Using Continuous Variables As Modeling Covariates for Net Coverage Estimation

Vincent Thomas Mule, Jr., Donald Malec, Jerry Maples and Teresa Schellhamer
U.S. Census Bureau,

Keywords: Dual System Estimation

Abstract: The Census Bureau uses dual system estimation as one method to evaluate the coverage of the decennial census. This estimation for previous censuses has used post-stratification to minimize the impact of correlation bias on the population estimates. For 2010, we are planning on using logistic regression modeling instead of post-stratification cells. Logistic regression gives us the option of using variables in the model as continuous variables instead of having to form groupings. This research presents our initial results of the impact of using certain continuous variables in the model development and the resulting population estimates.

For the 2010 Census Coverage Measurement (CCM), we are exploring using logistic regression modeling instead of post-stratification cells in the estimation. We believe by using logistic regression that we can potentially utilize more variables than we have used in the past in trying to minimize the impact of correlation bias and high variances. Logistic regression gives us the option of using variables in the modeling as main effects and not having to introduce any unnecessary interactions. The use of post-stratification in the past required some of the cross-classifications of variables to be collapsed because of small sample sizes and/or high variance estimates of the post-stratification cell.

In addition to potentially utilizing more variables, logistic regression can also use variables in the model as continuous variables. When we examined the previous research, we saw that area-level rates, age and household size were used as continuous variables to evaluate the 1990 Post-Enumeration Survey (PES) post-stratification. This analysis and presentation focuses on our research on using age in the modeling as a continuous variable. Mule, Schellhamer, Malec and Maples (2007) documents the analysis of the other variables. The source of the data is the A.C.E. Revision II data.

"This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau."

Relationship of Match and Correct Enumeration Status to Age

We started by examining the relationship of both correct enumeration and match status to age by graphical methods. This allowed us to examine what representations might be the best to use for these types of variables.

Figure 1 shows the relationship between match rate and age for the P-sample cases. The figure shows a nonlinear trend from 0 to 17, followed by a sharp decrease from 17 to 20, then a nonlinear trend from 20 to the top coded age of 90 in this figure. The match rate of each age is shown with a 95 percent confidence interval.

The figure also shows age heaping of the match rates for ages that end in 0 or 5, especially between 20 and 50. Since Figure 1 includes both nonmovers and outmovers, one hypothesis was this might be caused by the outmover cases who were all obtained by proxy. This has methodological implications because we are changing from the PES-C procedure to handle movers used in 2000 to the PES-B procedure for 2010. The PES-C procedure matched the outmovers in the sample cluster search area. The PES-B procedure uses inmovers instead and matches them at their census day address. If this heaping result was caused entirely by outmovers then these patterns would be less likely to show up in our 2010 results and we would like to account for that when using 2000 data and results for our research.

Proxy-response was used in the separate post-stratification of the correct enumeration status for the E sample. One of the criticisms of the A.C.E. Revision II estimates was the lack of balancing of erroneous enumerations and nonmatches because of the separate post-stratifications. These resulted in large overcount estimates for areas with high concentrations of proxy census enumerations. Using an indicator of ages ending in 0's or 5's between 20 and 50 in the correct enumeration and match status modeling may be beneficial. This indicator could be used in both the correct enumeration and match models and may explain some of the proxy results shown in these figures.

Figure 2 shows the correct enumeration rate by individual years of age. These figures show mostly the same results for correct enumeration rates as was shown for the match rate analysis. The same methodology was used as for the creation and analysis of the match rate figures. The figure shows that the correct enumeration rate appears to have a minimum point estimate at 19. The match rate figures showed a minimum point estimate of 20.

This analysis showed for age that the relationship to both rates was nonlinear and changed for different ages. These graphs led to our decision to consider using a four-piece spline approach to model the different parts. The four parts were a curvilinear relationship from 0 to 17, linear from 17 to 20, curvilinear from 20 to 50 and then linear from 50 on.

Using Age As a Continuous Variable

Age has been used as a continous variable in previous evaluations of census coverage. Alho et al (1993) used a third degree polynomial representation of age in their model. We started our research on age as a continuous variable by using this representation.

Figure 3 shows the results of using age as a third degree polynomial to model the match rate. The predicted match rate of the third degree polynomial is shown in comparison to the estimates by individual ages. We can see that the modeling does better for the 30+ population than the below 30 population. Alho et al (1993) stated that they found less of an effect with using this representation than they were originally expecting.

ACE Revision II formed five categories of age: 0-9, 10-17, 18-29, 30-49 and 50+. We can always use these categories again. However, there are some trends in the age results of match and correct enumeration rates that we would like to see if they provide any improvement to the estimation.

One solution is to model age by constructing a spline model with polynomial pieces of different degrees. Smith (1979) presents a methodology of spline regression using truncated polynomial functions. This approach allows standard multiple regression procedures to be used to show these relationships. Survey weights and replication techniques can then be used to generate standard errors that account for the complex sample design.

Malec et al. (1997) used piecewise linear spline modeling for age in their examination of the small area inference for binary variables in the National Health Interview Survey. This approach allowed changes in slopes at ages 15, 25 and 55 and different slopes for males and females at ages 15 and 25. Nandram and Choi (2005) used a similar join-point regression modeling of age in their hierarchical Bayesian nonignorable nonresponse regression models for small areas.

In examining the match and correct enumeration results by age, we can see four distinct parts that can possibly be represented by a continuous spline:

1. quadratic relationship from 0 to 17,
2. linear relationship from 17 to 20,
3. quadratic relationship from 20 to 50
4. linear relationship from 50 on.

Using Smith's notation of "+" functions to show the truncated polynomial functions, this four-part relationship can be expressed by the following six covariates in a logistic regression model for match rate. The same model form can be used for the correct enumeration rate as well.

4

$$\mathit{Logit}(\mathit{Mrate}) = B_0\mathit{Int} + B_1 \times \mathit{Age} + B_2 \times \left(\mathit{Age}^2 - (\mathit{Age}-17)_+^2\right) + B_3 \times (\mathit{Age}-17)_+ + B_4 \times (\mathit{Age}-20)_+ + B_5 \times \left((\mathit{Age}-20)_+^2 - (\mathit{Age}-50)_+^2\right) + B_6 \times (\mathit{Age}-50)_+$$

The “+” function assigns the maximum value of either the term in the parentheses or 0 based on the age of the person. For example, the first function is $(\text{Age-17})_+$. If the person record is 7 years old then this person will receive a value of 0 from this function. If the person record is 20 years old then this function will return a value of 3.

Stepwise Exploratory Model Building

We used a stepwise exploratory model building approach to examine what first and higher order interactions should be included in the models. We started by using Race/Origin domain, Age, Sex and Tenure (ROAST).

The seven categories of the Race/Origin domain variable are:

- Non-Hispanic White and Other
- Non-Hispanic Black
- Hispanic
- Non-Hispanic Asian
- Native Hawaiian and Pacific Islander
- American Indian on American Indian Reservation
- American Indian off American Indian Reservation

These have been the foundation of the post-stratifications for past coverage measurement evaluations. We examined two different ways to represent Age and Sex. The first used the seven age/sex categories used for the March 2001 estimation. The second used the four piece spline relationship seen in the graphical analysis and sex. These stepwise models had between 30 and 40 main effects and interactions in the models.

We then generated population estimates using these two sets of selected models. For comparison, we generated estimates from a 104 cell post-stratification using just the ROAST variables.

The N2 estimator uses only the sample data. The data-defined records in the census that were eligible for matching are represented by the E sample. Based on results of the modeling and the characteristics of each E-sample case, we can estimate a predicted probability of the correct enumeration and match status. This estimator may be more appealing than the N1 estimator if good covariates are only available for the sample cases and not all of the enumerations in the census. This may be more beneficial in future research when additional variables are explored.

This estimator produces higher standard error estimates for individual levels than the A.C.E. Revision II standard errors.

The formula for the N2 estimator is:

$$\hat{N}_2 = \sum_{j \in ESample} \frac{\pi_{ce(j)}}{\pi_{m(j)}}$$

where $w_{e(j)}$ is the adjusted sampling weight of the E-sample case,
$\pi_{ce(j)}$ is the predicted correct enumeration probability from the model and
$\pi_{m(j)}$ is the predicted match probability from the model.

In this analysis we compare the estimates from two different sets of research models to the research estimate using the 104 post-stratification by computing the relative difference by the following formula:

$$\frac{Model\ Estimate - 104\ Estimate}{104\ Estimate}$$

Standard error estimates were computed using a 100 random group jackknife methodology. Results for the overall housing unit population and the seven race/ethnicity domains are shown in Table 1.

Our results showed:

- For National and Tenure estimates, this comparison showed no significant differences for two sets of population estimates based on the new models.

- For Race/Origin domain estimates, both sets of new estimates showed significantly lower estimates for the Hispanic and Non-Hispanic Black domains. Both also showed significantly higher estimates for the Non-Hispanic White and Other domain.

- We generated graphs of the relative differences of the two models estimates compared to the post-stratification estimates. These graphs showed that using age in a continuous form as compared to groupings generated different relative differences within the age groupings.

Future Research

This documents some of the completed work on examining age for logistic regression modeling. We will continue to explore this topic by examining the following:

- Investigate other representations of the age spline. The current representation results in a continuous function and only a linear relationship from 50 to 80. We could explore allowing different intercepts or "jumps" for parts of the spline that are consistent with Smith's methodology. We can also explore if a quadratic form might be better for the 50+ population modeling.

- Using partial residual plots as suggested by Landwehr, Pregibon, and Shoemaker (1984) to help determine which additional variables should be included in the model. These plots can also examine if the variables should be transformed in certain ways to help performance.

- Continue exploring the addition of other variables that were significant predictors for the 2000 post-stratification. Region, Metropolitan Statistical Area (MSA), Type of Enumeration Area (TEA) and mail return rate were used for the Non-Hispanic White and Other post-stratification. We will also examine using housing unit size and mail return rate as continuous variables in the regression. These other variables may help improve the estimation of the Non-Hispanic White and Other domain since the variables considered in this initial research were used in the analysis of minority populations in the 1990 Census. We can also examine using an indicator for ages that end in "0" or "5" to possibly capture some age heaping because of proxy reporting.

- The spline representation used for age may also be a candidate representation for other continuous variables like the cluster rates.

References

Alho J., Mulry, M., Wurdeman, K., and Kim, J. "Estimating Heterogeneity in the Probabilities of Enumeration for Dual System Estimation" Journal of the American Statistical Society, September 1993.

Landwehr, J., Pregibon, D., and Shoemaker, A. "Graphical Methods for Assessing Logistic Regression Models" Journal of the American Statistical Association, March 1984.

Malec, D, Sedransky, J., Moriarity, C., and LeClere, F., "Small Area Inference for Binary Variables in the National Health Interview Survey" Journal of the American Statistical Association, September 1997.

Mule, T., Schellhamer, T., Malec, D. and Maples, J. "Using Continuous Variables as Modeling Covariates for Net Coverage Estimation" DSSD 2010 Census Coverage Measurement Memorandum Series 2010-E-09-R1, March 6, 2007.

Nandram, B and Choi, J., "Hierarchcial Bayesian Nonignorable Regression Models for Small Areas: An Application to the NHANES Data" Survey Methodology, 73-84, 2005.

Smith, P., "Splines As a Useful and Convenient Statistical Tool" The American Statistician, May 1979.

Figure 1: Match Rate by Age

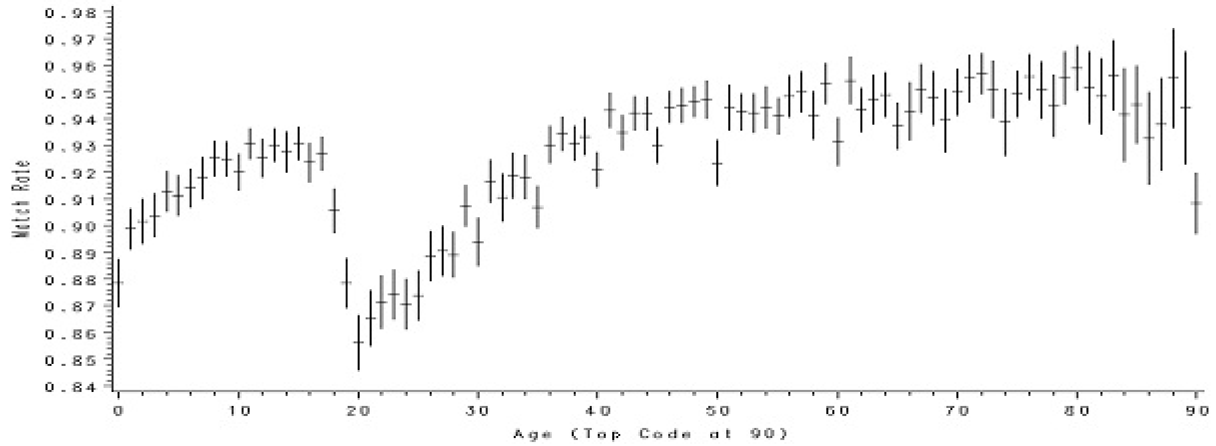

Figure 2: Correct Enumeration Rate by Age

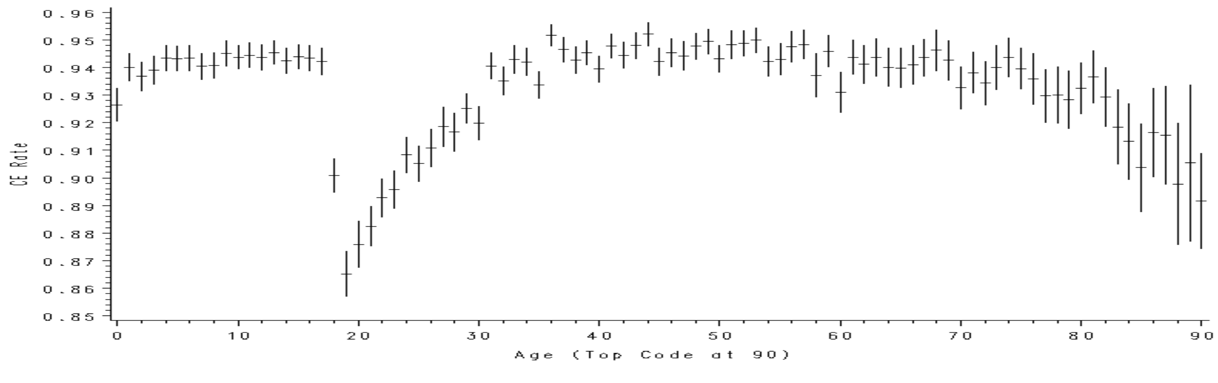

Figure 3: Individual Age Match Rate Estimates vs. Predicted 3rd Degree Polynomial Rates

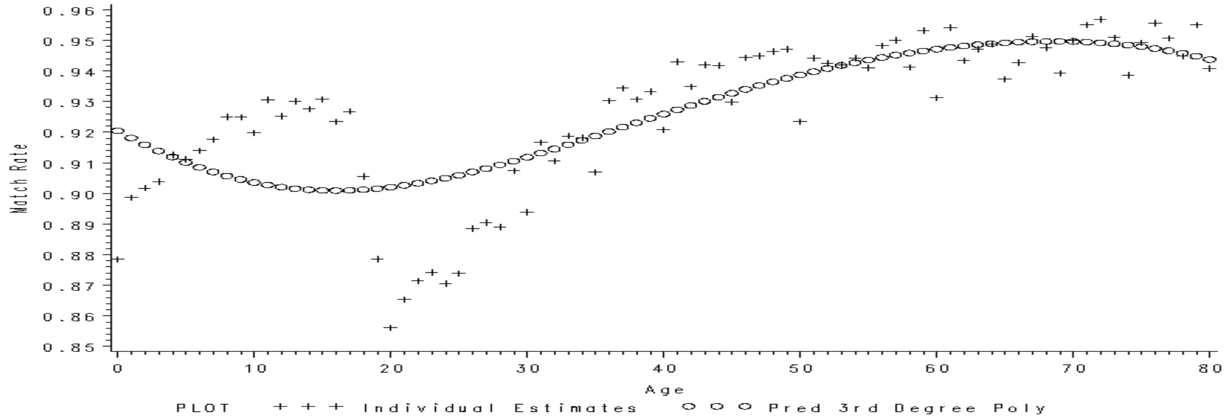

Figure 4: Individual Age Estimates of Match Rates vs. Predicted Results of the Age Models

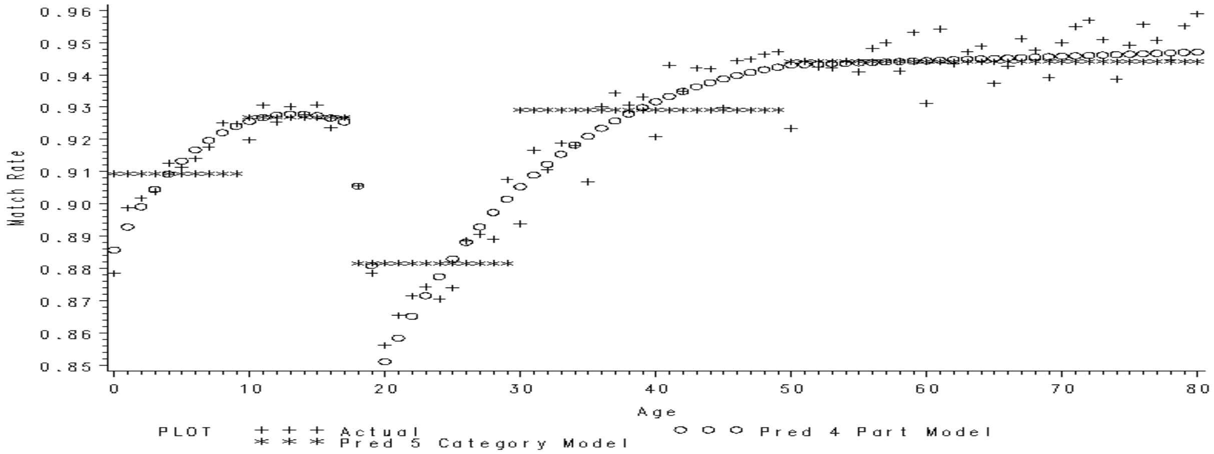

Table 1: Differences of Models from 104 Post-stratification

Relative Difference of Stepwise Models from the 104 Post-stratification Estimate for the Race/Ethnicity Domain Population Estimates		
Race/Ethnicity Domain	ROAST Models Using 8 Age/Sex Categories	ROAST Models Using 4 Part Spline & Gender
Overall HU Population	-0.0018% (0.0027%)	+0.0009% (0.0033%)
AIAN on AIR	-0.013% (0.043%)	+0.025% (0.046%)
AIAN off AIR	-0.070% (0.098%)	-0.113% (0.117%)
Hispanic	-0.045%[1] (0.011%)	-0.054%[1] (0.016%)
Non-Hispanic Black	-0.056%[1] (0.013%)	-0.036%[1] (0.012%)
NHPI	-0.075% (0.191%)	-0.055% (0.209%)
Non-Hispanic Asian	-0.007% (0.023%)	-0.008% (0.023%)
Non-Hispanic White and Others	+0.016%[1] (0.003%)	+0.019%[1] (0.004)