

Discussion of:  
Coverage Measurement for the  
2010 Census  
by Tom Mule

Mike Cohen  
Committee on National Statistics  
WSS Seminar: January 23, 2008

# How I Know Something About This

- I am assisting the Committee on National Statistics' Panel on Coverage Evaluation and Correlation Bias in the 2010 Census, to whom Tom has presented some of this material.
- Much of my insight is stolen from the panel members: Bob Bell, Larry Brown, Rod Little, Xiao-Li Meng, Jeff Passel, Don Ylvisaker, and Alan Zaslavsky --- (final report by early summer).
- However, my comments today are my own. And all mistakes are my own.

# History of Use of Poststratification

- Mentioned in first paper by Sekar and Deming (1949)
- Important findings by Marks, Seltzer, Krotki *Population Growth Estimation* (1974), Jabine and Bershad (1968), and others
- Used in 1980, 1990, and 2000 census coverage measurement programs
- A lot is known about how it works and there is a great deal of comfort with it

# Advantages of Post-stratification for Coverage Measurement

- Poststratification is a special case of logistic regression when predictors are indicator variables for values of poststratification variables
- Homogeneous cells help ameliorate correlation bias
- Estimates for post-strata serve as foundation for small-area estimation via synthetic estimation

# Disadvantages of Poststratification for Coverage Measurement

- Variables have to be categorical--- some continuous variables that might be useful are age, household size at the individual level, and at an aggregate level: percentage vacant, renter, multi-units, minority – they have to be modified to be used
- Each additional variable adds multiplicatively to the number of cells, since you ‘kind of’ have to add all interaction terms, so only limited number of variables can be used if want to keep variances of estimates for each poststratum reasonable

# More Disadvantages of Poststratification for Cov. Meas.

- Collapsing can be used to 'delete' ineffective separations of cells, but it is not easy to figure out whether or not to do it
- Not easy to determine how to combine information from poststrata that share some characteristics but not others to reduce variances.
- With Logistic Regression, the model and associated tools make these things easy.

# Comments on Proposed Logistic Regression in 2010

- Except for age, the current predictors in the logistic regression models proposed by the Census Bureau are essentially all indicator variables for levels of factors used to define the poststrata in 2000, and their interactions, which limits the improvement that might be gained. Nice to see interest expressed today in expanding the variables included. (Though Schindler (2006) suggests otherwise)
- Small-area estimates are the sum of the ratios of the estimated correct enumeration rate divided by the estimated match rate applied over individuals --- so predictors are limited to census short form variables and contextual variables

# Comments on Age as a Factor

- Incredibly interesting relationship between match rate and correct enumeration rate as a function of age, especially the discontinuity around 17-20 yr olds
- Clearly, the spline approach is preferable to any categorical (stepwise function) of age and likely preferable to any small degree polynomial
- So why isn't it coming up as a more important modification? --- Plots may suggest not THAT many people aren't satisfied by the four-level step function.

# Challenges in applying logistic regression

- Model building in a predictive framework --- *predictive to support small domain use* --- cross validation useful way to deal with this
- Logistic regression diagnostics --- how can things like useful transformations be discovered?
- Incorporation of sample design in both model development and model assessment is needed and can be difficult for more complicated models (random effects, classification trees)
- Missing data methods are needed and their application is a little unclear involving both E- and P-samples
- What to do about non-data-defined observations?

# Some Questions and Remarks

- One possibility that was not mentioned that the Census Bureau might consider --- though it looks a bit ad hoc --- is the use of completely different log regression models in different types of areas (urban – rural), or for men and women.
- A big issue is what is the right objective function to determine whether a variable is beneficial?:
  - A. More homogeneous match and correct enumeration rates
  - B. Tom has looked at relative changes --- not sure I understand what that does for you
  - C. Goodness of fit (but what if more parameters?)
  - D. Cp, etc.
  - E. Cross-validation, which I like and the Bureau has used, computationally a lot of work
  - F. DSE fit --- might be best. Using artificial populations since don't want to be too dependent on 2000 data. A huge amount of work.
  - Tough problem --- what is best objective function?

# Some Questions and Remarks (continued)

- Impacts of outliers --- doesn't seem to be that big of a deal to me – it seems like the impact would be limited given the large samples – though I need to read Diffendal, Zaslavsky, Belin, Schenker (ARC 1994).
- The members of a household seem to have correlated match and correct enumeration indicator functions. Would this dependence cause overdispersion? Is this something GEE would address?

# Log. Reg. diagnostics--- this could help in selecting transformations

- Ed Fowlkes (1987) Biometrika, pp. 503-515 “Some diagnostics for binary logistic regression via smoothing”

Used Multi-LOWESS to provide smoothed dependent variable

Then you get something more like normal residuals when you subtract the fitted values

Allows you to create nice partial residual plots, combine with ACE (Breiman and Friedman) for transformations

Also allows you to create local deviance plots – a decomposition into pure error and lack of fit

A little better set of methods than Landwehr, Pregibon, and Shoemaker (1984)

Problem: incorporation of sample design in with diagnostics (though weights may be doable).

# Things I would encourage more of:

- A little more work on diagnostics
- Broader look at variables that might prove effective
- Broader look at other approaches besides logistic regression, especially classification and regression trees

# Things I would encourage more of:

- More work along the lines of Malec and Maples on random effects for census enumeration areas, or states, or even counties --- not discussed today --- very promising but very hard work --- unlikely to be ready in time for 2010, but you never know.

# Strong Support for General Approach

- Strong Praise: This is likely to provide major benefits in improved net undercoverage estimation and better understanding of the correlates of net undercoverage
  - My remarks have suggested some relatively major changes. This may have been silly on my part since the Census Bureau has to move cautiously in changing from poststratification to logistic regression, because:
    - as noted, there are some complications in doing this
    - there is real value in being able to make cross-census comparisons, which this change complicates
- Along those lines, I wonder about impacts on small-area estimates --- very likely to be very positive, but unclear if they will be produced so that may be moot.

# Finishing Up

- A pleasure to be asked to comment on this important and very successful research program. Congratulations Tom and others.
- There are already a lot of substantial changes in the coverage measurement program in 2010 that I have not mentioned so it is to their credit that they are undertaking another change with everything else on their plates
- I hope I haven't added unnecessarily to their "to do" list, but maybe some of the considerations I have raised will prove to be useful.
- Thanks