A Beginner's Guide to Government Data

Christopher Monsour

CAS Predictive Modeling Seminar
17 March 2008

TRAVELERS

Running Themes

- Government sources for survey data
  - Often free
  - Sometimes easy to access (DataFerrett)
  - Vendors often re-package it
    - But may want to look at the government documentation
- Micro-data
  - Sometimes micro-data is only released to academics or is released to them first
  - Nevertheless, many government surveys actually have data publicly released at the individual response level

Department of Transportation / BTS / FWHA

- National Household Travel Survey (every 5 years or so)
  - 1990, 1995, and 2001 microdata available on the web
  - 2001 sample has approx. 66,000 households
    - Only about 26,000 sampled nationally
      - Remainder from specific areas requested (purchased) by local governments
  - Detailed information on all travel by members of a household for 1 day, plus:
    - Detailed information on all trips over 50 miles for a 4 week period
    - Two odometer readings per vehicle, approximately 2 months apart
    - And of course demographics of the household and its vehicles
  - Available for free download

Department of Transportation / BTS / FWHA

- Motor Carrier Management Information System
  - Crash records (including some characteristics, such as type of cargo)
  - Inspections records (inspections performed, types of violations)
  - Certain fields not made available to the public
    - Privacy reasons: Information that would identify drivers
    - National security reasons: Hazardous materials information
  - Nominal fee
- Much else available
  - E.g., Motor vehicle recall databases
  - See Bureau of Transportation Statistics….Databases

Bureau of Labor Statistics—with microdata

- National Longitudinal Surveys
  - Truly unique
  - Initiated about twice a generation
  - Most recently in 1979 and 1997
  - Begin with sample of about 10,000 teenagers
  - Follow-up for annually or biennially for the next 25-50 years
  - Education, employment, family-life, salary, etc., history

Bureau of Labor Statistics—*sans* microdata

- Unlike the Census Bureau, the BLS tends *not* to release microdata for most of its surveys
- Some important examples include:
  - Safety and health statistics (injuries, illnesses and fatalities)
  - JOLTS (Job Openings and Labor Turnover Survey)
- These are done by industry, today using NAICS codes
  - NAICS = North American Industry Classification System
  - Old codes were SIC = Standard Industrial Classification
- BLS also has a lot of econometric series, like CPI and PPI

Current Population Survey

- Conducted monthly by the Census Bureau
- The BLS publishes many of the results
  - Underlies the unemployment rate
  - Respondents interviewed 4 months in a row and then again 4 months a year later
  - Certain questions (like income) asked only once in each 4 month period
  - 60,000 interviewed / month
- Microdata is available from both the Census Bureau and the National Bureau of Economic Research
- NBER page with the supplements data is http://www.nber.org/data/cps.html

Current Population Survey--Supplements

- Supplements—In many months additional questions are asked
  - Some of these are one-off, but many have become annual:
    - January (was Feb) of even years—Displaced Worker, Employee Tenure, and Occupational Mobility Supplement
    - March—"The March Supplement"—Social and Economic Supplement (includes noncash benefit information & migration)—comprises most of the Annual Demographic Supplement (ADS)
- Other recurring supplements
  - Veterans
  - Voting & Registration
  - Computer & Internet Use
  - Tobacco Use
  - Fertility & Marital History
  - School Enrollment

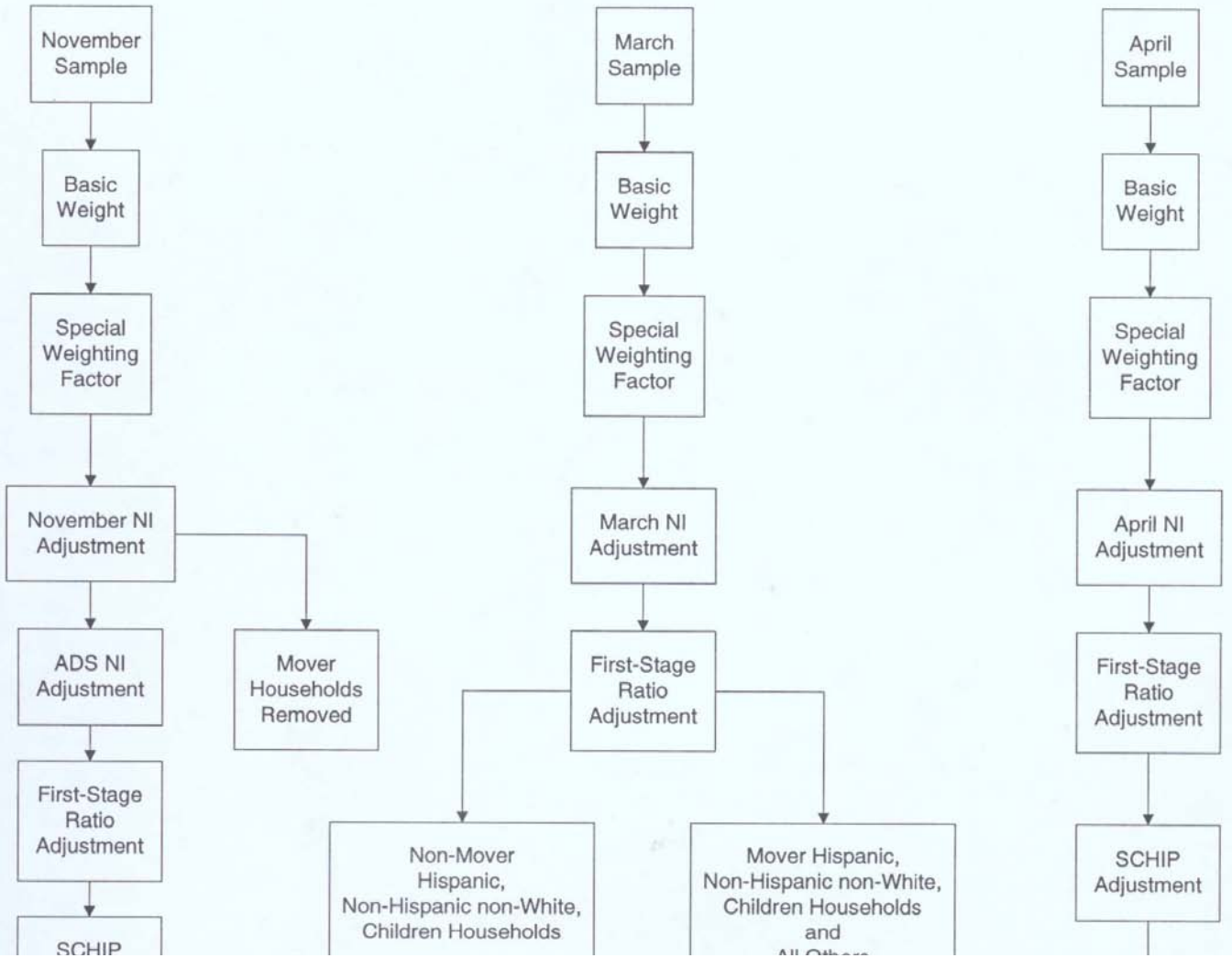
Items to be Aware of in Using Survey Micro-Data

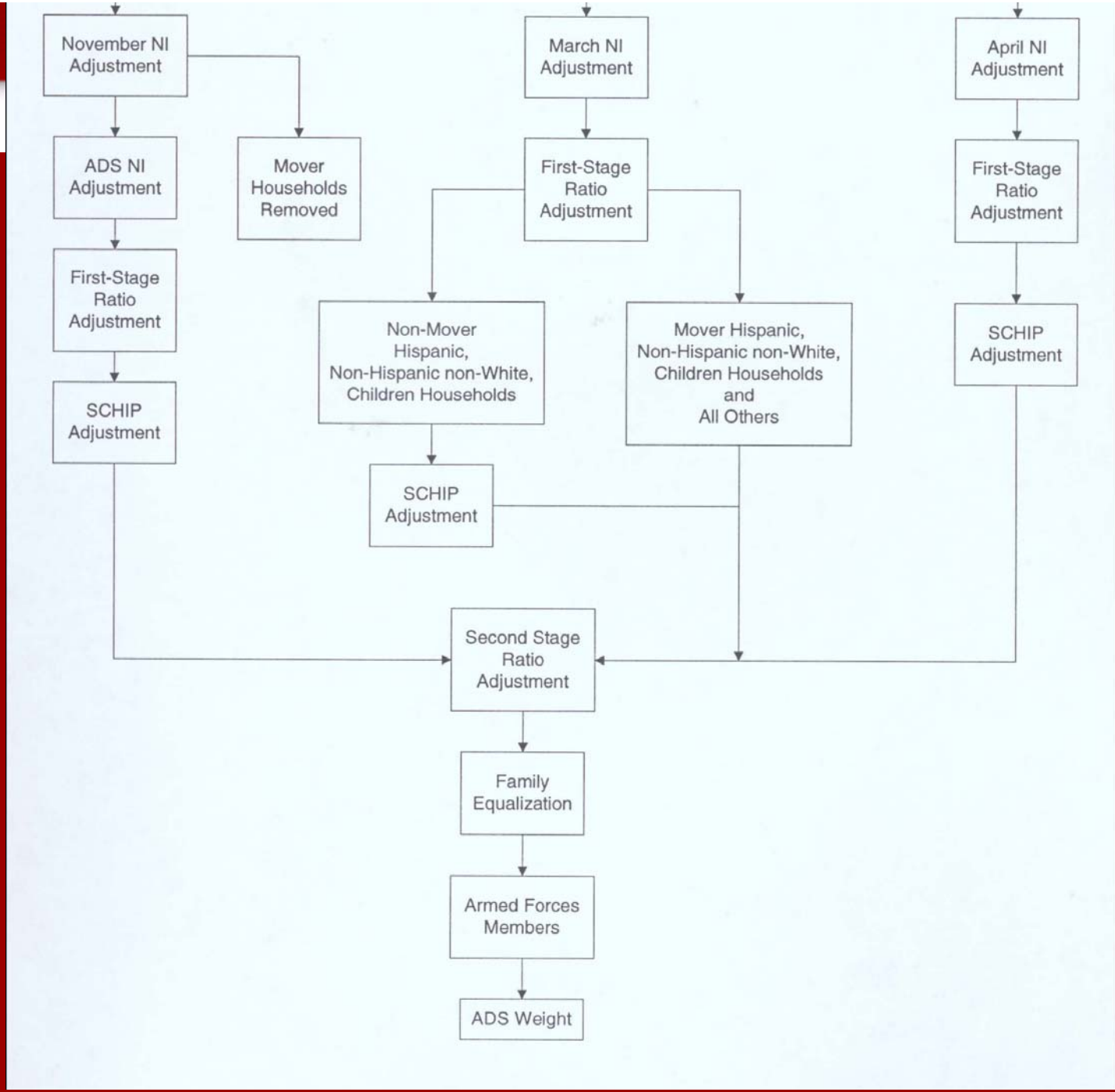
- Top- and bottom- coded values
- Missing value reasons
- Units of measurement
  - For instance, NLSY & CPS allow the respondent to report earnings on any calendar basis (monthly, weekly, hourly…)
- Weights (read the documentation for the survey in question)
  - Sampling weights AND replicate weights (not always available…)
    - Details of the sampling may be concealed to avoid identifying individuals
  - If a dataset contains multiple tables, there may be multiple sets of weights (e.g., household, person, vehicle)…be sure you are using the right ones
- Technical Documentation
- Questionnaire: How were the questions phrased

Why should you read the documentation?

- The answer doesn’t fit on this slide

Figure 11-1. Diagram of the ADS Weighting Scheme




November NI Adjustment
March NI Adjustment
April NI Adjustment
ADS NI Adjustment
Mover Households Removed
First-Stage Ratio Adjustment
First-Stage Ratio Adjustment
First-Stage Ratio Adjustment
SCHIP Adjustment
Non-Mover Hispanic, Non-Hispanic non-White, Children Households
Mover Hispanic, Non-Hispanic non-White, Children Households and All Others
SCHIP Adjustment
SCHIP Adjustment
Second Stage Ratio Adjustment
Family Equalization
Armed Forces Members
ADS Weight

Decennial Census and American Community Survey

- The American Community Survey ramped up to full speed in 2005, with approximately 3,000,000 households
- 2000 was the last Census long form
- ACS estimates will be released annually
  - But for areas with 25,000-60,000 people, estimates will be based on the latest 3 years
  - For areas with less than 25,000 people, estimates will be based on 5 years

Decennial Census and American Community Survey

- An obituary of Chip Alexander, principal creator of the ACS, at
  http://www.amstat.org/Sections/Srms/Proceedings/y2003/Files/JSM2003-000486.pdf
  references many ACS vs Census long-form differences

- PUMS (public-use microdata samples)—from both ACS and long-form
  - PUMA (public-use microdata areas) are smallest geographic unit identified
  - Data fuzzed about somewhat prior to release

Additional Banking, Financial, and Economic Data Resources

- The FDIC and NCUA websites have detailed financial data on banks and credit unions
- The St. Louis Fed website has an extensive collection of financial time series under the title “FRED”
- Detailed GDP information is available from the Bureau of Economic Analysis

Formats

- Some data available in much more friendly formats than others
  - For example, some data is in SAS datasets or SAS transport files (such as
  - Sometimes the data may be in flat files, but SAS formats are available
  - Sometimes a record layout is given, sometimes a SAS input statement is given (the latter with the CPS supplements from the NBER website)
  - Data Ferrett will download ASCII and SAS inputs and formats
    - Federal Electronic Research and Review Extraction Tool
    - Dataferrett.census.gov