Using Telephone Exchange-Level Data To Adjust For Non-Response: Application In An Establishment Survey

Stephen Williams and Ronghua (Cathy) Lu
Mathematica Policy Research, Inc., P.O. Box 2393, Princeton, NJ 08543-2393

Abstract

Telephone surveys are being challenged by increasing usage of answering machines, call waiting, and other screening and call-blocking services. These popular innovations are significantly influencing how telephone surveys are conducted. For the national establishment survey described in this paper, for example, resolving by telephone whether or not a sample name and address is a potential eligible establishment and, if so, whether or not it is actually in the target population were major challenges. In general, resolving whether or not a telephone number is linked to a potential eligible, determining eligibility for the study, and obtaining participation in a survey is becoming more difficult. Hence, not only are survey methods being revised, but also more effective methods are being sought to adjust for the lower rates of successfully screened numbers and completed responses. The application of one method with application to an establishment survey is the topic of this paper. We conclude that the use of summary information about population demographics at the telephone exchange level can improve the accuracy of survey results.

Key Words: Nonresponse, Weighting, Propensity Modeling, Establishment Survey, Telephone Survey

1. Introduction

This paper deals with statistical analysis of survey data obtained through telephone interviews with post secondary educational establishments (*the study*). The sample establishments were selected through a complex multi-stage design (*the survey*). Such analyses must reflect the sampling design and methodology in order to obtain valid results. In this particular setting, the consideration of the sampling design and survey implementation are particularly key for obtaining an unbiased characterization of the target population. Hence, the details of the survey objectives, definitions, design, and implementation are needed to understand the problems facing the analyses of the survey data. The development of the sampling frame and the response and resolution of sample units are noteworthy subtopics described below. The objective of the paper is to describe the methodology pursued for resolving the problem of sample unit attrition primarily through nonresolution of telephone numbers and inability to determine if the unit is a member of the target population.

2. Background

A viable sampling frame did not exist for the survey and was a challenge to create. As a result, the frame was somewhat outdated and had issues regarding quality of information about the units on the frame. For the sample selected from the frame, many of the telephone numbers could not be resolved as to whether or not the number was linked to an establishment belonging to the study population. The potential bias to survey results from this sample attrition was a serious concern. An effective method was needed to adjust the sampling weights for this sample attrition.

In similar situations involving a frame with minimal reliable information we have used summary socio-demographic information about the household population to model response propensities. We use Genesys Sampling Systems to generate random digit dialing (RDD) numbers for many of our telephone surveys at Mathematica Policy Research, Inc. (MPR) Hundred-blocks (i.e., 100 numbers identified by the first eight digits of a ten-digit telephone number: NPA-NXX-00xx) containing one or more listed telephone numbers are used to generate candidate telephone numbers; a list assisted method (Potter, et al. 1991). Three issues are typically of concern in telephone surveys: (1) some numbers cannot be resolved as to whether or not they are assigned to a potentially study-eligible unit; (2) for those identified as potential eligibles, the screening cannot be completed for some; (3) and for those screened to be eligible, some do not complete the interview.

In some of the RDD surveys, we also use another service provided by Genesys. They maintain a database that summarizes household population characteristics summarized at the telephone exchange level (the NXX level in the above characterization of a telephone number). This information is used as covariates in logistic regression models to predict the probabilities associated with each level of sample attrition. These probabilities (*propensities*) are then used to adjust the sampling weights to reduce bias related to the attrition. Individual unit information is usually more effective for this purpose than aggregated data, but when little is known about the individual units this is often a useful alternative. The effectiveness of using data at the exchange level is influenced by the fact that an exchange contains

10,000 potential telephone numbers, representing a relatively large, diverse, and non-contiguous area. This and the fact the data relate to the household population rather than directly to the educational establishments in the target population of the survey, meant this solution was somewhat exploratory. The hypothesis was that the characteristics of such establishments would be correlated to their environment, that is, the adjacent household characteristics.

Response propensities estimated with logistic regression models have been widely used in a number of major survey settings to adjust for sample attrition. These models have utilized explanatory information (covariates) from various sources, including sampling frame information, small-area statistics from the Census, the Area Resource File (county-level information from numerous sources), information from previous rounds in longitudinal surveys, and information about telephone-based areas such as telephone exchange and telephone "hundred blocks". The latter is the topic investigated in this paper. Further, the application of propensity models ranges from their use to identify weighting classes, which are then used to adjust sampling weights for nonresponse, to using the model predictions themselves to directly adjust the weights of respondents. Here we consider using the model predictions themselves to adjust sampling weights.

Related Applications

A literature review of the applications of propensity modeling or methods to adjust for nonresponse is not presented here. But we mention some directly related studies and one recent example of an application of propensity models to form weighting class cells.

Response Models In Rdd Surveys: Utilizing Genesys Telephone-Exchange Data

This paper (Williams et al. 2004) presented results from the third round of the Community Tracking Study (CTS) Household Surveys conducted by MPR. The CTS is a periodic national study of the rapidly changing health care market and the effects of these changes on people. Funded by the Robert Wood Johnson Foundation, the study is conducted by the Center for Studying Health System Change (**www.hschange.com.**) That paper investigated the feasibility of using logistic regression models to predict for each sampled RDD number, the likelihood of resolving the number and the likelihood, if the number is a household telephone number, that the household will complete the questionnaire. The RDD numbers in this example are generated through the Genesys system, which, as noted above, contains a substantial amount of auxiliary demographic information about households at the telephone-exchange level. We used these auxiliary data as covariates in "propensity" models to adjust the sampling weights for nonresponse. The methodology used there is basically the same as that used in the current paper except that earlier study was based on an RDD household survey rather than a list-frame establishment survey. This CTS survey was the first of this type that revealed a problem with using weighted solutions for the logistic regression. As in that case, the present application to the educational establishment population also utilized unweighted solution to the models--discussed below.

The results of our analysis in that study indicated that aggregate demographics from the Genesys file showed promise for adjusting RDD sampling weights. The significant variables in the resolvability model were education, home ownership, income, race and age; demographics such as education, income and telephone listings per "working hundred block" were important in the screening model; while income, age, race and number of listings in working hundred blocks were used in the response model

Resolvability, Screening, And Response Models In RDD Surveys: Utilizing Genesys Telephone-Exchange Data (Lu, et al. 2002)

This paper also investigates the feasibility of using logistic regression models to predict for each sampled RDD number in a household survey setting, the likelihood of resolving the number and the likelihood, if the number is a household telephone number, that the household will complete the screening questions, and finally that the household will complete the questionnaire. This was a survey, conducted by MPR, of New Jersey residents. A total of nearly 36,000 RDD numbers were assigned to telephone interviewers, but about 28 percent could not be resolved. This results in uncertainty about the number of eligible households in the study population and important inference domains. Further, about 35 percent of identified households did not complete a screener and 20 percent of the screened households did not complete an interview. These results, while not unusual for telephone surveys, are potential sources of survey bias.

The RDD numbers as well as the exchange-level information used in this example were also from the Genesys system. The results of our analysis in that study, similarly to the previous study, indicated that aggregate demographics from the Genesys file showed promise for adjusting RDD sampling weights.

A Comparison Of Two Methods To Adjust Weights For Non-Response: Propensity Modeling And Weighting Class Adjustments (Carlson and Williams 2001)

A common method used to adjust sampling weights for non-response involves forming weighting classes of homogeneous sample members. Within each cell, the weights of the respondents are inflated to account for the non-respondents. A relatively new approach involves developing logistic regression models to predict response, using a potentially much broader set of predictive variables than can be used in the weighting class methodology. We applied both of these methods when computing weights for round two of the Community Tracking Study (CTS) Household Survey.

This earlier paper explored the differences resulting from these two methods. Very little difference was found between the propensity method and the weighting cell method when looking at the CTS household survey RDD re-interview sample (a substantial portion of the sample consists of households interviewed in the prior round of this longitudinal survey). The expected benefits of the propensity modeling were not seen. This is likely due to two main reasons: (1) the number of weighting cells here was so large (over 300) that the weighting cell approach nearly approximated the response surface nature of the propensity modeling approach, and (2) the screener and interview response rates among the household survey re-interview portion of the sample was high to start with, allowing for very little variation in the non-response adjustments.

Compensating for Provider Nonresponse Using Response Propensities to Form Adjustment Cells: the National Immunization Survey

This application is an example of using propensity models to form weighting classes (DHHS 2001). The use of logistic propensity models simply to identify variables for use in weighting classes is common, this particular application presents a slightly different application in that the estimated propensities are themselves used to form weighting classes, not just to identify variables for forming classes. A national model for predicting the probability that a sampled child has provider-reported vaccination history is used (basically this is a predictor of provider nonresponse based on household characteristics). The individual records are sorted according to value of predicted probabilities within each of 78 geographic-based strata. From 1 to 3 weighting classes were formed within each stratum, based on the frequencies of these predicted values in the stratum. The weights for respondents within each of these classes were then adjusted to the total sample weights in the class.

3. The Establishment Survey

Survey Objectives

The primary objective of this establishment survey is to estimate the number and characteristics of eligible institutions in the US and Puerto Rico. Eligible institutions are non-Title IV post-secondary educational establishments that provide a formal instructional program whose curriculum is designed primarily for students who are beyond the compulsory age for high school. The eligible establishments include programs whose purpose is academic, vocational, and continuing professional education, and exclude avocational (leisure) and adult basic education programs.

Institutions that have a Program Participation Agreement (PPA) with the Department of Education to disseminate Title IV funding are excluded from the target population. The majority of these responding institutions are required by law to respond because of their participation in, or applications for participation in, Federal Financial Assistance Programs authorized by Title IV of the Higher Education Act of 1965 (as amended in 1998). Postsecondary institutions that do not have a PPA (referred to as non-Title IV institutions) are largely unknown to the Department of Education, and thus were the focus of this data collection effort. Since little was known about the characteristics and number of the survey-eligible establishments, no comprehensive list existed for use as a sampling frame.

Logistics

The development, conduct, and analyses transcended several years and several contract research organizations. MPR's role was to prepare the data files for analysis, develop the analysis weights, and produce the basic summary statistics. Because of the complexity of developing the sampling frame for these institutions and the time required for this work, the frame was relatively outdated when the sample was selected and fielded. Hence, frame quality resulted in complications for the response rates and the calculation of analysis weights. Under such circumstances, the methods used to adjust for unresolved sample units is key for obtaining valid survey results.

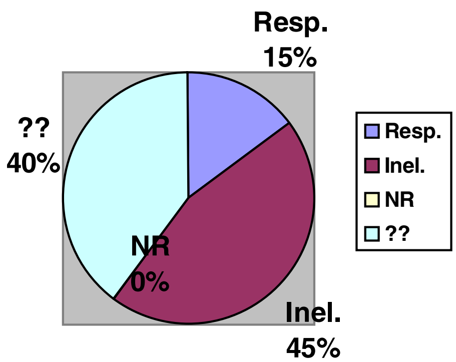
Frame Construction

As mentioned above, no list of the target establishments existed. To create a sampling frame, therefore, the first logical step was to construct a complete frame based on geographic areas. A total of 3,208 such primary sampling units (PSUs) were formed for the first stage frame. Those were stratified and a probability sample of 791 were selected. Multiple sources were used to create a list of potential eligible establishments for each of the sample PSUs. Numerous sources were used but the two major ones were commercial lists of establishments and involved the use of both key word and Standard Industrial Code searches. Names that were obviously not potentially eligible and multiple listings were removed. This process was extremely demanding and involved both automated and manual processes. Frame quality was bound to be a concern.

Survey Outcome

A telephone survey of the approximately 13,000 sample candidate establishments was conducted. Not surprisingly, considering the frame construction process and the time required to field the sample, the outcome required special attention to the calculation of analyses weights and analyses of survey responses.

Figure 1. Survey Outcomes



Approximately 45 percent of the sample establishments were not eligible, 15 percent were eligible respondents, 0.2 percent were eligible nonrespondents. The unusually small number of nonrespondents was attributed to the complexity and length of the screening questions—once this part was completed the interview was nearly complete. Finally, nearly 40 percent of the sample cases were unresolved as to whether or not they were eligible. These are the cases that required special attention in the calculation of analysis weights and the main focus of this paper.

Of the 40 percent that were unresolved, most (32%) were field coded as wrong number or not a working number, we refer to these as ***unlocated***. Often this classification in telephone surveys is final-coded as ineligible, but because of the large number in this group and because we had learned that many were related to incorrect frame information this was not a reasonable solution (for example, as the frame aged many of the telephone exchange numbers changed). The remaining 8 percent were incomplete screening, ring-no-answer, busy, or some type of automated answering machine—we refer to this group as ***unscreened.***

Figure 1, above, shows the response experience of the telephone numbers and sample units released to the telephone interviewers. To adjust for the approximately 40 percent of unresolved cases we investigated the feasibility of using Genesys information to develop propensity models for resolvability of the telephone numbers for the sample units.

Genesys

The information available to us in the Genesys sample files includes demographic characteristics of the household population at the telephone exchange level, such as age, income, race, education, and home ownership. Each exchange contains 10,000 potential telephone numbers. Many of these numbers are not assigned to households and the area covered by an exchange is large, diverse, and non-dense. This type of information is not ideal for developing propensity models, but is worth considering because of its availability and the paucity of reliable information about individual establishments. We initially explored the merit of using such information to adjust for missing information on several surveys described above. In this paper, we investigate the feasibility of the method for an establishment survey.

We now investigate the feasibility of using auxiliary demographic information available on the Genesys files to predict telephone number “location” and “screening” completion, the former being by far the major source of sample attrition in the establishment survey. The results of the survey and the Genesys data were the basis for constructing these two models. In addition to the Genesys data, geographic stratification and other frame information, although minimal, also were used as candidate independent variables (covariates). We hoped to use these auxiliary data in “propensity” models to adjust the sampling weights, in order to reduce the potential for bias that can result from the missing information.

3. Study Methodology

Propensity modeling -- We use propensity models, which are an increasingly popular method for adjusting for non-response: that is, creating a logistic regression model that predicts the likelihood of response versus non-response. Two propensity models were considered. First, a location model for all sample telephone numbers, to predict the likelihood of the number being a valid telephone number, and (2), a screening model for the likelihood that we are able to determine eligibility given an eligible number. A separate model to predict the likelihood whether or not the establishment will complete the interview was not needed because of the very small number of nonresponding known eligibles.

Constructing covariates for the propensity models The variables available for each sample case from the Genesys files were a few dozen demographic characteristics at the telephone exchange level in which the telephone number resided, and one variable at the working bank level which was household listings per working hundred block. They are mostly continuous variables, such as the median household income in the exchange area. Continuous variables in the models, however, use only the linear relationship in a model (unless higher-order terms are used). We therefore transformed most of the continuous variables into a set of binary variables in order to capture nonlinearity.

After creating binary variables for most of the continuous variables, we had approximately 70 candidate covariates for the models. Interactions were also investigated. Two sets of models were run: 1) using the dichotomy of whether or not the number is located (a valid telephone number) and 2) the dichotomy of establishments with valid numbers as to whether or not they were screened successfully. The chi-square p-values for the regression coefficients were used to reduce the number of variables and to identify a manageable number of candidates for interaction variables.

Using stepwise logistic regression in SAS -- Once the initial candidate variables for the two models were identified, we used stepwise unweighted logistic regression (SAS Logistic) to reduce the list of variables in the two models before we considered possible interactions. Both forward and backward solutions for stepwise logistic regression in SAS were used, setting the significance level at 0.3. Somewhat different sets of variables are identified in the two different solutions. Part of the problem with forward solution is that once a variable is admitted, it is retained even though it becomes insignificant as a result of other variables or interactions entered into the model.

In a previous paper we found that the use of weighted (using sampling weights) logistic regressions were problematic (Williams and Lu 2004). This was obvious from the result of the Hosmer-Lemeshow (HL) statistic, which evaluates model fit. The HL statistic had a p-value of approximately 0.001, which indicates a terrible fit. The predicted and actual outcomes of groups used to obtain the HL statistic are summarized for approximately ten groups. The predicted values in this previous survey were all larger than the observed for the number of unresolved cases. This indicates an upward bias, but our models were based on asymptotically unbiased maximum likelihood estimators.

Since we had used sampling weights to estimate the model coefficients, the results were unbiased for the population. But our sample was not proportional because of oversampling of prior round respondents. Hence, what we really needed were models that are effective for estimating the *conditional* probabilities of the sequential stages of attrition.

To explain, we note that to obtain an interview for a given establishment, three events must occur:

A. the establishment is selected in the sample,
B. the telephone number is a valid number, and finally,
C. the screening interview is completed.

The joint probability of these events is (unless we assume independence—bad assumption when we oversampled prior round respondents in this case):
P(ABC)=P(A)*P(B|A)*P(C|AB).
P(A) is known, determined by the design, but we must estimate the two conditional probabilities. Hence, we need a model that produces valid estimates for our particular sample, not for the population. Unweighted regression analysis produces such a model for resolvability models.

Subsequently, we learned of a corroborating simulation study by Little and Vartivarian (2003). Although the main focus of their study was on weighting class adjustments for nonresponse, the results generalize to propensity modeling, which is itself an extension of the weighting class method. They found that, although weighted adjustments produced an unbiased estimate of the population response rates this does not ensure unbiased estimates of the variables of interest. In particular, they assert that the correct approach is to model nonresponse based on design and other variables and to use the inverse probability estimated from the model as the response weight. And they finally conclude that using weighted response analyses is either incorrect or unnecessary depending on the relationships between the variables of interest, design factors, and response rates.

Initially, the resolve and response models were run with a significance level of 0.3. After identifying the interaction factors, the significance level for the final model was reduced to 0.1. At that level, the stepwise logistic regressions reduced the list of variables (including interaction terms) to 26 for the location model and 16 for the screening model. Some main effects are included, even though not significant, because an interaction with the main effect is significant at that level. Table 2 presents the variables in these models. The variables most important to the location model included age, income, population density, and geographic region. The variables most important to the screening model included the education, income, county size, and geographic region.

Running the final models in SUDAAN – The final stage in some of the past application has been to enter the reduced sets of variables into the weighted logistic procedure in SUDAAN, and calculate the final parameter estimates using the design features (primarily stratification and disproportionate sampling). We did not do this final step for the same reason we did not used weighted solution for the models, but expect little change in the conclusions had this been completed in this case. The more serious problems with weighted solutions occur when the sampling weights and the attrition rates are correlated, as when oversampling is used in groups that have consistently higher or lower attrition. While the final models were not run in SUDAAN in this case it was used in the final statistical analyses for the survey.

3. Study Results

Table 2 shows the variables that were retained in the two comprehensive models for the sample, using a significance cutoff of 0.1. Fourteen main effects and 12 interactions were retained in the location model and 12 main effects and 4 interactions were retained in the screening model. Table 1 shows some statistics we used to test the goodness of fit and predictive power of the models. The first one tests the global null hypotheses (Likelihood ratio statistics) that all the explanatory variables have coefficients of zero. We can see from the table that P-values are very small for both models in both models, so we reject the null hypothesis and conclude that at least one of the coefficients for explanatory variables in each model is not zero. Note that the degrees of freedom in these tests are equal to the number of covariates retained. Note also that both models have used a significance level cutoff of 0.1.

The next step tests model goodness of fit, using two statistics: Hosmer-Lemeshow (HL) test and Deviance. HL test is based on grouping predicted probabilities into approximately ten cells. Larger p-values for the HL statistic indicate better fit (we don't want to reject the null). We can see from the table that both models pass this test. The deviance compares the fitted model with a saturated model, and again, larger p-values indicate better fit. The models are also good fits according to the deviance test, although the location model appears weakest according to this test. The deviance statistic can be unreliable if there are many or continuous explanatory variables in the model, because too many covariate profiles allow small cell counts, this may be a factor because the location model has 1,139 unique profiles (only 291 for the screening model). (Allison 1999).

The last statistics describe how well we can predict the dependent variable based on the values of the independent variables. The model concordance shows that the models are effective. The spread between concordance and discordance is about 20 points or more, indicating relatively strong models. Three measures of association: Sommer's D, Gamma and Tau-a, can all range from zero to one, with larger values indicating better association between the predicted and observed values. These statistics are all based on the concordance/discordance numbers. Tau-a tends to be closest to the generalized R-square, which is a measure that is of limited use in the logistic setting and has much different meaning than the R-square we are familiar with in linear regression with continuous dependent variables. We see the association measures are relatively low. For predictive power, the screening model appears to be slightly better than the location model.

4. Conclusions and Limitations

The results indicate that aggregate demographics from the Genesys file are useful for adjusting sampling weights. Although the test statistics vary, the location and screening models appear very comparable. All models, as expected, are weaker than experienced for propensity models based on un-aggregated data. The significant variables in the model were income-related, population age, geographic region, and population density.

REFERENCES

Allison, P. (1999) Logistic Regression Using the SAS System: Theory and Applications, SAS Institute, Inc., Cary, NC

Carlson, B. and S. Williams (2001) "A Comparison Of Two Methods To Adjust Weights For Non-Response: Propensity Modeling And Weighting Class Adjustments" Proceedings of the American Statistical Association, Alexandria, VA

Hosmer, D.W. and Lemeshow, S. (2000) Applied Logistic Regression 2nd Edition, John Wiley & Sons: New York

Little, R.J.A., and S. Vartivarian. (2003) "On Weighting the Rates in Non-Response Weights." *Statistics in Medicine*, vol. 22, pp. 1589-1599.

Lu, R.., J. Hall, and S. Williams (2002). "Resolvability, Screening And Response Models In Rdd Surveys:…". Proceedings of the American Statistical Association, Alexandria, VA

Potter, F.J., McNeill, J.J., Williams, S.R. and Waitman, M.A., (1991) "List-Assisted RDD Telephone Surveys", Proceedings of the American Statistical Association, Alexandria, VA. pp 117-122.

Marketing Systems Group Genesys Sampling Systems Methodology, (2000) Philadelphia, P.A. MSG

Williams, S., R. Lu, and J. Hall (2004). "Response Models In Rdd Surveys: Utilizing Genesys Telephone-Exchange Data Proceedings of the American Statistical Association, Alexandria, VA

Table 1. Statistics for Effectiveness of Location and Screening Models

	Locate	Screen
Global Chi-Square (Likelihood Ratio)		
χ-Square	**447**	**91**
Degrees Freedom	**26**	**16**
P-Value	**<0.001**	**<0.001**
Goodness of Fit (P-values; χ2 tests)		
H-L Test	**0.93**	**0.93**
Pearson	**0.20**	**0.78**
Predictive Power Measures		
Concordance	**60.0%**	**57.1%**
Discordance.	**38.2%**	**38.1%**
Somer's D	**0.22**	**0.19**
Gamma	**0.22**	**0.20**
Tau-a	**0.09**	**0.02**

Table 2. Model Covariates

Location Model		
Variable	**Coeffic.**	**P value**
Intercept	0.935	<.0001
ageone3:high % in age 0-17	-0.256	<.001
ageSIX3:high % in age 55-64	-0.247	0.022
ageSEVN1:low % in age 65+	-0.114	0.038
INCONE3: high % in <$10K	0.375	0.041
INCFIV1: low % in $35-49.9K	-0.327	<.0001
INCFIV3: high % in $35-50K	-0.129	0.073
MEDHMV3= high home value	0.410	<.0001
TTLPOP1: small population	-0.074	0.162
TTLPOP3: large population	0.108	0.146
WHTPNT3: high % white	-0.035	0.625
ncntysiz1: large metro county	0.081	0.194
REGION1	-0.512	<.0001
STATUS2:suburb of MSA	-0.068	0.181
STATUS5: not an MSA	2.673	<.0001
Interactions	0.206	0.038
ageone3*ncntysiz1	-0.239	0.281
ageSIX3*HHLST1	0.394	<.0001
HHLST1*REGION3	-0.125	0.246
WHTPNT3*EDUYR3	0.275	0.001
ncntysiz1*EDUYR3	0.374	0.093
INCONE3*ncntysiz1	-0.380	0.085
INCONE3*REGION3	-0.294	0.026
MEDHMV3*REGION1	-1.541	0.001
MEDHMV3*STATUS5	0.660	<.0001
WHTPNT3*REGION1	0.288	0.005
WHTPNT3*REGION3	-1.056	0.013
ncntysiz1*STATUS5	0.935	<.0001

Screening Model		
Variable	**Coeffic.**	**P value**
Intercept	2.381	<.0001
ageone1: low % in age 0-17	-0.254	0.049
EDUYR1: low education	-0.345	0.015
INCTWO3: high % in $10-15K	0.384	0.017
INCTHR1: low % in $15-25K	0.206	0.160
INCFOR1: low % in $25-35K	-0.384	0.009
MEDRNT1: low median rent	0.236	0.047
RENTPNT1: low % rentals	0.407	0.025
ncntysiz1: large metro county	0.160	0.068
ncntysiz3: medium metro cty	0.380	0.008
REGION2	0.277	0.003
REGION3	0.407	<.0001
STATUS5: not an MSA	0.508	0.029
Interactions		0.018
ageone1*EDUYR1	2.326	0.028
EDUYR1*INCFOR1	-2.060	0.020
ncntysiz3*STATUS5	-1.171	0.099
RENTPNT1*REGION3	-0.461	0.035