Paper SDA-11

Developing a Model for Person Estimation in Puerto Rico for the 2010 Census Coverage Measurement Program

Colt S. Viehdorfer, U.S. Census Bureau, Washington, DC

This report is released to inform interested parties of ongoing research and to encourage discussion. Any views expressed on technical issues are those of the author and not necessarily those of the U.S. Census Bureau.

ABSTRACT

One of the goals of the 2010 Census Coverage Measurement (CCM) program is to estimate net coverage error for persons in housing units for the 2010 Census, with Puerto Rico results being calculated independently from the rest of the United States. General logistic regression is being used for the estimation of net error for the 2010 Census as opposed to using a post-stratification method, as was done in previous census coverage measurement surveys. Unlike post-stratification, logistic regression allows the use of continuous variables. This paper will outline the steps that I took to develop a logistic regression model for net coverage error estimation in Puerto Rico, using data from the 2000 Census and its coverage measurement survey. I will explain how I determine the main effects to be included in the model using various exploratory and statistical techniques and will also examine different model selection procedures for deciding on a final model, which will include main effects and interactions. The main effects that I have chosen to use in the model in this paper will be proposed to be used as the main effects for the model for the 2010 CCM. However, specific interaction terms to include in the model will be determined using the procedures outlined in this paper once the actual 2010 CCM data becomes available.

INTRODUCTION

Estimation of net coverage error for persons in housing units using dual system estimation requires data from the census and also from a sample taken independently of the census. This is often a post-enumeration survey. In 2000, the post enumeration survey was called the 2000 Accuracy and Coverage Evaluation (A.C.E.), and now it is the 2010 CCM. For both the A.C.E. and the CCM, two samples are identified during the post enumeration survey. For the survey, the Primary Sampling Unit (PSU) is the block cluster. The P sample consists of housing units and persons in housing units that are included in the selected block clusters and subsampled sections of the block clusters. As previously mentioned, housing units in these selected areas are identified independently of the census list. Housing units and persons in housing units that are listed in the census and are within the identified block clusters and block cluster segments identified by the P sample are part of the E sample. Since the two samples are identified independently, some housing units and persons in housing units could be in one sample but not the other.

Two important components required to estimate net coverage error are the correct enumeration (CE) rate and the match rate. In the broadest sense, a correct enumeration is a person who was correctly listed in the census. Otherwise, the enumeration is determined to be erroneous. E sample cases are assigned an enumeration status. A match is a P sample person who is determined to be the same person as someone listed in the census. Otherwise, the person is a nonmatch. Two models will be developed, one for estimating E-sample correct enumerations and the other for estimating P-sample matches. Various SAS® procedures will be used for model development. The SURVEYLOGISTIC procedure and the LOGISTIC procedure will be used mostly for the model development. Other SAS procedures are used for other parts of the analysis along with some SAS macro coding.

STUDY PLAN

Logistic regression will be used for estimation of net error for the 2010 Census as outlined in Griffin (2005).

MODELING VARIABLES

The same variables that were used in 2000 A.C.E. Puerto Rico person estimation and are discussed in Haines (2001) will also be examined for 2010 CCM. Modifications and transformations will be considered for certain variables. The following variables are examined:

Tenure – same two categories used in 2000 Puerto Rico person estimation:
Owner, Renter
MSA (Metropolitan Statistical Area) – same three categories used in 2000 Puerto Rico person estimation:
1=San Juan CMSA (Combined MSA), 2=Other MSA, 3=Non-MSA

Return Rate Indicator – Return Rate is a variable measuring the proportion of housing units in the mailback universe in each tract that returned a census questionnaire. The same two categories used in 2000 Puerto Rico person estimation:

High, Low

Age/Sex Groups – same seven categories used in 2000 Puerto Rico person estimation:

1=under 18, 2=18-29 male, 3=18-29 female, 4=30-49 male, 5=30-49 female,
6=50+ male, 7=50+ female

Since 2010 CCM Puerto Rico person estimation is using logistic regression instead of post-strata to estimate correct enumeration and match rate, continuous forms of return rate and age will be considered as alternatives to the categorical variables used in the 2000 A.C.E. Additionally, transformations will be examined for continuous variables. Research involving continuous forms of the age variables includes considering the use of age splines in the model. The same age splines documented in Mule (2007) are considered. Alternate variables for research for 2010 CCM are:

Sumratesq – In place of Return Rate Indicator, return rate percentage is rounded to the nearest five percent. Then, the square of the new variable is applied to the model.

Age Splines – In place of the AgeSex categorical variable, consider continuous Age Splines:
quadratic from 0-17, linear from 17-20, quadratic from 20-50, linear from 50-80 (top coded at 80).

MODELING PROCEDURE

The variables included in the previous section will be added to the models predicting correct enumeration rate and match rate and tested for parameter significance. This procedure will be performed using PROC SURVEYLOGISTIC because it adjusts variance estimates for the different design effects implicit in the A.C.E and CCM's complex sample design. Prior to testing the effects, continuous age and return rate variables will be examined in place of the AgeSex post-stratification variable and Return Rate Indicator binary variable. A final model can be determined once 2010 CCM data is available by testing for significance of the parameters and also by applying a cross-validation technique which will be discussed later in the document. Estimates of persons in housing units can be determined using various estimators, one of which is the N2 estimator that will be discussed later.

RESULTS

SELECTING FORM OF RETURN RATE VARIABLE

Since a logistic regression model is being used for Puerto Rico person estimation instead of post-stratification, using a continuous form of the return rate variable in the model may improve the model fit. For both the E and P samples, the return rate percentage is rounded to the nearest five. The new variable is called Sumrate. A logistic model is run for correct enumeration rate and match rate using MSA, Tenure, and AgeSex as variables. A predicted probability is output by PROC LOGISTIC. By comparing the means of the predicted probabilities to the means of the actual response, a determination can be made as to whether a continuous return rate variable should be added to the model. Means are calculated using the MEANS procedure. Output for the E and P samples are given below:

E Sample
(z is the actual response with z=1 for a correct enumeration and z=0 for an erroneous enumeration)

Sumrate	N		Mean
0.05	121	z	0.9256198
		P_1 Predicted Probability: z=1	0.9125171
0.25	123	z	0.8831249
		P_1 Predicted Probability: z=1	0.9169033
0.3	122	z	0.7815712
		P_1 Predicted Probability: z=1	0.9337588
0.35	1196	z	0.8357773
		P_1 Predicted Probability: z=1	0.9257771

0.4	3158	z	0.9200214
		P_1 Predicted Probability: z=1	0.9280897
0.45	3762	z	0.9060313
		P_1 Predicted Probability: z=1	0.9283294
0.5	10517	z	0.9342703
		P_1 Predicted Probability: z=1	0.9316507
0.55	9355	z	0.9458351
		P_1 Predicted Probability: z=1	0.9328271
0.6	5409	z	0.9468011
		P_1 Predicted Probability: z=1	0.9330523
0.65	2411	z	0.9581819
		P_1 Predicted Probability: z=1	0.9377649
0.7	73	z	0.9499572
		P_1 Predicted Probability: z=1	0.9349648

P Sample
(z is the actual response with z=1 for a match and z=0 for a nonmatch)

Sumrate	N		Mean
0.05	109	z	0.6972477
		P_1 Predicted Probability: z=1	0.8194689
0.25	64	z	0.6875000
		P_1 Predicted Probability: z=1	0.8402204
0.3	131	z	0.9541985
		P_1 Predicted Probability: z=1	0.8441112
0.35	993	z	0.8207452
		P_1 Predicted Probability: z=1	0.8596813
0.4	2949	z	0.8338420
		P_1 Predicted Probability: z=1	0.8555739
0.45	2983	z	0.8132752
		P_1 Predicted Probability: z=1	0.8620764
0.5	8967	z	0.9188134
		P_1 Predicted Probability: z=1	0.8619546
0.55	8541	z	0.8675799
		P_1 Predicted Probability: z=1	0.8670806
0.6	4509	z	0.9046352
		P_1 Predicted Probability: z=1	0.8620895
0.65	1938	z	0.8921569
		P_1 Predicted Probability: z=1	0.8587995
0.7	54	z	0.8888889
		P_1 Predicted Probability: z=1	0.8806360

By comparing the means of the actual and predicted values for both the P and E samples, it is apparent that there are large differences in the means of the predicted probabilities using the model and the means of the actual responses. Based on these results Sumrate will be included in the logistic regression model.

Transformations are to be considered for the Sumrate variable in order to get a form of the variable that best fits the data. For the P and E samples, the square root of the Sumrate variable along with the square of the Sumrate variable will be compared to the original Sumrate using -2 times the log likelihood and the Wald Chi-square statistic output by PROC SURVEYLOGISTIC. An identical log likelihood can be obtained by PROC LOGISTIC; however, the SURVEYLOGISTIC procedure gives a different value for the Wald Chi-square because it takes into consideration the sample weights unlike the LOGISTIC procedure. Output is given for each of the three models in the P and E samples. Note that MSA, Tenure, and AgeSex are included in all of the models. Since all models have the same number of degrees of freedom, direct comparisons can be made.

E sample

	-2 LogLikelihood	Wald Chi-square
Sumrate	1741268.2	233.4236
Sqrt(Sumrate)	1744578.2	223.2374
Sumrate**2	1739375.0	220.1463

P sample

	-2 LogLikelihood	Wald Chi-square
Sumrate	2723250.0	208.6158
Sqrt(Sumrate)	2724310.6	213.7273
Sumrate**2	2723947.2	189.4723

The difference between the highest log likelihood and the lowest log likelihood for the E sample is approximately five thousand. However, the difference between the two most extreme values on the P-sample side is only about five hundred, meaning the choice of transformation is less important when predicting match rate. The squared version of the Sumrate variable has the lowest log likelihood for the E sample, indicating the best fit. Since there is such a small difference between values for the P sample, the squared transformation of the Sumrate variable will also be used in order to maintain consistency with the E sample. Transformation of the Sumrate variable should again be examined once 2010 CCM data is available.

SELECTING FORM OF AGE VARIABLE

Again, by using a model-based approach to estimation a continuous age variable can be added to the model in place of the seven AgeSex post-strata if the continuous variable improves the fit of the model. First, a similar approach will be used for age as was used for return rate to determine if a continuous variable should be considered. Age is rounded to the nearest five and means are produced for the actual value and the predicted values using MSA, Tenure, and the square transformation of the Sumrate variable (Sumratesq) for both the E and P samples.

Although the means are not given here in order to conserve space, examination of the means for the different age groups provides indication that there is enough of a difference between the means that a continuous age variable warrants consideration. Looking at residual plots can provide further insight into whether or not a continuous variable should be used and can also give clues as to what type of continuous variable should be used. Figure 1 graphs the residuals at each age rounded to the nearest five for a logistic regression model predicting match rate using MSA, Tenure, and Sumratesq. Figure 2 shows a similar graph for the residuals of a model with the same predictors, but the model is attempting to predict correct enumeration rate. The code used to calculate the residuals and to plot them is given here:

```
proc sort data=match1; /* P sample output data set from logistic regression */
   by sumage;  /* sumage is age rounded to the nearest 5 */
run;
```

```
data cumage(keep=cumm cumx cumy diff);
  set prapxrat;
  by sumage;
  if first.sumage then do;cump=0;cumm=0;cumx=0;cumy=0;end;
  cump + zwgt;         /* zwgt is the weight variable  */
  cumm + zwgt*z;       /* z=1 for matches and 0 for nonmatches  */
  cumx + zwgt*p_1;     /* p_1 is modeled probability of a match  */
  cumy + zwgt*sumage;
  diff=(cumm-cumx)/cump;
  if last.sumage then do;
    cumm=cumm/cump;
    cumx=cumx/cump;
    cumy=cumy/cump;
  output;
  end;
run;

title "Match Rate Residual By Age";
axis1 label=("Residual")
      order=(-.05 to .04 by .01);
axis2 label=("Age")
      order=(0 to 80 by 5);

proc gplot data=cumage;
  plot diff*cumy/haxis=axis2 hminor=1
                 vaxis=axis1
                 overlay regeqn;
       symbol1 interpol=sm50 value=dot;
run;
quit;
```

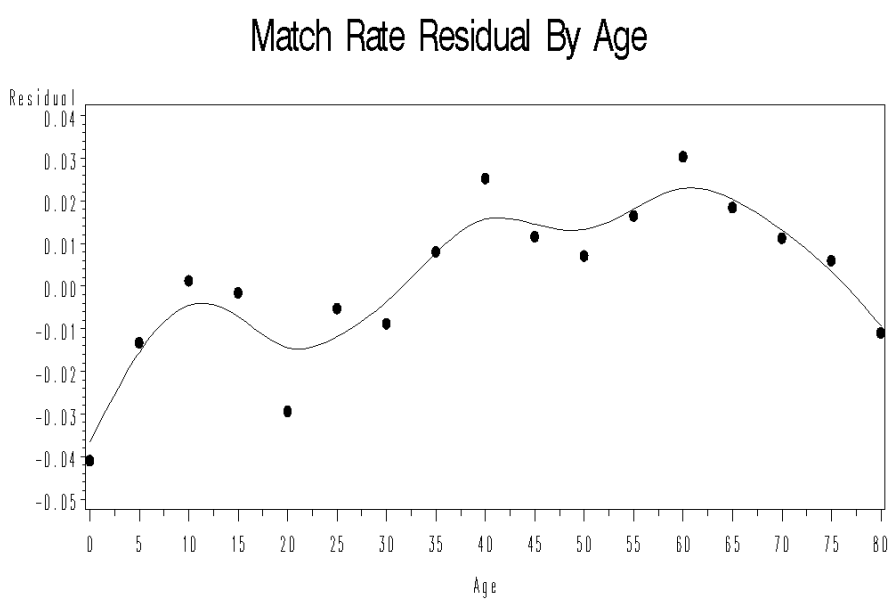

Figure 1. *No Age Variable*

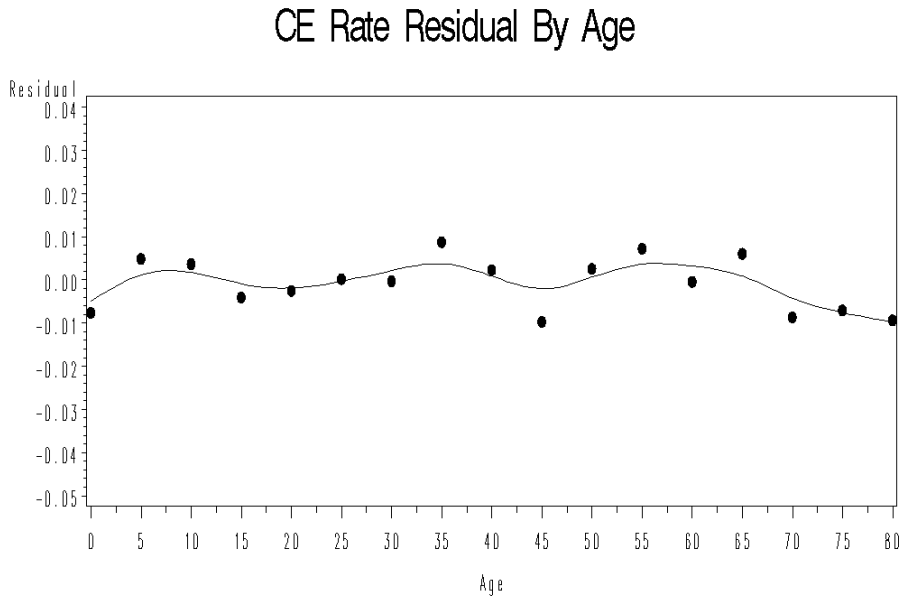

Figure 2. *No Age Variable*

A model that fits the data well will have nearly flat residuals, and it is apparent by looking at the graphs that neither of the models produces flat residuals. The model that predicts match rate has larger residuals than the model predicting correct enumeration rate. The plots indicate that the age splines discussed in Mule (2007) might be appropriate to add to the model. Age splines were created using the following code:

```
data agesplines;
    set zp;
    xx00=age;
    if xx00 gt 80 then xx00=80;
    if xx00 gt 17 then xx17=xx00-17; else xx17=0;
    if xx00 gt 20 then xx20=xx00-20; else xx20=0;
    if xx00 gt 50 then xx50=xx00-50; else xx50=0;
    xx02=(xx00**2) - (xx17**2);
    xx22=(xx20**2) - (xx50**2);
    xx52=(xx50**2);
run;
```

Figure 1 gives indication that adding a quadratic 50+ age term to the P-sample model in addition to the age splines already in the model may be appropriate. To try to determine the best age variable, four models are fit for the E and P samples. A model without any age variable, a model with the AgeSex categorical variable, a model with Age Splines, and a model with Age Splines crossed with a sex variable are all fit. As previously mentioned, an additional model will be fit specifically to the P sample with a 50+ quadratic term added to the Age Splines. To compare the various models, the log likelihood (divided by the sample weight) and the Wald Chi-square statistic will be taken from output given by PROC SURVEYLOGISTIC. In addition, a cross-validation will be performed on each of the models as described in Griffin (2005). The data is divided into only twenty different groups for this analysis due to the small sample size in Puerto Rico. Similar to the log likelihood, a lower value for cross-validation indicates a better model fit. The log likelihood will almost certainly be better for a model with more parameters, thus the cross-validation measure is used to determine if the model is over-fitting the data. To calculate the cross-validation value, the following code was used:

```
/* Example modeling match rate with P-sample data */
%macro xval;
     %do j=1 %to 20;
       data loop20;
            set zp; /* The P-sample data set */
            if group~=&j then z2=z; /* Only 19 groups get dependent variable */
            else z2=.;
       run;
```

```
      /* Models the probability of a match */
      proc logistic data=loop20 noprint desc;
        class msagroup(ref='1') tenure2(ref='1') agesex;
        weight zwgt;
        model z2(ref='1') = msagroup tenure2 sumratesq xx00 xx17 xx20 xx50
                                                  xx02 xx22;
        output out=out&j (where=(z2=.)) predprobs=I;
      run;

    /* Calculates the log penalty function of each observation */
    data loss&j(keep=zwgt z ip_1 LP);
         set out&j;
       LP=zwgt*(z*log(ip_1) + ((1-z)*log(1-ip_1)));

    run;

    /* Finds the sum of all log penalty functions and sums the sampling weights */
    proc summary data=loss&j;
      var LP zwgt;
      output out=measures&j sum(LP)=totalp  sum(zwgt)=groupwgt;
    run;



    /* Calculates the log penalty function for the replicate */
    data grouppen&j;
         set measures&j;

    LPF=(1/groupwgt)*totalp;

    run;
%end;
%mend xval;
%xval
```

The final cross-validation result is calculated by combining all of the data sets containing the replicate log penalty functions using the APPEND procedure and then taking the mean of the twenty log penalty functions. The cross-validation is similar to a replicated log likelihood.

Output for the P sample is given below:

P sample

	Wald	LogLik	Cross-validation
No Age	164.0074	-.3965	-.40078
AgeSex	189.4723	-.3955	-.40034
Age Splines	195.4618	-.3955	-.40032
Age Splines Crossed With Sex	204.0474	-.3951	-.40015
Age Splines With 50+ Quadratic	199.4051	-.3953	-.40032
Age Splines With 50+ Quadratic Crossed With Sex	209.3888	-.3949	-.40020

The results seem to indicate that the model containing the Age Splines crossed with Sex to be the model that best fits

the data by having the lowest value for the cross-validations. However, it is interesting to note that after fitting this model and plotting the residuals, the residuals still have a quadratic shape as shown in Figure 3. Notice that by adding the 50+ quadratic term to the model and crossing it with Sex as well, the residual values move closer to zero as shown in Figure 4. Yet, the model containing the Age Splines crossed with Sex seems to fit the data best and the 50+ quadratic term may not be adding much predictive power to the model. The residuals are most likely not deviating enough from the center to warrant adding the extra term although the issue should be examined once 2010 CCM and census data becomes available.

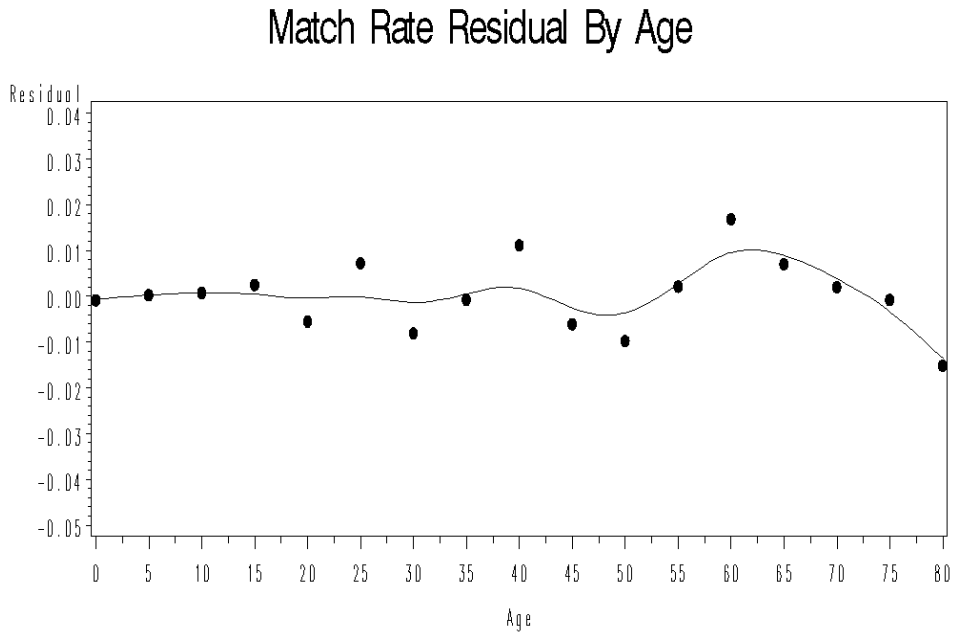

Figure 3. *Age Splines Crossed with Sex*

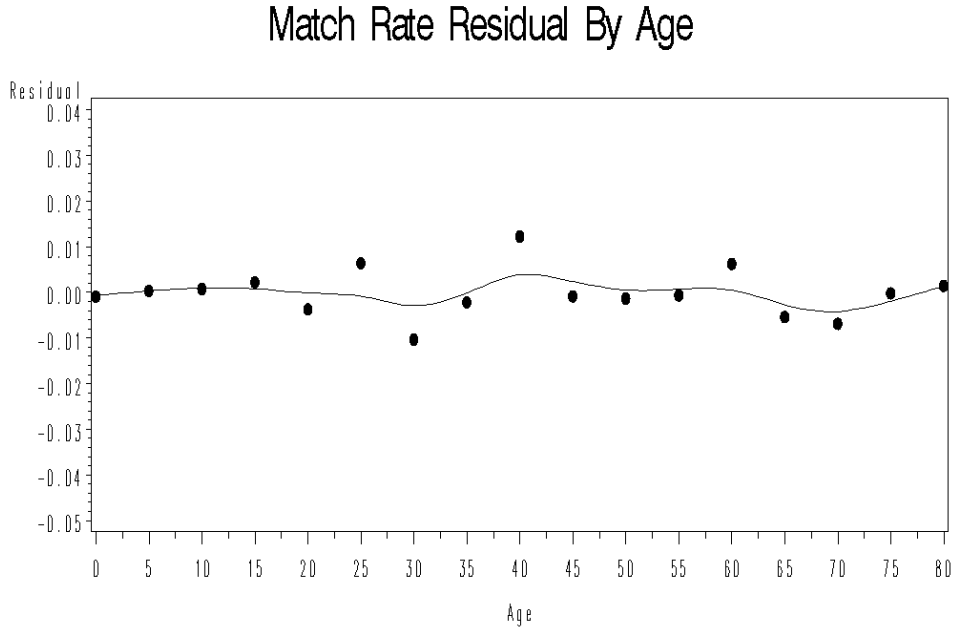

Figure 4. *Age Splines Crossed with Sex including 50+ Term*

Now, the results are shown for the model predicting correct enumeration rate:

E sample

	Wald	LogLik	Cross-validation
No Age	193.3320	-.2464	-.24698
AgeSex	220.1463	-.2461	-.24695
Age Splines	223.7641	-.2460	-.24683
Age Splines Crossed With Sex	230.9296	-.2459	-.24690

These results indicate that the model containing the Age Splines may actually fit the data better than the model where Age Splines are crossed with Sex. Notice that there is very little gain in the log likelihood even when the additional parameters are added, and the cross-validation value goes higher after crossing the Age Splines with Sex. However, to keep the models consistent, the Age Splines crossed with Sex model should be used as was determined by the P-sample model unless 2010 CCM data shows a larger difference between the models. Figure 5 shows the residual plot for the model.

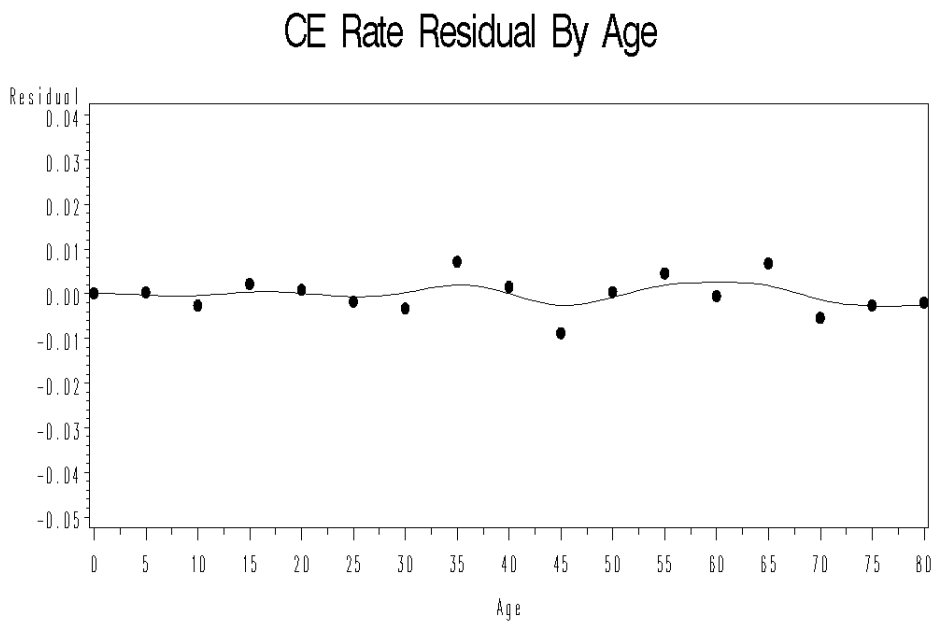

Figure 5. *Age Splines Crossed with Sex*

VARIABLE SELECTION

Two models were run using the possible main effects. As mentioned previously, interactions will be tested once 2010 CCM and census data becomes available. The analysis was again performed using PROC SURVEYLOGISTIC in order to get accurate estimates of the standard errors to test significance of parameter estimates. By accounting for the sample weights and selecting variables in this manner, the fit of the model should be improved over that of a model chosen using PROC LOGISTIC.

E sample

Effect	DF	Wald Chi-Square	Pr > ChiSq
MSA	2	12.8926	0.0016
TENURE	1	12.8958	0.0003
sumratesq	1	127.4800	<.0001
xx00	1	6.5757	0.0103
xx17	1	0.1703	0.6799
xx20	1	0.3131	0.5758
xx50	1	1.3473	0.2458
xx02	1	8.0053	0.0047
xx22	1	0.7881	0.3747
xx17*SEX	1	0.2884	0.5913
xx20*SEX	1	0.2309	0.6308
xx50*SEX	1	1.1264	0.2885
xx22*SEX	1	0.2314	0.6305

P sample

Effect	DF	Wald Chi-Square	Pr > ChiSq
MSAGROUP	2	28.7813	<.0001
TENURE2	1	37.4921	<.0001
sumratesq	1	62.4980	<.0001
xx00	1	5.6953	0.0170
xx17	1	1.0682	0.3013
xx20	1	2.0791	0.1493
xx50	1	1.2679	0.2602
xx02	1	3.5313	0.0602
xx22	1	0.9031	0.3420
xx17*SEX	1	1.4923	0.2219
xx20*SEX	1	0.9718	0.3242
xx50*SEX	1	1.5239	0.2170
xx22*SEX	1	0.3635	0.5466

The results show that MSA, Tenure, and Sumratesq are all significant variables when predicting both match and correct enumeration rate. While some individual Age Splines are not significant, the splines should be interpreted as a single variable. As a comparison, when AgeSex is placed into the model it is determined to be highly significant.

MODEL SELECTION

Once 2010 CCM data is available, interactions can be tested in the models. In addition to determining the significance of the parameters, cross-validations should be performed on the different models as was done earlier in this paper.

POPULATION ESTIMATION

Various estimators can be used to calculate the population of Puerto Rico. One such estimator is called the N2 estimator. While other estimators may provide better estimates and lower variances, this estimator is useful for model comparison and is fairly simple to calculate. The N2 estimator is calculated by taking a ratio of the modeled correct enumeration rate to the modeled match rate for each E-sample case and multiplying the ratio by the weight of the case. The weighted ratios are then summed over all E-sample cases. More information on the N2 estimator and other estimators is found in Griffin (2005).

Variance is calculated using a simple jackknife variance procedure. There is another version of N2 estimator that uses ratio adjustment to lower the variance of the estimates, but the N2 estimator is useful for simple comparison. The data is divided into twenty groups based on the last two digits of the cluster number. After the twenty groups have been created, the variance can be calculated using the following code:

```
%macro n2var;
     %do j=1 %to 20;
       data loopzp;
            set zp; /* P-sample data set */
            if group~=&j then z2=z;
            else z2=.;
       run;
       data loopze;
           set ze;    /* E-sample data set */
           if group~=&j then z2=z;
           else z2=.;
       run;

/* Models the replicate match probabilities */
proc logistic data = loopzp outmodel=matches2 noprint;
  class tenure2 msagroup agesex;
  model z2(event='1')= tenure2 msagroup sumratesq xx00 xx17 xx20 xx50 xx02 xx22;
  weight zwgt;
run;

/* Adds the modeled match probabilities to full E-sample data set */
proc logistic inmodel = matches2;
  score data=ze out=model9;
run;

/* Models the replicate correct enumeration probabilities */
proc logistic data=loopze outmodel=correct2 noprint;
  class tenure2 msagroup agesex;
  model z2(event='1') = tenure2 msagroup sumratesq xx00 xx17 xx20 xx50 xx02 xx22;
  weight zwgt;
run;

/* Adds the modeled correct enumeration probabilities to full E-sample data set */
proc logistic inmodel=correct2;
   score data=model9 out=model8;
run;

/* Calculates an N2 value for each observation in the E sample */
data model7;
  set model8;

N2reps=zwgt*((p_12)/(p_1));  /* p_12 is the CE rate, p_1 is the match rate */

run;

/* Creates a data set holding the replicate N2 estimate */
proc summary data=model7;
  var N2reps;
  output out=estimate&j sum(N2reps)=N2hatreps;
run;

     %end;
%mend n2var;
%n2var
```

After the twenty replicates are calculated, the twenty data sets are combined into one data set using PROC APPEND. Then, the variance is calculated by taking the difference between the full model N2 estimate and each replicate N2 estimate, squaring each difference, multiplying each difference by 19/20, and then taking the sum.

To show how the N2 estimate can be used, a simple example using four models using different versions of the Age variable is examined to see what impact the variable is having on the overall population estimates. All four models use Tenure, MSA, and Sumratesq in the model. Between the largest population estimate and the lowest population estimate there is only a difference of about 2,287 persons out of a total of approximately 3.8 million. The results can

provide an indication as to how much the choice of a form of variable is actually impacting the estimates. In this example, there seems to be little difference between the four variables in the estimate of the overall population.

Model	N2	Var(N2)	SE (N2)
Splines	3822982.52	2275350172.5	47700.63
Splines*Sex	3823055.50	2269840462.9	47642.84
AgeSex	3822288.46	2273858859.8	47685.00
No Age	3820767.89	2263159692.6	47572.68

CONCLUSION

The main goal of this paper was to find the best variables to use for modeling match and correct enumeration rate for persons in housing units in Puerto Rico for 2010 CCM net error estimation. Using logistic regression for net error estimation allows more flexibility when developing a model than has been available in the past. One such example is the ability to use continuous variables to improve model fit as this paper demonstrates. The variables presented in this paper will be proposed to be used as the main effects for models modeling correct enumeration and match rate in the 2010 CCM. Interaction terms for a 2010 CCM model will be explored once 2010 CCM data is available using the techniques covered throughout this paper. The best possible model can be developed using PROC SURVEYLOGISTIC along with cross-validation and various exploratory techniques. Finally, estimates can be calculated using different models to see if any large changes occur in the estimates. Based on the results from this paper, all of the covariates are quite strong predictors and form a strong set of main effect variables.

REFERENCES

Griffin, Richard (2005) “Net Error Estimation for the 2010 Census,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-01, U.S. Census Bureau, April 18, 2005.

Haines, Dawn (2001) “Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Dual System Estimation (Puerto Rico) - Re-Issue of Q-31,” DSSD Census 2000 Procedures and Operations Memorandum Series Q-39, U.S. Census Bureau, March 7, 2001.

Mule, Thomas, Schellhamer, Malec, Maples, and Griffin (2007) “Using Continuous Variables as Modeling Covariates for Net Coverage Estimation,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-09, U.S. Census Bureau, February 8, 2007.

ACKNOWLEDGMENTS

I would like to thank Douglas Olson for his contributions to this paper.

CONTACT INFORMATION

The author appreciates questions and comments. Contact the author at:

Colt Viehdorfer
U.S. Census Bureau
4600 Silver Hill Road
Washington, DC 20233
(301) 763-6796
Colt.S.Viehdorfer@census.gov