMPTION). EXAMPLE: JOB-TRAINING PROGRAM IN A STABLE ECONOMY.

NOTE: IF THE FIRST VARIABLE IS UNITY, THEN $\beta_1 = \text{mean}$. IF THE FIRST VARIABLE IS UNITY, THEN IT IS CUSTOMARY TO START THE INDEX AT 0 (SO THAT THE MEAN IS $\beta_0$).

Will usually drop the observation index, and write:

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_k + e$$

MATRIX NOTATION: THE PREVIOUS EQUATION IS WRITTEN AS

$$y = x' \beta + e$$

where $\beta' = (\beta_1,\ldots,\beta_k)$ and $x' = (x_1,\ldots,x_n)$ (boldface denotes vectors; prime denotes matrix transpose; absence of a prime on a vector denotes a column vector; prime on a vector denotes a row vector). (Note: Contrary to standard vector notation, Wooldridge defines the model in terms of a row vector $x$: $y = x\beta + e$.)

and the full sample is written as

$$y = X\beta + e$$

where $y' = (y_1,\ldots,y_n)$, $X' = (x_1,\ldots,x_n)$ and $e' = (e_1,\ldots,e_n)$.

X IS CALLED THE DATA MATRIX. IN MOST INSTANCES, $x_i=1$, AND $\beta_1$ IS THE MEAN.

THE “LEAST-SQUARES” (GAUSS-MARKOV) ESTIMATE OF $\beta$ MINIMIZES THE SUM OF SQUARES OF THE RESIDUALS:

$$\Sigma_{i=1}^N (y_i - x_i \beta)^2.$$

UNDER THE ASSUMPTION THAT THE MATRIX $X'X$ (THE “CROSSPRODUCTS” MATRIX) IS OF FULL RANK (INVERTIBLE), IT IS GIVEN BY

$$\beta = (X'X)^{-1} X'y.$$ 

(THE “INVERTIBLE” CONDITION MEANS THAT THERE CANNOT BE ANY LINEAR DEPENDENCIES IN THE x VARIABLES. FOR EXAMPLE, INCLUDE ONLY $n-1$ INDICATOR VARIABLES (“DUMMY” VARIABLES) FOR A CATEGORICAL VARIABLE HAVING $n$
CATEGORIES. THE ESTIMATOR IS DEFINED IN TERMS OF CONDITIONAL INVERSES OTHERWISE.)

THE LEAST-SQUARES ESTIMATES ARE OF INTEREST BECAUSE THEY ARE SIMPLE IN FORM (MATRIX OPERATIONS) AND HAVE DESIRABLE STATISTICAL PROPERTIES (E.G., UNBIASEDNESS, CONSISTENCY, ASYMPTOTIC EFFICIENCY, ASYMPTOTIC NORMALITY). THEY WORK WELL FOR MANY PROBLEMS.

IN THE CASE OF A MODEL WITH A MEAN AND A SINGLE EXPLANATORY VARIABLE,

\[ y = \beta_0 + \beta_1 x + e \]

THE LEAST-SQUARES ESTIMATES ARE

\[ \hat{\beta}_0 = \bar{x} = \text{sample mean} \]
\[ \hat{\beta}_1 = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{\text{(sample covariance of } x \text{ and } y)}{\text{(sample variance of } x)} \]

WHY IS THE ASSUMPTION THAT THE EXPLANATORY VARIABLES AND THE ERROR TERMS ARE UNCORRELATED IMPORTANT?

CONSIDER THE SIMPLE EXAMPLE:

Model \( y = \beta x + e \)

where \( y, x \) and \( e \) have zero mean, but \( x \) may be correlated with \( e \).

Multiply through by \( x \):

\[ xy = \beta x^2 + xe. \]

Take expectations:

\[ \text{Exy} = \beta \text{Ex}^2 + \text{E}(xe) \]

or, because of zero means,

\[ \text{cov}(x,y) = \beta \text{var}(x) + \text{cov}(x,e). \]

Solving for \( \beta \):

\[ \beta = \frac{\text{cov}(x,y)}{\text{var}(x)} - \frac{\text{cov}(x,e)}{\text{var}(x)}. \]

But the correct estimate for \( \beta \) is the first term on the right-hand side. The second term, which is the coefficient of the regression of \( e \) on \( x \), is a bias.

ENDOGENEITY (E.G., SIMULTANEOUS EQUATIONS) IS A CAUSE OF CORRELATION BETWEEN REGRESSORS AND ERROR TERMS.

Income-determination model consisting of a consumption function and an income identity:

\[ C_t = \alpha + \beta Y_t + u_t \]
\[ Y_t = C_t + Z_t \]

where

C = consumption expenditure
Y = income
Z = nonconsumption expenditure
\( u = \) stochastic disturbance term
\( t = \) time period.

Substituting the first equation in the second:

\[ Y_t = \alpha + \beta Y_t + Z_t + u_t \]

or

\[ Y_t = \alpha/(1-\beta) + Z_t/(1 - \beta) + u_t/(1 - \beta) \]

so that \( Y_t \) is influenced by \( u_t \).

\[ E(Y_t) = \alpha/(1 - \beta) + Z_t/(1 - \beta) \]

and

\[ E(u_t[Y_t - E(Y_t)]) = E(u_t^2)/(1 - \beta) \neq 0. \]

The disturbance term and the explanatory variable in the consumption equation are correlated, and direct application of OLS to that equation will not yield unbiased or consistent estimates of \( \alpha \) and \( \beta \).

IN MANY APPLICATIONS, THE \( x \) ARE FIXED EFFECTS (FIXED NUMBERS, NOT RANDOM VARIABLES), AND THEREFORE CERTAINLY UNCORRELATED WITH THE MODEL ERROR TERMS. UNDER THE STATED ASSUMPTIONS, THE LEAST-SQUARES ESTIMATES ARE UNBIASED AND MINIMUM VARIANCE.

IN ECONOMIC (AND EVALUATION) APPLICATIONS, THE \( x \) ARE RANDOM VARIABLES. UNDER THE STATED ASSUMPTIONS, THE LEAST-SQUARES ESTIMATES ARE CONSISTENT AND ASYMPTOTICALLY NORMALLY DISTRIBUTED.

VARIANCE OF ESTIMATORS DEPENDS ON ASSUMPTIONS (FIXED EFFECTS, HOMOSCEDASTICITY).

IF FIXED EFFECTS AND CONSTANT VARIANCE, COVARIANCE MATRIX OF THE \( \beta \)'s IS

\[ \sigma^2(X'X)^{-1} \]

WHERE ESTIMATE \( \sigma^2 \) BY
\[ \sigma^2 = \frac{1}{n-k} (y - X\hat{\beta})^T (y - X\hat{\beta}) . \]

THIS MODEL IS CALLED A LINEAR MODEL BECAUSE IT IS LINEAR IN THE PARAMETERS. THE EXPLANATORY VARIABLES MAY BE NONLINEAR (E.G., \( x^2 \), \( x_1x_2 \), \( \ln(x) \)).

THE EXPLANATORY VARIABLES MAY BE CONTINUOUS OR DISCONTINUOUS. THE RESPONSE VARIABLE IS MORE CONSTRAINED (SO THAT LINEARITY ASSUMPTION HOLDS).

FOR EXAMPLE, AN INDICATOR (BINARY, DICHOTOMOUS) VARIABLE THAT DENOTES SEX:

\( x = 1 \) for males, 0 for females

OR A DESIGN VARIABLE:

\( x = 1 \) for region A, 0 otherwise.

OR THE INTERACTION EFFECT OF TREATMENT AND TIME:

\( x_1 \) (treatment) = 1 if treated, 0 if not treated

\( x_2 \) (time, survey round) = 0 for time 1 (before), 1 for time 2 (after).

Measure of impact = interaction effect of treatment and time = coefficient of \( x_1x_2 \) in regression model

THREE KEY ASSUMPTIONS:

**HOMOSCEDASTICITY:** VARIANCE OF THE MODEL ERROR TERMS IS CONSTANT
**INDEPENDENCE:** MODEL ERROR TERMS ARE INDEPENDENT OF EACH OTHER AND OF THE \( x \)'s (NO ENDOGENEITY; SUTVA)
**LINEARITY:** THE MODEL IS LINEAR IN THE PARAMETERS

IF THESE ASSUMPTIONS ARE VIOLATED, THE PARAMETER ESTIMATES MAY BE SERIOUSLY BIASED, AND MODEL PREDICTIONS MAY BE ABSURD (E.G., LINEAR PROBABILITY MODEL (LPM)).

APPROACHES WHEN ASSUMPTIONS ARE VIOLATED

**HETEROSCEDASTICITY**

VARIANCE-STABILIZING TRANSFORMATION

IF THE “RAW” DATA DO NOT SATISFY THE HOMOSCEDASTICITY ASSUMPTION, IT MAY BE POSSIBLE TO TRANSFORM THE RESPONSE VARIABLE (\( y \)) SO THAT IT DOES. EXAMPLE: LOGARITHM OF INCOME MAY HAVE CONSTANT VARIANCE.
HETEROSCEDASTICITY-CONSISTENT (HETEROSCEDASTIC-ROBUST) STANDARD ERRORS (EIKER-HUBER-WHITE (BOSE ROY JOHNSON HOTELLING? – IN USE AT UNC CHAPEL HILL IN 1962) SANDWICH ESTIMATOR).

IF MODEL IS NOT WELL-SPECIFIED IN TERMS OF CAUSAL VARIABLE, INCLUDE DESIGN PARAMETERS (E.G., CLUSTER VARIABLE).

FIXED EFFECTS VS. RANDOM EFFECTS

ENDOGENEITY

ENDOGENEITY: CORRELATION OF AN x WITH e. USUALLY THE RESULT OF:

OMITTED VARIABLES
MEASUREMENT ERROR ("ERRORS IN VARIABLES")
SIMULTANEITY (SIMULTANEOUS-EQUATION MODEL)

METHODS FOR DEALING WITH ENDOGENEITY:

INDIRECT LEAST-SQUARES: PUT THE SIMULTANEOUS EQUATION IN "REDUCED FORM" (A SINGLE EQUATION), AND APPLY ORDINARY LEAST SQUARES (OLS).
METHOD OF INSTRUMENTAL VARIABLES: TWO-STAGE LEAST SQUARES (2SLS)
FULL-INFORMATION MAXIMUM LIKELIHOOD (FIML)

METHOD OF INSTRUMENTAL VARIABLES (IV).

IF AN x IS CORRELATED WITH THE e’s, REPLACE IT WITH A VARIABLE (z) THAT, CONDITIONAL ON THE OTHER x’s, IS CORRELATED WITH IT BUT NOT WITH THE ERROR TERMS. SUCH A VARIABLE IS CALLED AN INSTRUMENTAL VARIABLE (IV).

CONCEPTUALLY, A "TWO-STEP" PROCEDURE IS USED: REGRESS THE ENDOGENOUS VARIABLE ON THE IV AND THE OTHER (EXOGENOUS) VARIABLES, AND REPLACE THE ENDOGENOUS VARIABLE WITH THIS ESTIMATE \(\hat{z}\) IN THE REGRESSION MODEL.
SHORTCOMING: ESTIMATES OF ERRORS NOT CORRECT.

CORRECT PROCEDURE (TWO-STAGE LEAST SQUARES, 2SLS): IF \(x_k\) IS THE ENDOGENOUS VARIABLE, THEN DENOTE \(z' = (x_1, x_2, \ldots, x_{k-1}, z_k)\)

\[ \hat{\beta} = (Z' X)^{-1}Z' y. \]

DIFFICULTIES WITH IV METHOD:

IV ESTIMATES ARE BIASED (BUT CONSISTENT)
IF IV IS NOT HIGHLY CORRELATED WITH THE ENDOGENOUS VARIABLE, THE METHOD WORKS POORLY: BIAS AND LOW PRECISION. MAY BE BETTER TO USE OLS EVEN THOUGH BIASED.

NONLINEARITY

EXAMPLE: THE CASE OF A BIVARIATE RESPONSE VARIABLE, E.G., LINEAR PROBABILITY MODEL (LPM). THIS MODEL DOES NOT FIT WELL. IF A LINEAR MODEL IS
USED IN THIS SITUATION, SOME OF THE PREDICTIONS WILL BE OUTSIDE THE RANGE (0,1). MUCH BETTER TO TRANSFORM THE MODEL PARAMETER (I.E., REPARAMETERIZE THE MODEL).

THIS LEADS US TO THE GENERALIZED LINEAR MODEL (ALSO GLM).

GENERALIZED LINEAR MODEL

THE GENERAL LINEAR MODEL IS OF THE FORM

$$E(y) = \mu = x'\beta.$$  

THE GENERALIZED LINEAR MODEL IS OF THE FORM

$$g(\mu) = x'\beta.$$  

THE FUNCTION $g(\mu)$ IS CALLED A LINK FUNCTION.

EXAMPLE. BINARY RESPONSE VARIABLE. LOGISTIC REGRESSION MODEL.

SUPPOSE $y$ IS A BINARY VARIABLE, E.G.,

$$y = 1 \text{ if treated, } 0 \text{ if not treated}$$

Define

$$p = \text{probability of treatment}$$

and the link function

$$g(p) = \log(p/(1-p)) = x'\beta$$

that is,

$$p = e^{x'\beta} / (1 + e^{x'\beta})$$

NOTE THAT IN THIS MODEL, $p$ IS RESTRICTED TO THE INTERVAL (0,1).

THE ONLY PROBLEM WITH THIS FORMULATION IS THAT THE METHOD OF LEAST SQUARES CANNOT BE USED TO ESTIMATE THE MODEL PARAMETERS. INSTEAD, THE METHOD OF MAXIMUM LIKELIHOOD IS USED.

ALTERNATIVES TO LOGISTIC REGRESSION MODEL: PROBIT REGRESSION MODEL

ADVANTAGES OF LOGISTIC REGRESSION MODEL (LOG ODDS MODEL):

FOR DISCUSSION OF THE GENERALIZED LINEAR MODEL, SEE THE FOLLOWING REFERENCES:

Generalized Linear Model

Response variable Y (binary).

Latent variable X (continuous, \(-\infty < x < \infty\))

Distribution function

\[ F(x) = P(X \leq x), \quad -\infty < x < \infty \]

For example, \( F(x) \) may be the normal distribution, in which case the model is called a probit model, or a logistic distribution, in which case the model is called a logit or logistic model. From Agresti (1996), “The probit model was introduced in 1934 for models in toxicology. The logistic regression model was not studied until about a decade later, but it is now much more popular than the probit. Partly this is because one can also interpret the logistic regression effects using odds ratios. Thus, one can fit those models to data from case-controlled studies, because one can estimate odds ratios for such data.”

Link function \( g(.) \) specifies how \( \mu = E(Y) \) relates to the linear function of the explanatory variables:

\[ g(\mu) = x'\beta \]

Mean function \( \mu(.) \) specifies how the linear function relates to the mean:

\[ \mu(x'\beta) = g^{-1}(x'\beta) \]

Simplest link function \( g(\mu) = \mu \) corresponds to general linear model.

Canonical parameter, \( \theta \): a “natural” parameter for a distribution, such as the mean, \( \mu \), for a normal distribution, \( p \) for a binomial distribution, and \( \lambda \) for a Poisson distribution.

Canonical link function: the one that expresses the canonical parameter (\( \theta \)) in terms of \( \mu \), i.e., \( \theta = b(\mu) \).

Table of canonical link functions (from Wikipedia
So, for example, if the raw data followed a Poisson distribution, we would use a Poisson loglinear model, which is a GLM that assumes a Poisson distribution for Y and uses the log link function:

$$\log \mu = \alpha + \beta x,$$

$$\mu = \exp(\alpha + \beta x).$$

### Canonical Link Functions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Name</th>
<th>Link Function</th>
<th>Mean Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Identity</td>
<td>$X\beta = \mu$</td>
<td>$\mu = X\beta$</td>
</tr>
<tr>
<td>Exponential Gamma</td>
<td>Inverse</td>
<td>$X\beta = \mu^{-1}$</td>
<td>$\mu = (X\beta)^{-1}$</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>Inverse squared</td>
<td>$X\beta = \mu^{-2}$</td>
<td>$\mu = (X\beta)^{-1/2}$</td>
</tr>
<tr>
<td>Poisson</td>
<td>Log</td>
<td>$X\beta = \ln(\mu)$</td>
<td>$\mu = \exp(X\beta)$</td>
</tr>
<tr>
<td>Binomial</td>
<td>Logit</td>
<td>$X\beta = \ln\left(\frac{\mu}{1 - \mu}\right)$</td>
<td>$\mu = \frac{1}{1 + \exp(-X\beta)}$</td>
</tr>
<tr>
<td>Multinomial</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

M-ESTIMATORS, NONLINEAR LEAST SQUARES (NLS) ESTIMATORS

MORE GENERAL FRAMEWORKS (THAN ORDINARY LEAST SQUARES) FOR PARAMETER ESTIMATION INCLUDE NONLINEAR LEAST SQUARES AND M-ESTIMATION.

SUPPOSE

$$E(y \mid x) = m(x, \theta_0)$$ for some parameter value $\theta_0$ (a "correctly specified model for the conditional mean")

ESTIMATE $\theta$ AS THE VALUE OF $\theta_0$ THAT MINIMIZES THE EXPECTED SQUARED ERROR BETWEEN $y$ AND $m(x, \theta)$, I.E., $\theta_0$ SOLVES THE PROBLEM

$$\min_\theta E[(y - m(x, \theta))^2].$$

THE NLS ESTIMATOR OF $\theta_0$, $\hat{\theta}$, SOLVES THE PROBLEM

$$\min_\theta n^{-1} \sum_{i=1}^n [y_i - m(x_i, \theta)]^2$$

MORE GENERALLY, LET $q(x)$ DENOTE A FUNCTION OF THE RANDOM VARIABLE $x$. AN M-ESTIMATOR OF $\theta_0$ SOLVES THE PROBLEM

$$\min_\theta n^{-1} \sum_{i=1}^n q(x_i, \theta).$$

THIS IS THE SAMPLE ANALOGY OF SOLVING THE POPULATION PROBLEM

$$\min_\theta E(q(x, \theta)).$$

(THE PRECEDING SUMMARY IS A GENERALIZATION. FOR NONLINEAR MODELS, IT CAN BE SHOWN THAT THE ESTIMATE FOR WHICH THE CONDITIONAL EXPECTATION OF THE REGRESSION-MODEL ERROR TERMS, GIVEN THE EXPLANATORY VARIABLES, IS THE LEAST-SQUARES ESTIMATOR. SEE WOOLDRIDGE OP. CIT, PP. 397-401.)

RECOMMENDED APPROACH TO MODEL SELECTION: SEEK A MATHEMATICAL REPRESENTATION THAT HAS HIGH “FACE VALIDITY” WITH RESPECT TO REPRESENTING THE CAUSAL MODEL. E.G., USE OF A FIRST-STEP LOGISTIC REGRESSION SELECTION MODEL FOLLOWED BY A SECOND-STEP OUTCOME REGRESSION MODEL. FIXED-EFFECTS VS. RANDOM –EFFECTS ASSUMPTION GUIDED BY CAUSAL MODEL.

THE CAUSAL MODEL MAY BE REPRESENTED BY STRUCTURAL EQUATIONS. THE STATISTICAL MODEL (ESTIMABLE MODEL) IS USUALLY A “REDUCED-FORM” SINGLE-EQUATION REPRESENTATION.

RECOMMENDED APPROACH TO ESTIMATION OF STANDARD ERRORS: FOR SIMPLE MODELS, USE STANDARD (CLOSED-FORM) ESTIMATORS. USE HETEROSCEDASTICITY-ROBUST ESTIMATORS WHEN APPROPRIATE. IF NEED TO ACCOUNT FOR DESIGN FEATURES (E.G., CLUSTER), SEEK A BETTER MODEL. FOR COMPLEX MODELS, USE BOOTSTRAPPING.

SURPRISING RESULT (pp. 500-502 and 925 of Wooldridge) (WILL USE LATER): CONSIDER A “TWO-STEP” M-ESTIMATOR, SUCH AS ESTIMATING THE PROBABILITY OF SELECTION IN THE FIRST STEP USING VARIATES z AND USING THAT ESTIMATE TO ESTIMATE OUTCOME IN THE SECOND STEP USING VARIATES v. ASSUME THAT THE FIRST-STEP ESTIMATOR IS A MAXIMUM LIKELIHOOD ESTIMATOR. UNDER THE CONDITIONAL INDEPENDENCE ASSUMPTION THAT v AND w (TREATMENT) ARE INDEPENDENT CONDITIONAL ON z, THEN THE STANDARD ERRORS OBTAINED BYignoring THE VARIABILITY IN THE FIRST-STEP ESTIMATOR ARE CONSERVATIVE.

ASSESSMENT OF QUALITY OF ESTIMATES

CANNOT INFER CAUSALITY FROM SAMPLE DATA, ANY MORE THAN CAN INFER RANDOMNESS FROM SAMPLE DATA. RANDOMNESS IS A PROCESS, NOT AN ASPECT OF THE DATA. CAUSALITY IS AN INTRINSIC PROPERTY OF PHYSICAL EXISTENCE, NOT AN ASPECT OF THE DATA.

So estimation of precision is not a problem, as long as the sample size is adequate.

Degrees of freedom: sample size less number of parameters estimated.

For complex models, prefer several hundred degrees of freedom for error-variance estimation.

Tukey: “With five parameters I can fit an elephant.” Disregarding “nuisance” parameters, a statistical model should be such a good representation of reality that only a few parameters are required. Example: Box-Jenkins time series models usually require just 2-5 parameters.

Overall model fit:
- Model face validity
- Measures of model fit (for comparing alternative models):
  - Coefficient of determination ($R^2 = \text{square of multiple correlation coefficient}$)
  - Various fit measures, such as:
    - Chi-square statistic
    - Root mean square error of approximation (RMSEA)

$$RMSEA = \sqrt{\frac{\chi^2}{n-1}}$$

Akaike information criterion (AIC): $= \chi^2 + k(k-1) - 2df$ where $k$ is the number of parameters and $df$ is the degrees of freedom of the model, or $-2 \ln(L) + 2k$ where $\ln(L)$ is the maximized log-likelihood of the model.

Bayesian information criterion (BIC) $= \chi^2 + \ln(n)[k(k-1)/2 - df]$ or $-2 \ln(L) + k \ln(n)$.

(Reference: Measuring Model Fit, by David A. Kenny, posted at http://davidakenny.net/cm/fit.htm (September 4, 2011).)

Alternative model representations

Because of multicollinearity (correlation among variables; lack of orthogonality) many alternative representations have similar fit (Simpson's paradox). Because of lack of forced changes, cannot interpret model coefficients as response to forced changes. Include explanatory variables, but do not read much into individual coefficients, other than forced-change ones (such as treatment). Estimates provide hypotheses, not conclusions.

Example:

Two degrees of freedom associated with a variable.

Representation 1: linear and quadratic components
ORTHOGONAL POLYNOMIALS
IF COMPONENTS ARE CORRELATED, EFFECTS ARE CONFOUNDED
IF COMPONENTS ARE ORTHOGONAL, EFFECTS ARE UNCORRELATED

If \( x = 0, 1 \) and 2, then \( x \) and \( x^2 \) are correlated, and effects confounded.

\[
\begin{array}{ccc}
\text{x} & \text{x}^2 \\
0 & 0 \\
1 & 1 \\
2 & 4 \\
\end{array}
\]

Cross-product (vector inner product): \( \mathbf{x}'\mathbf{x}^2 = 0 \times 0 + 1 \times 1 + 2 \times 4 = 0 + 1 + 8 = 9 \)

\[
\text{sd}(x) = **
\]

\[
\text{sd}(x^2) = **
\]

\[
\text{cov}(x, x^2) = **
\]

\[
\text{corr}(x, x^2) = **
\]

If define \( x \) with three values, \(-1, 0 \) and 1, then \( x \) and \( x^2 \) are orthogonal.

\[
\begin{array}{ccc}
\text{x} & \text{x}^2 \\
-1 & 1 \\
0 & 0 \\
1 & 1 \\
\end{array}
\]

Cross-product: 0

\[
\text{sd}(x) =
\]

\[
\text{sd}(x^2) =
\]

\[
\text{cov}(x, x^2) = 0
\]

\[
\text{corr}(x, x^2) = 0
\]

IN THIS CASE, WITH ZERO CORRELATION BETWEEN \( x \) AND \( x^2 \), THE LINEAR AND QUADRATIC EFFECTS ARE UNCONFOUNDED.

WHEN INCLUDE INTERACTION EFFECTS IN A MODEL, MAKE SURE TO DE-MEAN THE COVARIATES, OR ELSE COEFFICIENT OF INTERACTION OF TREATMENT AND TIME WILL NOT BE UNBIASED ESTIMATE OF IMPACT. WILL DISCUSS LATER.

Continuing the example, suppose that a variable has three levels, involving two sub-variables:

Type of ownership of a firm:
- Private
- Public for-profit
- Public not-for-profit

Define indicator variables \( x_1 \) for private/ profit and \( x_2 \) for-profit/not-for-profit.

\[
\begin{array}{cc}
\text{x}_1 & \text{x}_2 \\
\text{Private} & -2 & 0 \\
\text{Pubic for-profit} & 1 & -1 \\
\text{Public non-for-profit} & 1 & 1 \\
\end{array}
\]
Cross-product: 

\[-2x_0 + 1x(-1) + 1x1 = 0.\]

Variables are orthogonal.

Model coefficients are not confounded, easy to interpret.

THESE EFFECTS (DEFINED BY SUBSTANTIVELY MEANINGFUL COMPARISONS (CONTRASTS)) ARE MUCH MORE INFORMATIVE (GREATER FACE VALIDITY) THAN THE LINEAR AND QUADRATIC COMPARISONS, WHICH HAVE NO SUBSTANTIVE MEANING. BOTH REPRESENTATIONS USE TWO DEGREES OF FREEDOM, AND WOULD PROVIDE IDENTICAL COEFFICIENT OF DETERMINATION (R²). THE INDIVIDUAL EFFECTS (SINGLE DEGREES OF FREEDOM) CAN BE INTERPRETED IN ONE REPRESENTATION, BUT NOT THE OTHER. (EXTREME EXAMPLE: 10 DISTRICTS. DEFINE 10 DISTRICT INDICATOR VARIABLES. WOULD NEVER CONSIDER REPRESENTING DISTRICT USING A NINE-DEGREE POLYNOMIAL.) DON’T BLINDLY INTRODUCE A MÉLANGE OF FUNCTIONAL FORMS INTO A MODEL (LINEAR, QUADRATIC, LOGARITHMIC). ("FLEXIBLE" REGRESSION MODELS.)

A NOTE ON UNBIASEDNESS:

AN ESTIMATED MEAN MAY BE UNBIASED EVEN THOUGH THE MODEL COEFFICIENTS ARE BIASED. FOR EXAMPLE, MAY WISH TO FORECAST (PREDICT) SALES (Y), AS A FUNCTION OF A NUMBER OF OBSERVED VALUES (X’s), INCLUDING ADVERTISING BUDGET. IN THIS CASE, THERE IS NO NEED FOR UNBIASED OR CONSISTENT ESTIMATES OF THE MODEL COEFFICIENTS. IF, HOWEVER, WE WISH TO PREDICT THE EFFECT OF A FORCED CHANGE IN THE ADVERTISING BUDGET, WE WANT A GOOD ESTIMATE OF THE EFFECT OF CHANGING THE BUDGET, IN WHICH CASE THE COEFFICIENT MATTERS VERY MUCH.

7.3. Impact Analysis

IMPACT ANALYSIS

WILL NOW DISCUSS IMPACT ESTIMATION IN THE CASE OF A FOUR-GROUP DESIGN (PRETEST/POSTTEST/COMPARISON GROUP).

THE PRESENTATION USES THE APPROACH AND NOTATION OF WOOLDRIDGE OP. CIT.

A. DESIGN-BASED ESTIMATES: ESTIMATES FOR HIGHLY STRUCTURED EXPERIMENTAL DESIGNS WITH RANDOMIZED ASSIGNMENT TO TREATMENT

[GRAPH, SHOWING SAME STARTING POINT FOR TREATMENT AND CONTROL GROUPS.]

MAIN FEATURE: NOT NECESSARY TO DEFINE THE CAUSAL MODEL (NO ECONOMIC THEORY, NO ECONOMIC MODEL (OTHER THAN TREATMENT VARIABLE)). WITH RANDOMIZED ASSIGNMENT TO TREATMENT, DON’T NEED ONE (OTHER THAN TREATMENT) TO INFER CAUSAL IMPACT. ESTIMATES OF IMPACT WILL BE BASED ON DESIGN PARAMETERS (AND PERHAPS A FEW COVARIATES), NOT ON A STRUCTURAL-EQUATION MODEL.
CASE 1 (SIMPLEST CASE): PRETEST/POSTTEST RANDOMIZED ASSIGNMENT TO TREATMENT, WITH SIMPLE RANDOM SAMPLING IN EACH GROUP. NO IMPORTANT HIDDEN VARIABLES. (OTHER THAN TREATMENT, NO FACTORS (EVENTS) THAT MIGHT AFFECT THE TREATMENT AND CONTROL OUTCOMES DIFFERENTLY AND SUBSTANTIALLY.)

MODEL:

Let t denote survey round: t = 0 for baseline (before, time 1) and t = 1 for endline (after, time 2).

Let T denote treatment: T = 0 for control unit and T = 1 for treatment unit (same value in baseline and endline).

Important note: The variable T may or may not be considered to be a function of time, T_t. In the preceding model formulation, a treatment unit is distinguished as such from Round 0, and the value of the treatment indicator does not change in time (for the same unit). In some model formulations, a treatment unit is not distinguished as such until the second round. While either formulation works well for the present case (randomized assignment to treatment), the convention adopted here works better for cases (considered later) in which randomized assignment is lacking, and it is necessary to distinguish the treatment and control groups in Round 0 (baseline). In the convention adopted here, impact is the interaction effect of treatment and time. In the alternative representation, impact in the present case would be the treatment effect.

Response y_{Tt}

Allow for a mean, \( \beta_0 \), a “round” (or time) effect, \( \beta_1 \), a “treatment” effect, \( \beta_2 \) (which is zero in the case of randomized assignment), and an interaction effect of round and treatment, \( \beta_3 \):

Control before: \( y_{00} = \beta_0 + e \)

Control after: \( y_{01} = \beta_0 + \beta_1 + e \)

Treatment before: \( y_{10} = \beta_0 + \beta_2 + e \)

Treatment after: \( y_{11} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + e \)

or

\[ y_{Tt} = \beta_0 + \beta_1 t + \beta_2 T + \beta_3 tT + e \]

Impact is the interaction effect of time and treatment, i.e., the coefficient, \( \beta_3 \), on the interaction variable \( tT \).

May obtain this estimate as a double difference, or as the coefficient of \( T \) in a regression model.

DD measure = \( (Ey_{11} - Ey_{10}) - (Ey_{01} - Ey_{00}) = \beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 - \beta_2 - \beta_0 - \beta_1 + \beta_0 = \beta_3 \).

DD estimator = \( (\bar{x}_{11} - \bar{x}_{10}) - (\bar{x}_{01} - \bar{x}_{00}) \).
THIS DESIGN CAN BE SUBSTANTIALLY IMPROVED ON, BY INTERVIEWING THE SAME 
HOUSEHOLDS IN BOTH SURVEY ROUNDS.

CASE 2: PRETEST/POSTTEST RANDOMIZED ASSIGNMENT TO TREATMENT, MATCHING 
OVER TIME.

EACH HOUSEHOLD IS A “BLOCK.” THIS DESIGN ACHIEVES A SUBSTANTIAL INCREASE 
IN PRECISION OVER THE PREVIOUS DESIGN, BECAUSE OF THE INTRODUCTION OF 
WITHIN-HOUSEHOLD CORRELATION INTO THE DESIGN.

Let \( h_i \) denote an indicator variable (“dummy” variable) for the \( i \)-th household:

\[ h_i = 1 \] if the unit is in the \( i \)-th household, 0 otherwise, \( i = 1, n_{hh} \) (where \( n_{hh} \) denotes the 
number of households).

The model becomes

\[
y_{th1t2...n_{hh}h1...n_{hh}} = \beta_0 + \beta_1 t + \beta_2 T + \beta_3 T + \beta_4 h_1 + + \beta_{n_{hh}h2} h_{n_{hh}1} + e
\]

A DIFFICULTY WITH THIS MODEL IS THAT IT INCLUDES A VERY LARGE NUMBER OF 
PARAMETERS (I.E., ONE FOR EACH HOUSEHOLD) THAT ARE OF NO INTEREST. THESE 
ARE CALLED “NUISANCE” PARAMETERS. IN A “STRAIGHTFORWARD” ANALYSIS, 
INCLUDING THESE PARAMETERS WOULD REQUIRE THE DEFINITION OF A LARGE 
NUMBER OF EXPLANATORY VARIABLES AND THE INVERSION OF A LARGE 
CROSSPRODUCTS MATRIX. STATISTICAL PROGRAM PACKAGES (E.G., STATA) CAN 
HANDLE THIS PROBLEM EFFECTIVELY (I.E., WITHOUT DEFINING THE NUISIBLE 
VARIABLES AND WITHOUT INVERTING AN UNNECESSARILY LARGE MATRIX).

AS BEFORE, THE IMPACT IS THE COEFFICIENT \( \beta_3 \) ON THE TIME x TREATMENT 
VARIABLE (tT).

THE PRECISION OF THIS DESIGN CAN BE INCREASED BY MATCHING THE TREATMENT 
AND CONTROL GROUPS IN SOME WAY. THE MOST FREQUENT WAY OF DOING THIS IS 
TO MATCH AT A LOW ADMINISTRATIVE LEVEL, SUCH AS VILLAGE, CENSUS 
ENUMERATION AREA OR DISTRICT.

CASE 3: PRETEST/POSTTEST “CLUSTER” RANDOMIZED ASSIGNMENT TO TREATMENT, 
WITH MATCHING OVER TIME. (I.E., A “DOUBLE-MATCHED” DESIGN, WITH TEMPORAL 
AND CROSS-SECTIONAL MATCHING.)

Let \( m_i \) denote an indicator variable (“dummy” variable) for the \( i \)-th match set (e.g., pair 
of matched villages):

\[ m_i = 1 \] if the unit is in the \( i \)-th match set, 0 otherwise, \( i = 1, n_{ms} \) (where \( n_{ms} \) denotes the 
number of match sets).

The model becomes

\[
y_{th1t2...n_{hh}1m1ms2...n_{ms}1} = \beta_0 + \beta_1 t + \beta_2 T + \beta_3 T + \beta_4 h_1 + ... + \beta_{n_{hh}h2} h_{n_{hh}1} + \beta_{n_{hh}h3} m_{1} + ... + \beta_{n_{hh}n_{ms}1} m_{n_{ms}1} + e
\]

75
AS BEFORE, STATISTICAL SOFTWARE PACKAGES CAN HANDLE THE SUBSTANTIAL NUMBER OF DUMMY VARIABLES ASSOCIATED WITH THE TREATMENT/CONTROL MATCH SETS.

AS BEFORE, THE IMPACT IS THE COEFFICIENT $\beta_3$ ON THE TIME $\times$ TREATMENT VARIABLE ($tT$).

NOTE ON MODEL SIMPLIFICATION:

THE HOUSEHOLD NUISANCE PARAMETERS MAY BE ELIMINATED BY TRANSFORMING THE DATA BY TAKING DIFFERENCES WITHIN HOUSEHOLDS, OR BY DE-MEANING THE DATA BY SUBTRACTING THE HOUSEHOLD MEAN FROM EACH OBSERVATION. IN THE CASE OF TWO SURVEY ROUNDS, THESE TWO APPROACHES PRODUCE IDENTICAL ESTIMATES. (IN TIME SERIES APPLICATIONS (WITH MORE THAN TWO TIME PERIODS), THEY PRODUCE DIFFERENT RESULTS.)

THE DIFFERENCING APPROACH REDUCES THE SIZE OF THE SAMPLE TO THE NUMBER OF HOUSEHOLDS HAVING DATA IN BOTH ROUNDS.

THIS TRANSFORMATION IS CALLED A “FIXED-EFFECTS” TRANSFORMATION (WILL BE ADDRESSED LATER).

EVEN THOUGH THE PRECEDING DESIGNS ARE RANDOMIZED, IT IS QUITE POSSIBLE THAT THERE ARE SOME OBSERVED VARIABLES THAT AFFECT OUTCOME THAT MAY HAVE CHANGED OVER TIME. THESE VARIABLES (COVARIATES) ARE INCLUDE IN THE MODEL.

NOTE: EVEN THOUGH THERE MAY HAVE BEEN RELATIVELY FEW MATCH VARIABLES AVAILABLE FOR DESIGN, AFTER THE SURVEY HAS BEEN CONDUCTED THERE ARE A LARGE NUMBER OF COVARIATES AVAILABLE FROM THE QUESTIONNAIRE.

CASE 4: PRETEST/POSTTEST “CLUSTER” RANDOMIZED ASSIGNMENT TO TREATMENT, WITH TEMPORAL AND CROSS-SECTIONAL MATCHING, ADJUSTING FOR COVARIATES.

THIS IS STILL CONSIDERED A “DESIGN-BASED” ESTIMATE, EVEN THOUGH COVARIATES ARE INVOLVED. THE COVARIATES ARE SIMPLY OBSERVED, NOT THE RESULT OF FORCED CHANGES. THEY ARE ASSUMED TO BE EXOGENOUS, FIXED EFFECTS. THE MODEL IS AN “ANALYSIS OF COVARIANCE” MODEL. SIMILAR TO THE USE OF RATIO AND REGRESSION ESTIMATES IN DESCRIPTIVE SAMPLE SURVEY. (INCREASE PRECISION, BUT LITTLE MEANING READ INTO THE COVARIATE COEFFICIENTS – SIMILAR TO NUISANCE PARAMETERS.)

ROLE OF COVARIATES: BETTER MODEL REPRESENTATION (INCREASE PRECISION, LOWER BIAS)

REPRESENT EVENTS THAT HAPPENED OVER THE COURSE OF THE STUDY (E.G., INTRODUCTION OF HEALTH CLINICS IN SOME DISTRICTS).

REPRESENT DIFFERENCES IN TREATMENT AND CONTROL UNITS, EVEN THOUGH RANDOMIZED ASSIGNMENT.
NOTE: FOR THE PRESENT EXAMPLE, FOR SIMPLICITY, THE COVARIATES WILL BE ASSUMED TO CONSTANT IN TIME. THIS ASSUMPTION WILL BE DROPPED LATER.

Let \((x_1,\ldots,x_k)\) represent \(k\) covariates.

The model becomes

\[
y_{thth2...nnhth-1ms2...nms-1} = \beta_0 + \beta_1 t + \beta_2 T + \beta_3 tT + \beta_4 h_1 + \ldots + \beta_{n_{hth}+2} h_{n_{hth}-1} + \beta_{n_{hth}+3} ms_1 + \ldots + \beta_{n_{hth}+n_{ms}+1} ms_{n_{ms}-1} + \beta_{n_{hth}+n_{ms}+2} x_1 + \ldots + \beta_{n_{hth}+n_{ms}+k} x_k + e.
\]

As discussed before, it is difficult to read much into covariate coefficients unless forced (random) changes have been made in the variables, and the variables are orthogonal (as in a highly structured designed experiment, such as a fractional factorial design).

For clarity in discussion, it is most useful is to represent only the forced-change variables explicitly in the model, and represent all others as a “generic” \(x'\beta\) term. Using this approach, the preceding model is shown, for example, as:

\[
y_{1t} = x'\beta + at + yT + \delta tT + e
\]

(where, because of randomization, \(y=0\)). In most cases, the individual-variable terms in \(x'\beta\) have little meaning (because of lack of forced changes and lack of orthogonality (confounding; Simpson’s paradox).

The preceding discussion has focused on the general linear model, which is appropriate, for example, for continuous response variables. For binary response variables, would use a generalized linear model, such as a logistic regression model.

The main point to the preceding is that, even if the design is based on randomized assignment to treatment, the model used to analyze the data will be a “regression-adjusted” (covariate adjusted) model.

It should not be anticipated that the use of a highly structured randomized-assignment-to-treatment design will enable the use of the simple double-difference estimator to estimate impact.

The measure of impact is the double-difference measure \((\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00})\), but the estimator of impact is rarely the sample double-difference estimator. The impact estimate is usually a coefficient in a regression model. This estimate is called a “regression-adjusted” double-difference estimate. This descriptor is somewhat misleading, because it is the model that is adjusted, not the measure, and the estimator (the regression coefficient) is not represented as an “adjustment” to the raw double-difference estimator.

This is why, when a design gets “messed up” (corrupted) somewhat by nonresponse, it usually doesn’t affect the analysis much – complex regression estimators will be required in any event.
NOTE THAT THE PRECEDING MODEL FOCUSES ON OUTCOME, NOT ON SELECTION OR ATTRITION. SELECTION BIASES ARE REMOVED BY RANDOMIZED ASSIGNMENT TO TREATMENT. IF ATTRITION IS A PROBLEM, IT WILL BE ADDRESSED IN THE “MODEL-BASED” APPROACH, TO BE DISCUSSED.

B. MODEL-BASED ESTIMATES. STATISTICAL ANALYSIS USED TO ESTIMATE PARAMETERS IN A CAUSAL MODEL

WITH THE DESIGN-BASED APPROACH, RANDOMIZED ASSIGNMENT TO TREATMENT WAS THE BASIS FOR CAUSAL INFERENCE. NO NEED FOR A CAUSAL MODEL (OTHER THAN TREATMENT).

MODEL-BASED APPROACH. CAUSAL MODEL IS SPECIFIED VIA CAUSAL DIAGRAM OR STRUCTURAL EQUATIONS. STATISTICAL ANALYSIS IS USED TO ESTIMATE MODEL PARAMETERS. STRUCTURAL-EQUATION MODELING.

IN PHYSICAL SCIENCES, INTRODUCE FORCED RANDOM VARIATION (CHANGES) INTO EXPLANATORY VARIABLES, ACCORDING TO A HIGHLY STRUCTURED DESIGN (E.G., FRACTIONAL FACTORIAL, LATIN SQUARE, PARTIALLY BALANCED INCOMPLETE BLOCKS).

IN SOCIAL AND ECONOMIC SCIENCES, CAN USUALLY INTRODUCE FORCED CHANGES INTO JUST A FEW VARIABLES (E.G., TREATMENT, SURVEY ROUND), AND OFTEN WORK WITH A BINARY TREATMENT VARIABLE (INCLUDED OR NOT INCLUDED IN THE PROGRAM; RECEIPT OR NON-RECEIPT OF PROGRAM SERVICES).

BECAUSE OF LACK OF RANDOMIZED ASSIGNMENT TO TREATMENT, AND LACK OF FORCED CHANGES IN EXPLANATORY VARIABLES, PROPERTIES OF ESTIMATES ARE BASED ON ASSUMPTIONS, SUCH AS CONDITIONAL INDEPENDENCE.

THE BIASING EFFECTS OF LACK OF RANDOMIZATION ARE MITIGATED THROUGH THE USE OF PROCEDURES SUCH AS:
- EX ANTE AND EX POST MATCHING
- USE OF MODELS THAT REPRESENT SELECTION AND ATTRITION
- USE OF MODEL-BASED ESTIMATORS

THE MATHEMATICAL FORM OF THE MODEL USED TO REPRESENT OUTCOME WILL BE AS IN CASE 4 DISCUSSED ABOVE (GLM), BUT THE ASSUMPTIONS WILL BE DIFFERENT.

TAKING EX ANTE MATCHING INTO ACCOUNT

EX ANTE MATCHING: MATCHING AS PART OF THE DESIGN, PRIOR TO THE SURVEY.

SUMMARY OF ROLE OF MATCHING

ASSUME SAME HOUSEHOLDS INTERVIEWED IN BOTH SURVEY ROUNDS. DISCUSSION HERE PERTAINS TO MATCHING OF TREATMENT AND CONTROL SAMPLES (I.E., CROSS-SECTIONAL MATCHING, NOT TEMPORAL MATCHING).
IN AN EXPERIMENTAL DESIGN WITH RANDOMIZED ASSIGNMENT TO TREATMENT, MATCHING IS USED PRIOR TO RANDOMIZED ASSIGNMENT TO TREATMENT, TO INCREASE PRECISION, NOT BIAS (SINCE BIAS IS REMOVED BY RANDOMIZED ASSIGNMENT TO TREATMENT. (“MATCHED PAIRS” DESIGN; LOCAL CONTROL.)

USE A MATCHING TECHNIQUE TO FORM MATCHED PAIRS (OR MATCH SETS OF MORE THAN TWO, TO ALLOW FOR REPLACEMENTS). ONE MEMBER OF EACH MATCH SET IS ASSIGNED TO TREATMENT AND ONE TO CONTROL. MATCHING IS USUALLY DONE USING ADMINISTRATIVE UNITS SUCH AS VILLAGES OR DISTRICTS, NOT AT THE HOUSEHOLD LEVEL (BECAUSE OF LIMITATIONS ON DATA FOR SURVEY DESIGN).

IN A QUASI-EXPERIMENTAL DESIGN, THE PURPOSE OF EX ANTE MATCHING IS TO CAUSE THE DISTRIBUTIONS OF VARIABLES THAT MAY AFFECT OUTCOME TO BE SIMILAR FOR THE TREATMENT AND CONTROL SAMPLES. THE PRIMARY GOAL IS TO REDUCE BIAS, BUT PRECISION IS ALSO AFFECTED.

IN A QUASI-EXPERIMENTAL DESIGN, EX ANTE MATCHING IS USED TO INCREASE PRECISION AND TO REDUCE BIAS (ASSOCIATED WITH LACK OF RANDOMIZED ASSIGNMENT TO TREATMENT).

MATCHING MAY BE DONE AT TWO LEVELS: MATCHING OF UNITS AND DISTRIBUTIONAL MATCHING. UNIT MATCHING IS PREFERRED SINCE IT CAUSES THE JOINT DISTRIBUTION OF MATCH VARIABLES TO BE SIMILAR (FOR TREATMENT AND CONTROL GROUPS), NOT JUST THE MARGINAL DISTRIBUTIONS.

EX ANTE MATCHING WAS DONE ON (MATCH) VARIABLES AVAILABLE PRIOR TO THE SURVEY. IN THE ANALYSIS PHASE, THE EX ANTE MATCHING IS “HISTORY.” IT MUST BE TAKEN INTO ACCOUNT IN THE ANALYSIS, BUT IS A GIVEN.

IF DISTRIBUTIONAL MATCHING WAS DONE, THE DESIGN IS NOT MATCHED CROSS-SECTIONALLY. DATA ARE ANALYZED AS IF THEY WERE TWO INDEPENDENT SAMPLES.

IF MATCHING WAS DONE ON HIGHER-LEVEL UNITS (E.G., VILLAGES OR HOUSEHOLDS), THEN THE MATCH SETS MUST BE SPECIFIED IN THE MODEL (AS WAS DISCUSSED EARLIER). (NOTE: MANY RESEARCHERS FAIL TO ACCOUNT FOR THE EX ANTE MATCHING, IN THE DATA ANALYSIS.)

THE ROLE OF EX POST MATCHING

EX POST MATCHING: MATCHING IN THE DATA ANALYSIS PHASE

TWO KINDS OF EX POST MATCHING:
   DATA TRIMMING, CULLING OR PRUNING, TO REDUCE MODEL DEPENDENCE MATCHING ESTIMATORS

   C. USE OF MATCHING TO REDUCE MODEL DEPENDENCE

MODEL DEPENDENCE: [DEFINITION**]

[GRAPH ILLUSTRATING MODEL DEPENDENCE]
EX POST MATCHING TO REDUCE MODEL DEPENDENCE

DATA TRIMMING: AS LONG AS THE TRIMMING IS DONE ON EXPLANATORY VARIABLES OF A MODEL, NO BIASES ARE INTRODUCED IN ESTIMATES IN ESTIMATES OF MODEL PARAMETERS, IF THE MODEL IS CORRECTLY SPECIFIED AND ESTIMABLE.

EX POST MATCHING IS USUALLY RESTRICTED TO REMOVAL OF NON-COMMON SUPPORTS IN EXPLANATORY VARIABLES THAT ARE BELIEVED TO HAVE A SUBSTANTIAL EFFECT ON OUTCOMES OF INTEREST. NO EFFORT TO MODIFY “SHAPE” OF DISTRIBUTIONS – THAT WILL BE TAKEN INTO ACCOUNT IN A REGRESSION MODEL.

DATA TRIMMING “LOSES” DATA. MIGHT HAVE DELETERIOUS EFFECT ON PRECISION (OR MIGHT HELP).

NOTE THAT EX POST MATCHING ADDRESSES BOTH SELECTION AND ATTRITION.

NOTE THAT EX POST MATCHING INCLUDES MANY MORE MATCH VARIABLES THAN EX ANTE MATCHING (I.E., ALL OF THE SURVEY VARIABLES, NOT JUST THE DESIGN VARIABLES).

A COMBINATION OF MATCHING AND REGRESSION ESTIMATION IS “DOUBLY ROBUST”: IF EITHER THE MATCHING IS CORRECT (I.E., RESULTS IN THE SAME DISTRIBUTIONS FOR TREATMENTS AND CONTROLS) OR THE REGRESSION MODEL IS CORRECT (I.E., CORRECTLY SPECIFIED), BUT NOT NECESSARILY BOTH, THEN THE ESTIMATES WILL BE CORRECT.

THE PRECEDING DISCUSSION RELATES TO MATCHING ON INDIVIDUAL VARIABLES. MATCHING MAY ALSO BE DONE ON COMPOSITE SCORES, SUCH AS A PROPENSITY SCORE. THIS WILL BE DISCUSSED IN THE NEXT SECTION.

D. ESTIMATORS BASED ON IGNORABILITY (CONDITIONAL INDEPENDENCE)

THIS SECTION WILL PRESENT RESULTS FOR A TWO-GROUP DESIGN (TREATMENT/CONTROL). A LATER SECTION WILL DESCRIBE MODIFICATIONS FOR THE FOUR-GROUP DESIGN (“PANEL DATA”: TREATMENT BEFORE, TREATMENT AFTER, CONTROL BEFORE, CONTROL AFTER).

WE WISH TO ESTIMATE THE (UNCONDITIONAL) AVERAGE TREATMENT EFFECT, ATE:

\[ \text{ATE} = E(y_1 - y_0) \]

AND/OR THE AVERAGE TREATMENT EFFECT ON THE TREATED, ATT:

\[ \text{ATT} = E(y_1 - y_0 | w=1) \]

A NUMBER OF ESTIMATORS WILL BE PRESENTED, OF VARYING COMPLEXITY. CHOICE OF MODEL DEPENDS MAINLY ON ASSESSMENT OF “FACE VALIDITY” (I.E., THE EXTENT TO WHICH THE VARIABLES AND STRUCTURE RESEMBLE REALITY), AND THE EXTENT TO WHICH MODEL TECHNICAL ASSUMPTIONS ARE SATISFIED).
FACTORS TO CONSIDER:

SELECTION FOR TREATMENT (BY PROGRAM PERSONNEL AND BY ELIGIBLE CLIENT)
NONRESPONSE (TO SURVEY QUESTIONNAIRE)
ATTRITION (LEAVING THE PROGRAM BETWEEN SURVEY ROUNDS)
EXPLANATORY VARIABLES
STRUCTURAL RELATIONSHIP OF OUTCOMES OF INTEREST TO EXPLANATORY VARIABLES (CAUSAL MODEL STRUCTURE)

STATISTICAL FACTORS:

ENDOGENEITY (CORRELATION BETWEEN EXPLANATORY VARIABLES AND MODEL ERROR TERMS)
HETEROSCEDASTICITY
LINEARITY

NOTE: THE FOLLOWING TAKES INTO ACCOUNT SELECTION (FOR TREATMENT) FROM ALL SOURCES.

NOTE: WILL PRESENT RESULTS FIRST FOR A TWO-GROUP (TREATMENT/CONTROL) DESIGN, THEN SHOW MODIFICATIONS FOR FOUR-GROUP (PANEL DATA) DESIGN.

ASSUMPTIONS FOR ESTIMATING ATE:

ASSUMPTION ATE1 (IGNORABILITY): CONDITIONAL ON x, w AND THE POTENTIAL OUTCOMES (RESPONSE) \( y_0, y_1 \) ARE INDEPENDENT. ALSO CALLED: CONDITIONAL INDEPENDENCE OF RESPONSE AND TREATMENT; IGNORABILITY OF TREATMENT; UNCONFOUNDEDNESS OF TREATMENT.

NOTE: IF w IS A DETERMINISTIC FUNCTION OF x, THIS ASSUMPTION CALLED SELECTION ON OBSERVABLES.

THE ASSUMPTION OF IGNORABILITY IS NOT TESTABLE, SINCE WE OBSERVE ONLY (y, w, x), AND NOT BOTH POTENTIAL OUTCOMES \( y_0 \) AND \( y_1 \).

(WEAKER) ASSUMPTION ATE1' (IGNORABILITY IN MEAN): \( E(y_0|x,w) = E(y_0|x) \) AND \( E(y_1|x,w) = E(y_1|x) \).

ASSUMPTION ATE2: (OVERLAP): FOR ALL x IN THE SUPPORT OF THE COVARIATES, 0 < \( P(w=1|x) < 1 \).

NOTE: \( p(x) = P(w=1|x) \) IS CALLED THE PROPENSITY SCORE.

NOTE: ASSUMPTIONS 1 AND 2 COMBINED ARE CALLED STRONG IGNORABILITY.

ASSUMPTIONS FOR ESTIMATING ATT:

ASSUMPTION ATT1' (IGNORABILITY IN MEAN): \( E(y_0|x,w) = E(y_0|x) \).

ASSUMPTION ATT2 (OVERLAP): FOR ALL x IN THE SUPPORT OF THE COVARIATES, \( P(w=1|x) < 1 \). (NOTE THAT THIS IS WEAKER THAN ASSUMPTION ATE2.)

GUIDANCE FOR SELECTION OF COVARIATES (x):

81
DO NOT INCLUDE VARIABLES THAT MAY BE AFFECTED BY TREATMENT (THIS WILL INTRODUCE A BIAS)
GOOD CANDIDATES: VARIABLES MEASURED PRIOR TO TREATMENT ASSIGNMENT, INCLUDING PAST OUTCOMES ON y.
DO NOT SELECT CANDIDATES FOR INSTRUMENTAL VARIABLES FOR w (I.E., VARIABLES THAT ARE RELATED TO w BUT INDEPENDENT OF UNOBSERVABLES THAT AFFECT \( (y_0, y_1) \))
CONSTRUCT A DIRECTED ACYCLIC GRAPH THAT DISPLAYS THE CAUSAL RELATIONSHIPS AMONG THE MODEL VARIABLES, FOLLOWING THE THEORY PRESENTED IN PEARL OP. CIT. CONDITION ON COVARIATES THAT SATISFY THE "BACK DOOR" OR "FRONT DOOR" ESTIMABILITY CRITERIA.

DEFINE THE TWO POTENTIAL-OUTCOME CONDITIONAL MEANS:
\[
\mu_0(x) = E(y_0|x) \\
\mu_1(x) = E(y_1|x)
\]
NOTE THAT THESE FUNCTIONS ARE UNKNOWN.

THE AVERAGE TREATMENT EFFECT CONDITIONAL ON x IS
\[
ATE(x) = E(y_1 - y_0|x) = \mu_1(x) - \mu_0(x)
\]
AND THE AVERAGE TREATMENT EFFECT ON THE TREATED CONDITIONAL ON x IS
\[
ATT(x) = E(y_1 - y_0|x, w=1).
\]
IDENTIFICATION PROBLEM: ARE ATE AND ATT IDENTIFIED (ESTIMABLE)?
FIRST APPROACH TO ESTABLISHING IDENTIFICATION (REGRESSION-FUNCTION-BASED APPROACH)
DEFINE
\[
m_0(x) = E(y|x, w=0) \\
and \\
m_1(x) = E(y|x, w=1).
\]
UNDER ASSUMPTION ATE1',
\[
ATE(x) = m_1(x) - m_0(x)
\]
AND UNDER THE ADDITIONAL ASSUMPTION ATE2,
\[
ATE = E[m_1(x) - m_2(x)].
\]
UNDER ASSUMPTION ATT1',
ATT(x) = m_1(x) - m_0(x),

AND UNDER THE ADDITIONAL ASSUMPTION ATT2,

ATT = E(m_1(x) - m_0(x)|w=1).

THE PRECEDING RESULTS, BASED ON REGRESSION FUNCTIONS E(y|x,w=0) and E(y|x,w=1) ARE THE BASIS FOR REGRESSION-ADJUSTMENT ESTIMATORS:

\[
\overline{ATE}_{reg} = N^{-1} \sum_{i=1}^{N} [\hat{m}_1(x_i) - \hat{m}_0(x_i)]
\]

\[
\overline{ATT}_{reg} = (\sum_{i=1}^{N} w_i)^{-1} \sum_{i=1}^{N} [\hat{m}_1(x_i) - \hat{m}_0(x_i)] .
\]

SECOND APPROACH TO ESTABLISHING IDENTIFICATION (PROPENSITY-SCORE-BASED APPROACH; PROPENSITY-SCORE WEIGHTING)

IT CAN BE SHOWN THAT

\[
\overline{ATE}(x) = \mu_1(x) - \mu_0(x) = E\left\{ \frac{[w-p(x)]y}{p(x)[1-p(x)]} | x \right\}
\]

AND

\[
\overline{ATE} = E\left\{ \frac{[w-p(x)]y}{p(x)[1-p(x)]} \right\} .
\]

THESE RESULTS, BASED ON THE PROPENSITY SCORE (p(x)), ARE THE BASIS FOR PROPENSITY-SCORE-BASED ESTIMATORS:

\[
\overline{ATE}_{psw} = N^{-1} \sum_{i=1}^{N} \left[ \frac{w_i y_i}{\hat{p}(x_i)} - \frac{(1-w_i)y_i}{1-\hat{p}(x_i)} \right] = N^{-1} \sum_{i=1}^{N} \left[ \frac{[w_i-p(x_i)]y_i}{\hat{p}(x_i)[1-\hat{p}(x_i)]} \right]
\]

\[
\overline{ATT}_{psw} = N^{-1} \sum_{i=1}^{N} \left[ \frac{[w_i-p(x_i)]y_i}{1-\hat{p}(x_i)} \right]
\]

NOTE: THE PRECEDING ESTIMATORS ARE SIMILAR TO THE HORVITZ-THOMPSON ESTIMATORS OF DESCRIPTIVE SAMPLE SURVEY. THE VARIANCE OF THE ESTIMATORS IS LARGE WHEN VALUES OF THE PROPENSITY SCORE ARE VERY SMALL OR VERY LARGE. IN PRACTICE, DROP OBSERVATIONS HAVING PROPENSITY SCORES BELOW .1 OR ABOVE .9.

ADDITIONAL PROPENSITY-SCORE-BASED ESTIMATORS:

REGRESSION-ADJUSTED PROPENSITY-SCORE-BASED ESTIMATE:

Regress \( y_i \) on \( 1, w_i, \hat{p}(x_i) \), \( i = 1, \ldots, N \)
MODIFIED REGRESSION-ADJUSTED PROPENSITY-SCORE-BASED ESTIMATE:

Regress \( y_i \) on \( 1, w_i, \hat{p}(x_i), w_i[\hat{p}(x_i) - \hat{\rho}], i = 1, \ldots, N \)

where \( \hat{\rho} \) is a consistent estimate of \( \rho = E[p(x_i)] = P(w_i = 1) \).

If \( \hat{p}(x_i) \) is from a logistic regression model that includes an intercept, the two estimates \( \bar{w} \) and \( N^{-1} \sum_{i=1}^{N} \hat{p}(x_i) \) are identical.

E. ESTIMATION OF STANDARD ERRORS

THE REGRESSION-ADJUSTMENT ESTIMATORS ARE SAMPLE AVERAGES OF REGRESSION ESTIMATES. COMPLICATED: USE BOOTSTRAPPING TO ESTIMATE THE STANDARD ERRORS OF THE ESTIMATES.


F. TESTING STATISTICAL HYPOTHESES

THREE STANDARD APPROACHES:

IMPACT ESTIMATOR IS A REGRESSION COEFFICIENT. TEST OF HYPOTHESIS OF ZERO IMPACT IS TEST OF HYPOTHESIS THAT THE COEFFICIENT IS ZERO. SHOWN IN ANALYSIS OF VARIANCE TABLE OUTPUT BY REGRESSION PROGRAM.

IMPACT ESTIMATOR IS A LINEAR CONTRAST (E.G., SINGLE DIFFERENCE). VARIANCE OF \( \hat{\beta} \) IS \( V = (XX')^{-1}\sigma^2 \). VARIANCE OF LINEAR FUNCTION \( c\hat{\beta} \) IS \( c'Vc \) (CALCULATED BY COMPUTER PROGRAM).

PRINCIPLE OF CONDITIONAL ERROR: ESTIMATE RESIDUAL VARIANCES FOR A “FULL” MODEL AND A “REDUCED” MODEL (UNDER THE NULL HYPOTHESIS). COMPARE VARIANCES USING AN F DISTRIBUTION (ANALYSIS OF VARIANCE). IF \( f \) REFERS TO THE “FULL” MODEL AND \( r \) TO THE “REDUCED” MODEL, AND SSE REFERS TO “SUM OF SQUARES,” THEN

\[
F = \frac{(SSE_r - SSE_f)/(t - 1)}{SSE_f/(N - t)}
\]

(THE RATIO OF TWO CHI-SQUARED VARIATES DIVIDED BY THEIR RESPECTIVE DEGREES OF FREEDOM). (SEE KUEHL, OP. CIT, P. 57 FOR DISCUSSION OF THIS FORMULA,)
G. ESTIMATION BASED ON COMBINING REGRESSION AND PROPENSITY-SCORE MATCHING

ESTIMATOR ON PAGE 932 OF WOOLDRIDGE OP. CIT.

H. MATCHING ESTIMATORS

SIMILAR TO THE PROPENSITY-SCORE-BASED ESTIMATION APPROACH (SAME IGNORABILITY/OVERLAP ASSUMPTIONS), BUT USING MORE GENERAL MATCHING TECHNIQUES

MULTIDIMENSIONAL MATCHING ON THE FULL SET OF COVARIATES
MAHALANOBIS-DISTANCE MATCHING
IMPORTANCE-SCORE MATCHING

PRECISION CAN BE SUBSTANTIALLY IMPROVED IF THE RELATIONSHIP OF OUTCOMES TO MATCH VARIABLES IS STRONG (SIMILAR TO STRATIFICATION IN A DESCRIPTIVE SURVEY). MAY BE MUCH MORE PRECISE THAN PROPENSITY-SCORE MATCHING.

APPROACH:

AS BEFORE, ESTIMATE CONDITIONAL ATE:

\[ \text{ATE}(x) = E(y|w=1,x) - E(y|w=0,x), \]

AND THEN AVERAGE OVER THE DISTRIBUTION OF \( x \) (BY AVERAGING OVER THE SAMPLE).

LACK OF OVERLAP (BETWEEN TREATMENT AND CONTROL GROUPS) CAUSES PROBLEMS (JUST AS FOR REGRESSION-ADJUSTMENT AND PROPENSITY-SCORE-BASED ESTIMATORS).

IF MATCH ON THE PROPENSITY-SCORE, THIS IS CALLED PROPENSITY-SCORE MATCHING (VS. PROPENSITY-SCORE WEIGHTING CONSIDERED EARLIER).

FOR BOOTSTRAP ESTIMATION OF STANDARD ERRORS, PREFERABLE TO MATCH ON SEVERAL NEAREST-NEIGHBORS, RATHER THAN A SINGLE ONE. (IF MATCH ON NEAREST NEIGHBOR, RESAMPLING Produces THE SAME BOOTSTRAP SAMPLE EVERY TIME.)

I. INSTRUMENTAL VARIABLE ESTIMATORS

IF THE IGNORABILITY ASSUMPTIONS CANNOT BE JUSTIFIED, THEN MAY USE THE METHOD OF INSTRUMENTAL VARIABLES TO ESTIMATE IMPACT.

DO NOT ASSUME INDEPENDENCE OF \((y_0, y_1)\) AND \(w\) GIVEN \(x\).

INSTEAD, ASSUME THE AVAILABILITY OF INSTRUMENTS \(z\) THAT ARE INDEPENDENT OF \((y_0, x)\). (THAT IS, \(z\) IS CORRELATED WITH \(w\) BUT INDEPENDENT OF OUTCOME \((y_0, x)\). WEAKER ASSUMPTIONS SUFFICE.)
PROCEDURE:

ESTIMATE THE BINARY RESPONSE \( P(w=1|x,z) = G(x,z;\gamma) \) BY MAXIMUM LIKELIHOOD, WHERE \( \gamma \) ARE PARAMETERS (E.G., OF A LOGIT OR PROBIT MODEL).

FROM THIS MODEL, ESTIMATE THE PROPENSITY SCORES \( \hat{G}_i \). DENOTE \( \hat{G} = (\hat{G}_1, \ldots, \hat{G}_n) \).

ESTIMATE THE EQUATION

\[
y = \delta + tw + x'\beta + e
\]

BY INSTRUMENTAL VARIABLES USING INSTRUMENTS 1, \( \hat{G} \), AND \( x \).

SEE WOOLDRIDGE PP. 937-940 FOR DETAILS.

NOTE THAT ESTIMATION OF THE PARAMETERS \( \gamma \) IN THE FIRST STAGE MAY BE IGNORED IN ESTIMATING THE 2SLS MODEL. (2SLS WITH GENERATED INSTRUMENTS.)

NOTE ALSO THAT SINCE \( G \) IS BEING USED AS AN INSTRUMENT FOR \( w \), THE MODEL FOR \( G \) DOES NOT HAVE TO BE CORRECTLY SPECIFIED.

REMEMBER THAT 2SLS ESTIMATES ARE BIASED. TO INVOKE CONSISTENCY, NEED LARGE SAMPLES.

J. INTENTION-TO-TREAT (ITT) ESTIMATOR

SUPPOSE THAT THE DESIGN IS AN EXPERIMENTAL DESIGN WITH RANDOMIZED ASSIGNMENT TO TREATMENT.

THE INTENTION TO TREAT (ITT) ESTIMATE OF IMPACT IS OBTAINED BY APPLYING THE ESTIMATION FORMULAS FOR THE EXPERIMENTAL DESIGN TO THE OBSERVED OUTCOME, WHETHER OR NOT TREATMENT WAS ACTUALLY RECEIVED.

BY DEFINITION, THERE IS NO "SELECTION" BIAS IN THIS ESTIMATOR.

LET \( z \) DENOTE THE INTENTION TO TREAT.

THE RELATIONSHIP OF ATE TO ITT IS

\[
ATE = \frac{ITT}{P(w=1|z=1) - P(w=1|z=0)}.
\]

IF \( P(w=1|z=0) = 0 \), THEN

\[
ATE = \frac{ITT}{P(w=1|z=1)}.
\]

THIS FORMULA IS KNOWN AS THE "BLOOM RESULT" (AFTER HOWARD BLOOM, 1994).

K. LOCAL AVERAGE TREATMENT EFFECT (LATE) ESTIMATOR
IN THE INSTRUMENTAL-VARIABLE CASE CONSIDERED ABOVE, AN INSTRUMENTAL VARIABLE \( z \) WAS INTRODUCED FOR THE TREATMENT VARIABLE, \( w \).

POTENTIAL OUTCOMES \((y_0, y_1)\) WERE INTRODUCED FOR THE OUTCOME, \( y \):

\[
\begin{align*}
y &= y_0 \text{ if } w=0 \\
y &= y_1 \text{ if } w=1
\end{align*}
\]

SO

\[
y = y_0 + w(y_1 - y_0).
\]

MAY ALSO CONSIDER POTENTIAL TREATMENTS:

\[
\begin{align*}
w &= w_0 \text{ if } z=0 \\
w &= w_1 \text{ if } z=1
\end{align*}
\]

IN WHICH CASE

\[
w = (1-z)w_0 + zw_1 = w_0 + z(w_1 - w_0)
\]

AND (SUBSTITUTING IN \( y = y_0 + w(y_1 - y_0) \))

\[
y = y_0 + w_0(y_1 - y_0) + z(w_1 - w_0)(y_1 - y_0).
\]

IT IS ASSUMED THAT \( z \) IS INDEPENDENT OF \((y_0, y_1, w_0, w_1)\).

IF WE DEFINE:

Never-takers: \( w_0 = 0, w_1 = 0 \)
Compliers: \( w_0 = 0, w_1 = 1 \)
Defiers: \( w_0 = 1, w_1 = 0 \)
Always-takers: \( w_0 = 1, w_1 = 1 \)

UNDER CERTAIN ASSUMPTIONS (CALLED THE “LATE” ASSUMPTIONS), IT CAN BE SHOWN THAT THE EXPECTED TREATMENT EFFECT ON THE COMPLIERS IS GIVEN BY

\[
\text{LATE} = E(y_1 - y_0|w_1=1, w_0=0) = (\text{Effect of } z \text{ on } y) / (\text{Effect of } z \text{ on } w) = [E(y|z=1) - E(y|z=0)] / [E(w|z=1) - E(w|z=0)].
\]

ESTIMATE LATE BY USING THE SAMPLE ANALOGS OF THE EXPECTATIONS.

INTERESTING ASPECT OF LATE: CANNOT OBSERVE THE COMPLIERS.

THE LATE ASSUMPTIONS ARE

1. \( P(w=1|z) \) is a nonconstant function of \( z \)
2. Independence of \( z \) and \((y_0, y_1, w_0, w_1)\)
3. Monotonicity: For all i, either \( w_{1i} \geq w_{0i} \) or \( w_{1i} \leq w_{0i} \)

For an experimental design, there is no need to consider covariates. The instrumental variable, \( z \), is selection for treatment in the design.

If the design has been corrupted, the Late may be calculated as the ITT, or covariates may be taken into account:

The formula for the Late when covariates are included is:

\[
E(y_1 - y_0|\text{complier}) = \frac{[E(y|x, z=1) - E(y|x, z=0)]}{[E(w|x, z=1) - E(w|x, z=0)]}.
\]

L. Regression-discontinuity ((RD) Design)

Sharp regression-discontinuity (SRD) design: There is a single covariate \( x_i \), that determines treatment:

\( w_i = I(x_i \geq c) \) (where \( I(.) \) is the unit indicator function).

Cannot apply previous methods, such as propensity-score weighting, because there is no overlap. (Ignorability holds because \( w \) is a deterministic function of \( x \), so \( E(y_g|x, w) = E(y_g|x), g=0,1 \).)

Use of regression adjustment would require extrapolation beyond the range of the data.

Approach: estimate the average treatment effect at the discontinuity point:

\[
\text{ATE}(c) = \tau_c = E(y_1 - y_0|x=c) = \mu_1(c) - \mu_0(c).
\]

No external validity.

Assumption: the mean functions \( \mu_1(x) - \mu_0(x) \) are continuous at \( c \).

Procedure: estimate the two regression functions above and below \( c \), and evaluate at \( c \). The impact is the difference in these values.

Linear regression estimator:

Define

\[
\mu_{0c} = \mu_0(c)
\]

And

\[
\mu_{1c} = \mu_1(c)
\]

Define

\[
y_0 = \mu_{0c} + \beta_0(x - c) + u_0
\]
\[ y_1 = \mu_c + \beta_1(x - c) + u_1 \]

SO THAT

\[ y = (1 - w)y_0 + wy_1 = \mu_c + \tau_c w + \beta_0(x - c) + \delta w(x - c) + r \]

where

\[ r = u_0 + w(u_1 - u_0) \]

and \( \delta = \beta_1 - \beta_0. \)

THE ESTIMATE OF IMPACT IS THE ESTIMATE OF \( \tau_c \) (THE JUMP IN THE LINEAR FUNCTION AT \( x = c \)).

TO ESTIMATE \( \tau_c \), REGRESS \( y \) on \( 1, w_i, (x_i - c) \) and \( w_i(x_i-c) \).

RESTRICT DATA TO A SMALL BAND AROUND \( c \).

FUZZY REGRESSION-DISCONTINUITY (FRD) DESIGN: THE PROBABILITY OF TREATMENT CHANGES DISCONTINUOUSLY AT \( x = c \):

\[ P(w=1|x) = F(x). \]

See pp 957-958 of Wooldridge op. cit. for details.

M. RESPONSES DISCRETE OR IN A LIMITED RANGE

SIMILAR METHODS

E.G., IF RESPONSE IS POISSON, USE POISSON DISTRIBUTION IN MAXIMUM LIKELIHOOD

CONTINUOUS, BUT DATA CENSORED OR TRUNCATED

DATA CENSORING (E.G., OBSERVATION IS “TOP-CODED” OR RESPONSE MISSING)

TRUNCATED (DISTRIBUTION IS “CUT OFF,” OBSERVATIONS DELETED; SELECTION ON BASIS OF RESPONSE)

EXAMPLE: CENSORED NORMAL REGRESSION MODEL

CENSORED TOBIT MODEL:

Latent variable \( y^* = x^\prime \beta + u \), where \( u \approx N(0, \sigma^2) \)

Observe \( y = \max(0, y^*) \)

ESTIMATION:

OLS INCONSISTENT (SIMILAR TO HIDDEN VARIABLE); USE MAXIMUM LIKELIHOOD.

(COMPLICATED, DETAILS OMITTED.)
EXAMPLE: TRUNCATED NORMAL REGRESSION MODEL

TRUNCATED TOBIT MODEL:

EXAMPLE: WAGE OFFER AND LABOR FORCE PARTICIPATION

Wage model: \( y_1 = x_1^\prime \beta_1 + u_1 \)

Labor force participation model: \( y_2 = I(x_2 \delta_2 + v_2 > 0) \) (indicator function)

where \((x, y_2)\) is always observed, and \(y_1\) is observed only when \(y_2 = 1\)

Not reasonable to assume wage = 0 when person is not a labor force participant

OLS INCONSISTENT: SELECTION BIAS.

USE TOBIT SELECTION MODEL:

\[
\begin{align*}
    y_1 &= x_1^\prime \beta_1 + u_1 \\
    y_2 &= \max(0, x_2^\prime \delta_2 + v_2)
\end{align*}
\]

TWO-STEP MODEL: ESTIMATE \( \delta_2 \) FROM THE MODEL. E.G., FIRST-STEP PROBIT MODEL

\[
P(y_{i2}=1|x_i) = \Phi(x_i^\prime \delta_2)
\]

using all observations.

Estimate \( \beta_1 \) using OLS on the selected sample. (Details complicated, omitted.)

N. MULTIVALUED (CATEGORICAL) TREATMENT VARIABLE

CATEGORICAL, NON-BINARY: E.G., THREE VALUES: CONTROL (0), TREATMENT 1 (1), TREATMENT 2 (2)

SIMILAR ESTIMATION TECHNIQUES APPLY: E.G., ESTIMATE AVERAGE TREATMENT EFFECT FOR EACH TREATMENT LEVEL VS. CONTROL.

O. MULTIPLE TREATMENT VARIABLES

INSTEAD OF SCALAR TREATMENT VARIABLE, \( w \), HAVE VECTOR TREATMENT VARIABLE, \( w \).

MAY CONSIDER MORE ELABORATE RESEARCH DESIGN, E.G., TWO VARIABLES AT TWO LEVELS = FOUR COMBINATIONS OF TREATMENT VARIABLES. THIS IS AN “ORTHOGONAL” DESIGN.

MODELS SIMILAR TO BEFORE. TREATMENT VARIABLES ARE “SPECIAL” (HIGH INTEREST) EXPLANATORY VARIABLES. MORE EFFECTS OF INTEREST.
SIMILAR ESTIMATION TECHNIQUES APPLY, ALTHOUGH SOMEWHAT MORE COMPLICATED.

P. PANEL DATA

SECTION A DESCRIBED THE FOUR-GROUP MODEL (TREATMENT BEFORE, TREATMENT AFTER, CONTROL BEFORE, CONTROL AFTER) IN THE CASE OF RANDOMIZED ASSIGNMENT TO TREATMENT:

MODEL:

\[ y_{Tt} = x'\beta + \alpha t + \gamma T + \delta tT + e \]

AND SECTIONS D-N DESCRIBED ESTIMATION FOR THE TWO-GROUP MODEL (TREATMENT/CONTROL), IN THE POTENTIAL-OUTCOMES CONCEPTUAL FRAMEWORK (LACK OF RANDOMIZED ASSIGNMENT TO TREATMENT).

THAT APPROACH WAS TAKEN TO SIMPLIFY THE PRESENTATION.

THIS SECTION DESCRIBES EXTENSIONS OF THE THEORY TO THE GENERAL FOUR-GROUP CASE (NON-RANDOMIZED ASSIGNMENT TO TREATMENT).

[GRAPH, SHOWING DIFFERENT STARTING POINTS FOR TREATMENT AND CONTROL GROUPS.]

MAIN FEATURES:

WITH RANDOMIZED ASSIGNMENT TO TREATMENT, IT WAS NOT NECESSARY TO DEFINE THE CAUSAL MODEL (ECONOMIC MODEL). LACKING THIS, IT IS NECESSARY TO SPECIFY THE CAUSAL MODEL. CAUSAL EFFECT WAS COEFFICIENT OF TREATMENT VARIABLE (w). COVARIATES WERE “INCIDENTAL” (DESIGN EFFECTS, ASSOCIATION).


CASE 1 PREVIOUS MODEL (RANDOMIZED ASSIGNMENT TO TREATMENT, NO COVARIATES):

\[ y_{Tt} = x'\beta + \alpha t + \gamma T + \delta tT + e \]

WHERE, BECAUSE OF RANDOMIZED ASSIGNMENT TO TREATMENT, \( \gamma = 0 \).

WE NOW ALLOW FOR THE FACT THAT, BECAUSE OF LACK OF RANDOMIZED ASSIGNMENT TO TREATMENT, THE TREATMENT AND CONTROL GROUPS MAY DIFFER AT BASELINE.
THE MODEL FOR THIS EXAMPLE (CASE 1) DOES NOT INCLUDE COVARIATES. REALISTIC MODELS WILL INCLUDE COVARIATES, IN ACCORDANCE WITH THE CAUSAL MODEL.

IN THE PREVIOUS CONSIDERATION OF COVARIATES (“CASE 4”), THEY WERE ASSUMED TO BE EXOGENOUS, FIXED EFFECTS. BOTH OF THESE ASSUMPTIONS WILL NOW BE DROPPED (I.E., THEY MAY NOW BE ENDOGENOUS AND RANDOM).

ASSUMPTIONS ARE NOW REQUIRED CONCERNING THE RELATIONSHIP OF EXPLANATORY VARIABLES (COVARIATES) AND ERROR TERMS OVER TIME (I.E., BETWEEN ROUNDS). THESE ASSUMPTIONS ARE COMPLICATED, BUT ESSENTIAL TO TAKE INTO ACCOUNT.

THESE ASSUMPTIONS ARE REPRESENTED IN THE CAUSAL MODEL.

FIXED-EFFECTS OR RANDOM EFFECTS

IN ECONOMETRIC APPLICATIONS AND EVALUATION STUDIES, COVARIATES (I.E., VARIABLES EXCLUDING TREATMENT AND DESIGN VARIABLES) ARE USUALLY ASSUMED TO BE RANDOM VARIABLES (“RANDOM EFFECTS”).

IN THIS CASE, ASSUMPTIONS MUST BE MADE ABOUT THE RELATIONSHIP BETWEEN THE COVARIATES AND THE MODEL ERROR TERMS.

A KEY CONSIDERATION IS OMITTED VARIABLES, OR UNOBSERVED EFFECTS. MANY OF THE UNOBSERVED EFFECTS ARE REMOVED BY DIFFERENCING (E.G., A SINGLE-DIFFERENCE OR DOUBLE-DIFFERENCE IMPACT ESTIMATE).

THE UNOBSERVED EFFECTS MODEL (UEM) IS:

\[ y_t = x_t' \beta + c + u_t, \quad t = 0,1 \]

WHERE \( c \) DENOTES AN UNOBSERVED TIME-CONSTANT VARIABLE (UNOBSERVED COMPONENT, LATENT VARIABLE, UNOBSERVED HETEROGENEITY INDIVIDUAL EFFECT, IDIOSYNCRATIC DISTURBANCES).

RANDOM-EFFECTS MODEL:

IF \( c \) IS A RANDOM VARIABLE, THE FOLLOWING ORTHOGONALITY ASSUMPTIONS (AMONG OTHERS) ARE REQUIRED TO OBTAIN CONSISTENT ESTIMATORS OF THE PARAMETERS (\( \beta \)):

\[ E(u_t | X, c) = 0, \quad t = 0,1 \] (the error terms are uncorrelated with \( X \), given \( c \))

\[ E(c | X) = E(c) = 0 \] (\( c \) is uncorrelated with \( X \))

where \( X = (x_1, x_2) \) (and the subscript \( t \) on \( x_t \) refers to time (\( t \))).

THE RANDOM-EFFECTS APPROACH TO ESTIMATING \( \beta \) IS EQUIVALENT TO PLACING \( c \) IN THE MODEL ERROR TERM.
IF \( c \) IS A RANDOM VARIABLE, THE COMPOSITE ERROR \( c + u_t \) IS SERIALLY CORRELATED. IN ORDER TO OBTAIN CONSISTENT (GENERALIZED) LEAST-SQUARES ESTIMATES, NEED TO ASSUME ORTHOGONALITY OF \( X \) AND \( c \).

THE PRECEDING ASSUMPTIONS MEAN THAT THE UNOBSERVED EFFECT IS UNCORRELATED WITH THE OBSERVED COVARIATES AND THE MODEL ERROR TERMS (I.E., \( c \) AND \( X \) ARE ORTHOGONAL).

THE VARIANCES OF THE ESTIMATES DEPEND ON THE VARIANCE OF \( c \) (SO THAT TESTING HYPOTHESES IS LIKE A RANDOM-EFFECTS ANOVA).

WHETHER THE ORTHOGONALITY ASSUMPTIONS ARE JUSTIFIED IS INFERRED FROM THE CAUSAL MODEL.

(THE OTHER ASSUMPTIONS REQUIRED FOR A FIXED-EFFECTS ANALYSIS ARE THAT A CERTAIN QUADRATIC FORM IS POSITIVE DEFINITE (OF FULL RANK), AND THAT, CONDITIONAL ON \( c \) AND THE COVARIATES, THE MODEL ERROR TERMS ARE UNCORRELATED AND HAVE CONSTANT VARIANCE, AND THE VARIANCE OF \( c \) CONDITIONAL ON THE COVARIATES IS A CONSTANT.)

**FIXED-EFFECTS MODEL:**

IF \( c \) IS A FIXED EFFECT, IT MAY BE ARBITRARILY RELATED TO THE COVARIATES \( X \).

THE KEY ASSUMPTION FOR THE FIXED-EFFECTS MODEL IS:

\[
E(u_t|X,c) = 0, \ t=0,1 \text{ (the error terms are uncorrelated with } X, \text{ given } c).\]

THE SECOND ASSUMPTION OF THE RANDOM-EFFECTS MODEL, VIZ.,

\[
E(c|X) = E(c) = 0 \text{ (} c \text{ is uncorrelated with } X)\]

IS NOT REQUIRED. IN OTHER WORDS, \( E(c|X) \) IS ALLOWED TO BE ANY FUNCTION OF \( X \).

THE FIXED-EFFECTS MODEL IS HENCE MORE “ROBUST” THAN THE RANDOM EFFECTS MODEL, WITH RESPECT TO TECHNICAL ASSUMPTIONS.

IN THE FIXED-EFFECTS MODEL, \( c \) IS CONFOUNDED WITH \( X; \beta \). HENCE WE HAVE TO ELIMINATE IT FROM THE MODEL. THIS IS DONE EITHER BY DE-MEANING THE OBSERVATIONS FOR MATCHED UNITS (HOUSEHOLDS) OVER TIME, OR BY DIFFERENCING MATCHED UNITS OVER TIME. (THIS REMOVES ALL UNOBSERVED VARIABLES THAT ARE CONSTANT WITHIN HOUSEHOLDS.) THESE TRANSFORMATIONS ARE CALLED FIXED-EFFECTS TRANSFORMATIONS OR “WITHIN” TRANSFORMATIONS.

IN THE FIXED-EFFECTS MODEL, THE LEAST-SQUARES ESTIMATES OF \( \beta \) ARE UNBIASED CONDITIONAL ON \( X \).

UNDER THE ASSUMPTION THAT THE MODEL ERRORS ARE UNCORRELATED CONDITIONAL ON \( X \) AND \( c \), THE LEAST-SQUARES ESTIMATES ARE EFFICIENT.

WHICH APPROACH TO USE?
FOR BOTH APPROACHES, DECIDING WHETHER THE TECHNICAL ASSUMPTIONS HOLD IS DIFFICULT TO JUSTIFY – MODELS ARE TOO COMPLEX.


FIXED-EFFECTS MODEL IS MORE ROBUST. INFERENTIAL SCOPE IS RESTRICTED TO THE FIXED VALUES OF THE EXPLANATORY VARIABLES. THIS CAN MAKE A BIG DIFFERENCE IF HAVE A RELATIVELY SMALL SAMPLE OF PSUs (E.G., DISTRICTS, VILLAGES). HAVE A MUCH BETTER CHANCE OF DETECTING EFFECTS.

SOLUTION:

TAKE FULL ADVANTAGE OF THE FACT THAT MANY EVALUATION STUDIES INVOLVE INTERVIEW OF THE SAME HOUSEHOLDS IN BOTH SURVEY ROUNDS.

THIS IS DONE VIA THE FIXED-EFFECTS TRANSFORMATION (DE-MEANING OR Differencing).

POTENTIAL PROBLEM: THIS MAKES USE OF DATA FOR WHICH INTERVIEWS ARE AVAILABLE IN BOTH SURVEY ROUNDS. MAKE EXCEPTIONAL EFFORTS TO KEEP SECOND-ROUND NONRESPONSE LOW.

FIXED-EFFECTS MODEL IS MUCH USED IN POLICY ANALYSIS AND EVALUATION. THE POLICY VARIABLE (TREATMENT VARIABLE, w) MAY BE CORRELATED WITH AVERAGE VALUE OF THE ERROR TERM FOR MATCHED UNITS BETWEEN ROUNDS. ALL THAT IS REQUIRED IS THAT THE DEMEANED ERROR TERMS BE UNCORRELATED WITH THE COVARIATES (SEE WOOLDRIDGE OP. CIT. pp. 278-279).

FIXED-EFFECTS APPROACH, BASED ON “WITHIN” DATA, IS USELESS FOR ASSESSING THE EFFECT OF A TIME-CONSTANT VARIABLE IN A PANEL-DATA STUDY. THE VARIABLE “DROPS OUT.” SOLUTION: USE RANDOM EFFECTS MODEL (“BETWEEN” MODEL), OR INCLUDE INTERACTIONS WITH SURVEY ROUND (WHICH DOES NOT DROP OUT). (E.G., SIZE OF FARM, SEX OF HEAD OF HOUSEHOLD, DISTANCE TO CLINIC, AND VILLAGE CHARACTERISTICS MAY DROP OUT, BUT THE INTERACTION OF THESE VARIABLES WITH SURVEY ROUND DOES NOT.)

RECOMMENDED APPROACH: CONSIDER AND COMPARE BOTH RANDOM-EFFECTS AND FIXED-EFFECTS MODELS. IN GENERAL, RELY ON FIXED-EFFECTS MODEL AND FIXED-EFFECTS TRANSFORMATION (SIMPLER ASSUMPTIONS, MORE ROBUST).

KEEP IN MIND: FIXED-EFFECTS MODEL CORRESPONDS TO A MORE RESTRICTED SCOPE OF INFERENCE. (E.G., IN RANDOM-EFFECTS MODEL, SAMPLE PSUs ARE VIEWED AS A RANDOM SAMPLE; IN FIXED-EFFECTS MODEL, THEY BECOME THE POPULATION OF INTEREST. RANDOM EFFECTS MORE DIFFICULT TO FIND (ALLOWS FOR VARIATION IN PSU MEANS). RANDOM-EFFECTS ANOVA: COMPONENTS OF VARIANCE, VARIANCE OF “REDUCED” MODEL INCLUDES RESIDUAL VARIANCE AND PSU VARIANCE.

VERY IMPORTANT NOTE: WHEN INCLUDING INTERACTION TERMS OF COVARIATES WITH TREATMENT, MUST DE-MEAN THE COVARIATE TERMS, OR WILL NOT OBTAIN AN UNBIASED ESTIMATE OF IMPACT.

NOTE ON MODEL SPECIFICATION:
MODEL SHOULD NOT DEPEND ON PSU (CLUSTER, FIRST-STAGE SAMPLE UNIT). IF IT DOES, LOOK FOR AN IMPROVED SPECIFICATION.

Q. TREATMENT OF MISSING VALUES

REASONS FOR MISSING VALUES:
- NONRESPONSE IN BASELINE (REFUSALS, NOT AT HOME, CANNOT LOCATE (DEFECTIVE FRAME))
- NONRESPONSE IN ENDLINE (SAME, PLUS DIED, MOVED)
- ATTRITION (WITHDRAWAL FROM PROGRAM)

NOTE: REPLACEMENT HOUSEHOLDS ARE MORE PROBLEMATIC IN PANEL SURVEYS (THE REPLACEMENT HOUSEHOLD MAY DIFFER ON MANY VARIABLES, INTRODUCING SUBSTANTIAL ERROR INTO THE INTER-ROUND HOUSEHOLD DIFFERENCE.)

IT IS ESSENTIAL TO KEEP TRACK OF ALL BASELINE UNITS (HOUSEHOLDS), TO IMPLEMENT SELECTION MODEL.

REFERENCES ON MISSING VALUES:


MISSING VALUES ARE A VERY SERIOUS PROBLEM IN MODEL-BASED ESTIMATION, BECAUSE OF CASEWISE DELETION (IF AN OBSERVATION DOES NOT HAVE VALUES FOR ALL VARIABLES, IT IS DROPPED FROM THE DATA SET).

IMPERATIVE TO IMPUTE MISSING VALUES.

APPROACHES:
- SUBSTITUTION OF MEANS (OR MEDIAN) FOR MISSING VALUES
- SUBSTITUTION OF REGRESSION ESTIMATE (BASED ON NON-MISSING VALUES) FOR MISSING VALUES (SUBSTITUTION OF “CONDITIONAL MEANS,” BUCK’S METHOD)
- LIKELIHOOD-BASED (MODEL-BASED) ESTIMATION OF MISSING VALUES: MAXIMUM-LIKELIHOOD AND BAYES MODELS.
COMPLETE DATA MATRIX $Y = (y_{ij})$  
MISSING-DATA INDICATOR MATRIX, $M = (M_{ij})$.

MISSING COMPLETELY AT RANDOM (MCAR): MISSINGNESS (DISTRIBUTION OF M) DOES NOT DEPEND ON THE VALUES OF THE DATA, MISSING OR OBSERVED.
MISSING AT RANDOM (MAR): MISSINGNESS DEPENDS ONLY ON THE NON-MISSING VALUES.
NOT MISSING AT RANDOM (NMAR): DISTRIBUTION OF M DEPENDS ON THE MISSING VALUES IN THE DATA MATRIX, $Y$.

EXPECTATION-MAXIMIZATION (EM) ALGORITHM FOR FINDING MAXIMUM-LIKELIHOOD ESTIMATES FROM INCOMPLETE DATA (AVOIDS ANALYTICAL PROBLEMS):
REPLACE MISSING VALUES BY ESTIMATED VALUES
ESTIMATE PARAMETERS
RE-ESTIMATE MISSING VALUES ASSUMING THAT THE NEW PARAMETER ESTIMATES ARE CORRECT
ITERATE PREVIOUS STEP (RE-ESTIMATION OF PARAMETERS) UNTIL CONVERGENCE

NOTE: BOOTSTRAP ESTIMATION PROCESS REQUIRES FULL DATA SET: MUST IMPUTE MISSING VALUES PRIOR TO APPLYING.

R. USE OF WEIGHTS

FOUR TYPES OF WEIGHTS (STATA):

FREQUENCY WEIGHTS ($f$weights) INDICATE DUPLICATED OBSERVATIONS
SAMPLING WEIGHTS ($p$weights) DENOTE RECIPROCALS (INVERSES) OF THE PROBABILITY OF INCLUSION BECAUSE OF THE SAMPLE DESIGN. E.G., “EXPANSION FACTORS” USED TO ESTIMATE POPULATION TOTALS.
ANALYTIC WEIGHTS ($a$weights) ARE INVERSELY PROPORTIONAL TO THE VARIANCE OF AN OBSERVATION. (E.G., OBSERVATIONS ARE MEANS, AND THE WEIGHTS ARE THE NUMBER OF UNITS ON WHICH THE MEAN IS BASED.)
IMPORTANCE WEIGHTS ($i$weights) INDICATE THE “IMPORTANCE” OF AN OBSERVATION (NO FORMAL DEFINITION).

IN EVALUATION, ATTENTION FOCUSES ON $p$weights.

IF A REGRESSION MODEL IS CORRECTLY SPECIFIED (IDENTIFIED), DO NOT NEED THE WEIGHTS. (SO THE LARGE VARIATION IN WEIGHTS THAT IS OFTEN A FEATURE OF ANALYTICAL DESIGNS IS NOT A PROBLEM.)

ESTIMATE MODELS WITH AND WITHOUT WEIGHTS: RESULTS SHOULD BE SIMILAR.

NOTE: THE SAMPLE DESIGN WILL SPECIFY THE PROBABILITY OF SELECTION FOR EACH PSU, BUT THE SELECTION OF WITHIN-PSU UNITS IS USUALLY DONE IN THE FIELD (NO SAMPLE FRAME FOR HOUSEHOLDS). IT IS IMPORTANT TO DO THE WITHIN-PSU SAMPLING IN A WAY SUCH THAT THE PROBABILITY OF SELECTION IS KNOWN FOR EACH SAMPLE UNIT.
FOR EXAMPLE, SELECT TWO OR THREE RANDOM STARTING POINTS, AND SELECT “EVERY K-TH” HOUSEHOLD IN RANDOM DIRECTIONS. KEEP TRACK OF NUMBER OF HOUSEHOLDS AT EACH POINT. (COMPACT SEGMENTS.)

NOTE: IF USE SYSTEMATIC SAMPLING WITH A SINGLE RANDOM START (WITHIN PSU), CANNOT ESTIMATE THE WITHIN-PSU VARIANCE. THIS IS NOT A PROBLEM IN DESCRIPTIVE SURVEYS IF THE FIRST-STAGE FPC IS SMALL. IT IS NOT A PROBLEM UNDER THE “RANDOM-EFFECTS” ASSUMPTION (INFINITE POPULATION). IT IS A POTENTIAL PROBLEM UNDER THE “FIXED-EFFECTS” ASSUMPTION, IF THE NUMBER OF PSUs IS NOT VERY LARGE. BEST TO USE SYSTEMATIC SAMPLING WITH AT LEAST TWO RANDOM STARTS.

S. ESTIMATION OF STANDARD ERRORS

IT IS ESSENTIAL TO BE ABLE TO ESTIMATE THE STANDARD ERRORS OF ESTIMATES, TO ESTIMATE CONFIDENCE INTERVALS AND TO CONDUCT TESTS OF HYPOTHESES.

USE IMPACT ESTIMATORS FOR WHICH STANDARD ERRORS CAN BE ESTIMATED (E.G., NEAREST-NEIGHBOR MATCHING ESTIMATOR PRESENTS A PROBLEM).

FOR SIMPLE MODELS, CLOSED-FORM FORMULAS ARE AVAILABLE. FOR COMPLEX MODELS, USE RESAMPLING TECHNIQUES (E.G., BOOTSTRAPPING, JACKKNIFE).

USE TRANSFORMATIONS AND MODEL SPECIFICATION ARTFULLY, SO THAT STANDARD ERRORS ARE SMALL AND MODEL ERROR TERMS ARE HOMOSCEDASTIC.

FOR TWO-STEP M-ESTIMATORS (E.G., FIRST-STEP LOGISTIC REGRESSION SELECTION MODEL + SECOND-STEP OUTCOME REGRESSION MODEL), MAY IGNORE THE FACT THAT THE FIRST-STEP PARAMETER ESTIMATE IS AN ESTIMATE, AND OBTAIN CONSERVATIVE ESTIMATES OF STANDARD ERRORS.

IN STATA, THERE ARE SEVERAL OPTIONS AVAILABLE FOR ESTIMATION OF STANDARD ERRORS:

- HETEROSCEDASTICITY-ROBUST ESTIMATES (EICKER-HUBER-WHITE “SANDWICH” ESTIMATOR)
- “CLUSTER” OPTION (RECOGNIZES CLUSTERING – SHOULD NOT BE NECESSARY IF MODEL IS WELL SPECIFIED)
- BOOTSTRAPPING (AS AN OPTION IN MANY PROCEDURES, BUT WILL HAVE TO WRITE A .ado FILE FOR COMPLEX ESTIMATORS (E.G., TWO-STEP SELECTION MODEL).

IF ASSUME A RANDOM-EFFECTS MODEL, AND UNEXPECTEDLY LITTLE IS SIGNIFICANT, THEN USE A FIXED-EFFECTS MODEL (BUT RECOGNIZE THAT THE SCOPE OF INFERENCE IS RESTRICTED).

INSTRUMENTAL-VARIABLE MODELS WORK WELL ONLY IF GOOD INSTRUMENTS ARE AVAILABLE (UNCORRELATED WITH MODEL ERROR TERM BUT HIGHLY CORRELATED WITH THE ENDOGENOUS VARIABLE, CONDITIONAL ON COVARIATES).
7.4. Analysis of Baseline Data

ANALYSIS OF BASELINE DATA

DESCRIPTIVE SURVEYS

MAIN ASPECTS:
RESPONSE RATES
DESCRIPTIVE STATISTICS
ERROR PROFILE

EXCERPT FROM COURSE, SAMPLE SURVEY DESIGN AND ANALYSIS (DAY 2)

PART III. HOW TO ANALYZE SURVEY DATA

I. STANDARD ESTIMATION PROCEDURES FOR DESCRIPTIVE SURVEYS

A. SUMMARY OF PROCEDURES

1. PRELIMINARY ANALYSIS (LARGE DATA SETS)


UNIVARIATE ANALYSIS:

- UNIT (QUESTIONAIRE) NONRESPONSE. CALCULATE PROPORTION OF SAMPLE UNITS RESPONDING. CATEGORIZE BY REASON. NOTE VARIABLES CORRELATED WITH NONRESPONSE.
- ITEM NONRESPONSE. DEFINE A NO-RESPONSE INDICATOR VARIABLE FOR EACH VARIABLE. NOTE VARIABLES HAVING HIGH LEVEL OF NONRESPONSE, AND VARIABLES CORRELATED WITH NONRESPONSE.
- FOR INTERVAL-SCALE VARIABLES, COMPUTE MEAN OR PERCENTAGE, MEDIAN, MIN, MAX, RANGE, VARIANCE, STANDARD DEVIATION, HISTOGRAM.
- FOR CATEGORICAL VARIABLES COMPUTE MIN, MAX, RANGE AND HISTOGRAM.
- FOR GEOGRAPHIC DATA, PLOT ALL SAMPLE UNITS ON A MAP.

BIVARIATE ANALYSIS:

- RECODE ALL INTERVAL-SCALE VARIABLES BY QUINTILE
- RECODE ALL CATEGORICAL VARIABLES HAVING MORE THAN 10 CATEGORIES INTO TEN OR LESS CATEGORIES (PREFERABLY 5 OR LESS). DEFINE INDICATOR VARIABLES FOR HIGH-INTEREST CATEGORIES.
- CALCULATE CRAMER COEFFICIENT OF ASSOCIATION FOR ALL VARIABLES (I.E., EACH VARIABLE WITH EACH OTHER). NOTE STRONG CORRELATIONS.
• CONSTRUCT TABLES OR GRAPHS (SCATTER PLOTS, LOESS / LOWESS (LOCALLY WEIGHTED SCATTERPLOT SMOOTHING) CURVES) TO ILLUSTRATE STRONG RELATIONSHIPS.

2. PLANNED ANALYSIS

• FOR TOTAL POPULATION AND SUBPOPULATIONS OF INTEREST, SUMMARIZE NONRESPONSE PATTERNS
• ADJUST SAMPLE WEIGHTS FOR NONRESPONSE
• CALCULATE ESTIMATES INTEREST (USE APPROPRIATE FORMULAS OR PROCEDURES)
• CROSSTABS OF INTEREST (BY STRATUM OR WEIGHTED)
• COMPUTE EXPANSION ESTIMATES
• CONDUCT TESTS OF SIGNIFICANCE

3. SPECIAL ANALYSES

• ALTERNATIVE ESTIMATORS (RATIO, REGRESSION)
• POSTSTRATIFICATION
• ADDITIONAL SUBPOPULATIONS
• REDEFINE NONRESPONSE STRATA, RECOMPUTE ESTIMATES
• RESAMPLING ESTIMATES OF VARIANCES
• TABLES OF GENERALIZED VARIANCES
• $\chi^2$ (CHI-SQUARED) TESTS FOR CLUSTER-SAMPLE CROSSTABS

ANALYTICAL SURVEYS

MAIN ASPECTS:
  RESPONSE RATES
  MEASURES OF PRECISION OF ESTIMATES VS. MEASURES OF MATCH QUALITY
  TEST OF RANDOMIZED ASSIGNMENT TO TREATMENT
  ASSESSMENT OF COMMON SUPPORT

MAIN POINT: IN A PANEL SURVEY, THE FINAL ANALYSIS (OF IMPACT) WILL PROBABLY INVOLVE THE FIXED-EFFECTS TRANSFORMATION (E.G., DIFFERENCING OF HOUSEHOLDS), AND MANY VARIABLES WILL "DROP OUT." THERE IS NO POINT TO CONSTRUCTING MODELS OF THE RELATIONSHIPS OF RESPONSE VARIABLES TO EXPLANATORY VARIABLES FOR THE BASELINE DATA, WHEN THESE MODELS WILL EVENTUALLY BE IRRELEVANT.

INSTEAD, FOCUS ON ASPECTS OF THE BASELINE DATA THAT ASSESS THE USEFULNESS OF THE DATA FOR THE ENDLINE ANALYSIS.

USUALLY WILL NOT HAVE RANDOMIZED ASSIGNMENT TO TREATMENT, OR IF DO, IT IS A "CLUSTER-RANDOMIZED" DESIGN. SO IT IS OF INTEREST TO COMPARE THE TREATMENT AND CONTROL POPULATIONS, TO ASSESS THE QUALITY OF THE RANDOMIZED ASSIGNMENT TO TREATMENT.
WHEN DO ESTIMATION OF IMPACT VIA REGRESSION ANALYSIS, A MAJOR CONCERN WILL BE THE COMMON SUPPORT OF VARIABLES THAT ARE BELIEVED TO AFFECT OUTCOME (OF THE DIFFERENCES, NOT OF THE ORIGINAL DATA).

SO, IT IS OF INTEREST TO COMPARE THE SUPPORTS OF IMPORTANT EXPLANATORY VARIABLES.

THIS CAN BE CUMBERSOME, SINCE, AFTER THE BASELINE, THERE WILL BE MANY EXPLANATORY VARIABLES, AND MANY OF THEM WILL BE RELATED TO OUTCOMES OF INTEREST. TO REDUCE THE MAGNITUDE OF THE PROBLEM, DEVELOP A PROPENSITY-SCORE MODEL (SELECTION, NONRESPONSE), AND COMPARE THE DISTRIBUTIONS OF THE PROPENSITY SCORES FOR THE TREATMENT AND CONTROL GROUPS. (THE PARTICULAR VARIABLES IN THE MODEL ARE NOT HIGHLY IMPORTANT, SINCE THEY WILL BE CORRELATED (CONFOUNDED).)

[GRAPH SHOWING TWO PROPENSITY-SCORE DISTRIBUTIONS]


7.5. Use of Stata Statistical Program Package

1. USE OF STATA STATISTICAL PROGRAM PACKAGE

REFERENCES

Stata Reference Manuals


SUMMARY DESCRIPTION OF STATA

INTERACTIVE, .do COMMAND FILES (COLLECTION OF STATA COMMANDS), AND .ado FILES (A PROGRAM, CONSISTING OF A NUMBER OF STATA COMMANDS).

NOTE: THE KHANDKER REFERENCE INCLUDES INTERACTIVE EXAMPLES ON THE WORLD BANK WEBSITE, AT http://go.worldbank.org/FE8098Bl60

EXAMPLES

LIST DATA:
sort by treatment
list treatment tothhinc

COUNT DATA:

count
count if treatment=0

SUMMARIZE DATA (SUMMARY STATISTICS, SUCH AS MEAN AND VARIANCE):

summarize tothhinc

FREQUENCY DISTRIBUTIONS (TABULATIONS)

tab tothhinc

tab tothhinc if Round=0

MULTI-WAY TABLES (TABLES)

table inc1 inc2 inc3, contents(freq) by(treatment)

GENERATE NEW VARIABLES

generate inctot=inc1+inc2+inc3

MODIFY VARIABLES

generate senior=1 if age>64
replace senior=0 if age<=65

LABELING VARIABLES

label variable senior “Senior (age>64)”

LABELING VALUES

label define sexlabel 0 “Female” 1 “Male”
label values sex sexlabel

GRAPHS

histogram inctot

twoway (scatter inctot incemp), ytitle(Total Household Income) xtitle(Employee Income)
title(Total Household Income vs. Employee Income)

TWO-SAMPLE RANDOMIZED DESIGN

TWO-SAMPLE t-Test
\begin{verbatim}
ttest inctot, by(treatment)

REGRESSION MODEL
Regress inctot treatment

LOGISTIC REGRESSION MODEL
logistic treatment education farmsize

PROPENSITY-SCORE MODEL
pscore treatment education farmsize

AVERAGE TREATMENT EFFECT ON THE TREATED USING THE “RADIUS MATCHING” MATCHING ESTIMATOR
attr inctot treatment, pscore(ps98) radius(0.001) comsup

FOUR-GROUP RANDOMIZED DESIGN
Generate new observations by differencing income for matching households by survey round (not shown).
Apply t-test to differences:

ttest inctotd, by(treatment)

Model

\[ y = \alpha + \beta T + \gamma t + \delta DDTt + e \]

(data setup not shown)
generate RndTrt=Round*treatment
regress inctot Round treatment RndTrt

May include covariates:
regress inctot Round treatment RndTrt age educ farmsize

FIXED-EFFECTS MODEL

_xtset panelvar timevar
_xtset hhid Round
_xtreg inctot Round treatment RndTrt, fe

May include covariates:
_xtreg inctot Round treatment RndTrt age educ farmsize, fe

INSTRUMENTAL-VARIABLE MODEL:
ivreg and xtivreg commands
\end{verbatim}
Suppose Treated is an endogenous variable, and P is an estimated propensity score for Treated.

xtivreg lnctot educ agemployees farmsize equipmentvalue farmrentalvalue (Treated RoundTreated = P RoundP), fe

REGRESSION DISCONTINUITY
rd_sharp, rd_fuzzy

RECOMMENDED .do FILE HEADER TEMPLATE:

version 10.0
set more off
clear
capture log close

*In the following line, specify the name of the folder in which the data files are located.
global direct1 "C:\DataAnalysis\Project1"

log using "$direct1\Do10Project1ImpactEstimation.log", replace

*Place file header following the preceding statement, so it shows up in the log file.

*File name: Do10Project1ImpactEstimation.do
*Project: Project 1 Program Evaluation
*Creator: Joseph Caldwell
*Date created: 22 January 2012
*Modifier: Joseph Caldwell
*Date modified: 23 January 2012
*Purpose: Estimate impact for Project 1 program.
*Input file: Project1Data.dta.
*Output files: Project1ImpactEstimation.dta, Project1PropensityScores.dta.
*Log file: Do10Project1ImpactEstimation.log

*Repeat, as comments, commands prior to turning log on:
*version 10.0
*set more off
*clear
*capture log close
*In the following line, specify the name of the folder in which the data files are located.
*global direct1 "C:\DataAnalysis\Project1"
*log using "$direct1\Do10Project1ImpactEstimation.log", replace

set memory 100m
set matsize 800
set linesize 100

REVIEW OF A TYPICAL PROJECT FINAL REPORT.

SURVEY OF REFERENCES
8. References

Reference List
Statistical Design and Analysis in Evaluation

Causal Modeling


Additional references on structural equation modeling:


Evaluation


**Econometrics**


**Matching**


**General statistics (undergraduate-level mathematical statistics)**


**Technical references on estimation and test of hypotheses:**


**General statistics (less mathematical)**


**Survey sampling, less mathematical**


**Survey sampling, mathematical**


**Survey sampling, additional references (mix of mathematical and less mathematical)**


Kott, Phillip S., *Sample Survey Theory and Methods: A Correspondence Course*, Sept. 12, 2006. May be downloaded free from [http://www.nass.usda.gov/research/reports/course%20notes%200906.pdf](http://www.nass.usda.gov/research/reports/course%20notes%200906.pdf) This course has the prerequisite of one previous college-level course in statistics. It uses Sharon Lohr’s *Sampling: Design and Analysis* as a required text.


Thompson, Steven K., Sampling, 3rd edition, Wiley, 2012

Statistical Power Analysis


References on Experimental Design and Quasi-experimental Design


Cook, Thomas D. and Donald T. Campbell, Quasi-Experimentation: Design and Analysis Issues for Field Settings Houghton Mifflin, 1979

References on use of Stata statistical program software


Additional Material on Sample Survey Design for Evaluation

Reference on analysis of categorical data (binary response):