

Data Ferrett And Microdata

Christopher Monsour

CAS Predictive Modeling Seminar
6 October 2008

Data Ferrett

dataferrett.census.gov

[The presentation here included a live demo of Data Ferrett—a tool for downloading both summarized and unsummarized data from government surveys. Most of this you can pick up by downloading Data Ferrett and using it hands-on. I should entice you, however, by mentioning that it is possible to get good metadata in formats compatible with various software tools such as SAS and SPSS. For instance, you can get a flat file download with SAS INPUT and PROC FORMAT statements appropriate for reading the data into SAS. Finally, you can also use DataFerrett to do *ad hoc* data tabulations without downloading the source data, and this runs pretty efficiently.]

Weights in Survey Microdata

- Probability Weights—for mean estimation
- Replicate Weights—for variance estimation
 - Not always provided
- Many ways these could have been computed
 - Read the survey documentation
 - Survey designs often complex
 - This makes determination of weights more complex

Typical Determination of Probability Weights

- Start with inverses of sampling probabilities (may all be equal, or may be intentionally oversampling certain strata..., e.g., a DoT survey may oversample counties or states that have paid for additional study)
- Adjust for differences between the primary sampling unit and the unit of interest
 - For instance, for a random dial survey
 - Divide by the number of telephones in the household
 - Multiply by the number of eligible persons in the household
- Adjust for non-response rates (e.g., by county)
- Balance to known subtotals for various combinations of variables. Often this is done for age, gender, race.
- Cap weights at a maximum value (more bias, less MSE)

Replicate Weights

- Variance estimation is usually done by a re-sampling or sub-sampling methods
- Why can't the user just do this?
- Because he needs to know **what the weights would have been** if the sub-sample (replicate) had been the entire sample
 - The necessary information on primary sampling units and strata might NOT be provided for privacy reasons
- So before the data are released, replicates are created, the weights computed for each replicate, and for each replicate, two fields are added to the data:
 - The weight with respect to that replicate
 - Membership (with multiplicity) in that replicate

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 20,000-65,000 people, estimates will be based on the latest 3 years
 - For areas with less than 20,000 people, estimates will be based on 5 years
- First release of 5 year data will be in 2010, based on 2005-2009 surveys
 - Tract and block group data will not be available until then
- Census Bureau's ACS Compass Products due out soon