Alternative Methods to Adjust for Non-response in the American Community Survey (ACS)

Robyn Sirkis, US Census Bureau
Washington, D.C. 20233-8700
robyn.b.sirkis@census.gov

Abstract

This paper discusses research on alternatives to the ACS weighting methodology to adjust for non-response, which remains unchanged since 1996. The study universe is the 2006 ACS tabulation sample. Methods using the inverse of the propensity score as the adjustment factor, the mean inverse of the propensity score as the adjustment factor, and forming the adjustment cells using the propensity scores were investigated. The methods were tested on a dataset where a sample of respondents were recoded to non-respondents and compared to a dataset that only contained respondents by analyzing several population, household and housing unit estimates. The variances were analyzed in order to ensure that there is not an increase due to the alternative methodologies.

Keywords: American Community Survey, Weighting, Estimation, Non-Response, Propensity Scores

1. Introduction

The objective of this research project was to find alternative methods to adjust for non-response. Research has been conducted on alternative non-interview adjustments on the ACS (Weidman, 2006). However, there have been additional non-interview adjustment techniques explored in the past that use logistic regression. Three of these techniques include using the inverse of the propensity score as the adjustment factor, using the mean inverse of the propensity score as the adjustment factor, and using the propensity score to form weighting adjustment cells (Aertker and Kalsbeek, 2006). The current methodology consists of several ratio estimation steps and we wanted to compare it to propensity score methods that use logistic regression models.

To test the alternative methods, a truth deck dataset was created by recoding a sample of the original respondents as non-respondents. This was accomplished by using a systematic sampling procedure. The weighting process was performed for each of these alternative methods to produce several population, household and housing unit estimates.

Then the weighting process was performed using only the respondents from the 2006 ACS production data. This produced the truth estimates. The truth estimates were compared to the estimates from each of the alternative methods that used the truth deck dataset. These alternative estimates were evaluated using the mean algebraic percent error and mean absolute percent error.

2. Background

There are two non-interview adjustment factors in the current production methodology, Non-Interview Adjustment Factor #1 (NIF1) and Non-Interview Adjustment Factor #2 (NIF2). There are also the Mode Non-Interview Adjustment Factor (NIFM) and Mode Bias Factor (MBF) which are included in the non-interview adjustment process. This process is implemented by weighting area, which is either a county or group of less populous counties. The background section discusses the ACS non-interview adjustment process. For more information on the entire ACS weighting procedure see Census, 2006.

Characteristics that have been shown in other surveys or censuses to be related to housing unit response can be used to adjust for non-response. Since nothing new is learned from the housing unit or person characteristics of the non-respondents during data collection, only characteristics known at the time of sampling can be used in adjusting for them. The characteristics currently used are census tract, building type (single unit and multi-units) and month of data collection. However, there is not a sufficient number of sample housing units in each cell of a three-way cross classification table formed by these variables for each weighting area to produce a reliable estimate of its non-response proportion. The reason is that tracts typically have small annual sample sizes and the sample is spread evenly across all twelve months of the year. As a result, there are two non-interview adjustment factors that each use two of the three characteristics for each weighting area. The vacant units and ineligible units are excluded from the non-interview adjustment process.

---

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

In NIF1, the housing units are placed into cells based on the cross classification of building type and census tract. In NIF2, the housing units are placed into cells based on the cross classification of building type and tabulation month. If a cell contains fewer than 10 interviewed housing units, it is collapsed with an adjoining cell until the minimum size of 10 interviewed housing units is achieved. A ratio adjustment of the number of interviews divided by the number of interviewed and non-interviewed housing units is performed in each final cell.

The housing unit addresses assigned to a given month form a panel. The data for each address can be collected via 3 modes over a 3-month period. In the mail mode, a questionnaire is mailed to the sample address for the household to complete and return by mail. Computer Assisted Telephone Interviewing (CATI) occurs in the second month when telephone interviewers attempt to contact the occupants of non-responding addresses and complete the ACS questionnaire by telephone. Computer Assisted Personal Interviewing (CAPI) involves taking a sub-sample of the remaining non-responding addresses and making a personal visit in the third month to households to complete the ACS questionnaire.

An assumption is that the non-responding housing units have characteristics more similar to the housing units that are conducted by CAPI than to the housing units that are conducted by CATI or that respond by mail. The purpose of MBF is to adjust for this difference while minimizing the impact on the variances. The first step in the calculation of MBF is to calculate an intermediate factor referred to as NIFM, which is not used directly as a weighting factor. The cross-classification cells are defined within weighting area by building type and tabulation month. The collapsing procedure is the same as NIF2 except that NIFM is applied only to CAPI interviews. In MBF, the housing units are placed into cells based on the cross classification of tenure (owned, rented, or temporarily-occupied), tabulation month, and marital status of the householder (married/widowed or single). If a cell has fewer than 10 interviewed housing units it is collapsed across marital status or tenure until it contains ten interviewed housing units. A ratio adjustment factor of the number of interviewed and non-interviewed housing units divided by the number of interviewed housing units is performed in each final cell.

3. Methodology

There are four non-interview adjustment methods that are discussed in this paper. The methods include using the inverse of the propensity score as the adjustment factor, using the mean inverse of the propensity score as the adjustment factor, formation of cells using the propensity scores, and the production method.

The logistic regression models were first estimated using the 2006 ACS production data with the original respondents and non-respondents. The purpose was to find a model that fits the original data because this is what would be used in production. Once the model fit the 2006 ACS production data well the model was estimated using the truth deck dataset.

3.1 Truth Deck Dataset

The input dataset to the non-interview weighting process consists of the housing units whose weights have been adjusted for their initial probability of selection including CAPI subsampling.

A truth deck dataset, which only contained respondent records, was created to test the alternative methods. The respondent records are those that are considered occupied or temporarily-occupied vacant interviews. A sample of records was recoded to non-respondent records by using a systematic sampling procedure. This was conducted for each weighting area, mode, and an ID that captures sample month and geography. The truth deck data was constructed so that the non-respondents would have the same distribution by mode and geography as the original non-respondents since the non-interviews that are conducted by CAPI have similar characteristics than the housing units conducted by CATI or those that respond by mail. Also taken into account was the differential response by geography.

3.2 Logistic Regression Models

Logistic regression models were produced to obtain the propensity scores, which were calculated using the following equation.

$$\text{Propensity Score} = \frac{\text{EXP(LOGIT)}}{(1 + \text{EXP(LOGIT)})}$$

The logistic regression models were estimated at several levels of geographies: weighting area, state, division, region, and urban/rural area, in order of preference. The logistic regression models were estimated using the weights and without using the weights.

The response variable was the housing unit status, which was either a respondent or a non-respondent. The respondent records were the occupied and temporarily-occupied vacant interviews. Cases not included in the model were vacants, ineligible units, and CAPI sub-sampled out cases (cases excluded from personal interviewing).

The variables that were selected as potential predictors were building type, tabulation date, metropolitan status, mode, urban/rural area. The additional variables related to telephone interviewing were the number of fax messages, number of hang-ups, number of refusals, number of times record was updated, number of times message indicating phone number not correct, number of times temporarily out of service message, number of unproductive calls, and the total number of calls.

A date variable was created and then separated into binary variables. The date variable counts the number of days between the date the outcome was determined for either the first mailing, second mailing, Telephone Questionnaire Assistance (TQA), or CATI, and the date it was first mailed. The starting date varies by month partly due to weekends or holidays and usually is the first of the month. The starting date is the same for the first mailing, second mailing, TQA, and CATI.

For CAPI, the date variable counts the number of days between the date the CAPI outcome was determined and the date CAPI began. There are several procedures for deciding whether the first mailing, second mailing, TQA, CATI, or CAPI date should be used. The date variable was categorized into several binary variables for different time periods determined by the distributions of the date variable. The categories are uniform in size. Several models were tested using a different number of date variables.

The tabulation date, which consisted of twelve months, was separated into twelve binary variables. The purpose was to determine if there were some months where there were more differences between the interviews and non-interviews. The mode variable was separated into three binary variables. This consisted of mail, CATI, and CAPI records.

There were two types of models: the main effects model, and the main effects with first-order interactions model. The stepwise logistic regression method was utilized which adds and removes variables from the model in such a way that each forward selection step may be followed by one or more backward elimination steps. The α value was set at 0.1 for the model, since the Census Bureau uses 90 percent level as the acceptable reliability. Significant variables identified using stepwise regression models were ultimately used in the final model.

In order to determine if the models were a good fit, certain statistics produced from the model were analyzed. These included the R-Squared statistic, Hosmer and Lemeshow goodness-of-fit test, and the Pearson Deviance Statistic. The R-Squared statistic, which refers to the fraction of variance explained by a model, gives some information about the goodness of fit of a model. The Hosmer and Lemeshow goodness-of-fit test was used to determine if the model fit the data well. This goodness-of-fit test is for the case of a binary response model. A small p-value (< 0.05) suggests that the fitted model is not an adequate model. An overall goodness-of-fit test called the Pearson or Deviance Statistic was also used to confirm if the model fits the data well. Small p-values lead us to reject the current model. If there was quasicomplete separation then the model was discarded. Quasicomplete separation is a numerical problem, which means there is little overlap in the distribution of the covariates between the two outcome groups.

3.3 Method Using Inverse of Propensity Score

The first method involved using the inverse of the propensity score as the non-interview adjustment factor called NIF. NIF would replace NIF1 and NIF2. NIFM and MBF will also be replaced if NIF better captures the characteristics of non-respondents. This applies to all of the alternative non-interview adjustment methods that use the propensity scores.

The occupied and temporarily occupied interviewed housing units were adjusted by the non-interview adjustment factor. The vacant and ineligible housing units get a factor of 1.0, and non-interviews get a factor of 0.0. This applies to all of the alternative non-interview adjustment methods.

3.4 Method Using Mean Inverse of Propensity Score

This method involves using the mean inverse of the propensity score as the adjustment factor. The mean of the propensity scores was calculated by weighting area and building type. There are 2005 estimation strata and two building types (single-unit and multi-unit). The mean of the propensity score was based on the respondent records (occupied and temporarily-occupied vacant interviews).

3.5 Method Using Propensity Score to Form Weighting Adjustment Cells

This method involves placing housing units into weighting adjustment cells based on the propensity scores to adjust for non-response. For each weighting area the respondent and non-respondent records were placed into adjustment cells based on the propensity scores. If a cell contains fewer than 10 interviewed housing units, it is collapsed with an adjoining cell until the collapsed cell meets the minimum size of 10 interviewed housing units. If a cell contains more than 10 interviewed housing units, all with the same propensity score, it is not collapsed with another cell. A ratio adjustment factor of the interviewed and non-interviewed housing units over the number of interviewed housing units is performed in each final cell.

3.6 Production Method

The current 2006 ACS production non-response adjustment method was also applied to the input truth deck dataset. The purpose was to compare it with the alternative methods.

3.7 Method for Constructing the Truth Estimates

The weighting process was performed using the 2006 ACS production method with only the respondent records. This produces the truth estimates. The truth estimates would be compared to the estimates from each of the alternative methods that used the truth deck dataset. The purpose was to determine how close the alternative methods were to the truth.. This was evaluated by examining the mean algebraic percent error and mean absolute percent error. These are defined in a section 3.8.

3.8 Evaluation

The weighting process was performed for each of the four alternative non-interview adjustment methods to produce the person weight and housing unit weight files. The alternative methods were compared against each other by comparing several population, housing, and housing unit estimates.

Several ACS base tables were produced for each alternative method at the national, state, and county levels. There are 3219 counties, which includes every state. Many of these tables were chosen based on research reported on the paper 2005 ACS Respondent Characteristics by Data Collection Mode (Joshipura, 2008). This research identified characteristics that differ significantly by mode. The description of the base tables and the number of estimates in the base tables used in this research is found in Table 1.

The base tables consist of total, subtotal, and dimension lines. The total line is the first line of every table and pertains to the total universe. The subtotal lines are aggregates of lower data lines. The dimension lines contain the most detail. For example, in the Sex by Age base table, the total is all females and males. One of the subtotal lines is all the males. One of the dimension lines is males under five years.

The base tables were produced using the intermediate non-interview adjustment weights. These are called the pre-controlled estimates meaning that the housing unit and population controls have not been applied. The purpose of using the pre-controlled estimates is to isolate the impact of the non-interview adjustment. The base tables were also produced using either the final person or housing unit weights. These are called the post-controlled estimates, which are the estimates that are published.

The mean algebraic percent error (MALPE) and mean absolute percent error (MAPE) were computed for each estimate and coefficient of variation (CV) across geography and table by type (total, subtotal, dimension). This was done for the four alternative methods. For example there were three MALPE and MAPE estimates, which are the types for each alternative non-interview adjustment method.

MALPE is defined as the mean of:

$$\frac{\text{Estimate or CV (alternative method)} - \text{Estimate or CV (Truth)}}{\text{Estimate or CV (Truth)}} \times 100$$

MAPE is defined as the mean of:

$$\left|\frac{\text{Estimate or CV (alternative method)} - \text{Estimate or CV (Truth)}}{\text{Estimate or CV (Truth)}}\right| \times 100$$

4. Findings

The logistic regression models were first estimated using the 2006 ACS production data with the original respondents and non-respondents. The three models with highest p-values from the statistics were chosen to go through the next step, which used the truth deck dataset instead of the original dataset. These models included first-order interactions but did not use the weights. None of the models using the weighting area or state geographies fit the data well.

4.1 Models Based on Original Data

The first model selected included the predictor variables building type, tabulation date, and urban/rural area. The geographic area used was a recoding of Census region. There are four regions, northeast, midwest, west, and south. However, the model did not fit well when midwest was included in the model. Therefore the sample for the midwest was allocated to the other three regions by state. Puerto Rico was placed in the south region. The p-values from the Hosmer-Lemeshow Statistics ranged from 0.79 to 0.93. The p-values from the Pearson Goodness-Of-Fit Statistics ranged from 0.13 to 0.48. The R-Square values ranged from 0.01 to 0.04.

The second model selected included the predictor variables building type, mode, and tabulation date. The tabulation date consisted of twelve binary variables. The geographic area selected was the binary urban/rural area. The p-values from the Hosmer-Lemeshow Statistics ranged from 0.81 to 0.98. The p-values from the Pearson Goodness-Of-Fit Statistics ranged from 0.16 to 0.26. The R-Square values ranged from 0.14 to 0.18.

The third model selected included the predictor variables building type and date. The date variable was separated into ten binary variables. This was determined by the distributions of the date variable. The geographic area selected was Census division. There are nine divisions. Puerto Rico was placed into the south division. The p-values from the Hosmer-Lemeshow Statistics ranged from 0.63 to 0.99. The p-values from the Pearson Goodness-Of-Fit Statistics ranged from 0.12 to 0.90. The R-Square values ranged from 0.15 to 0.27.

4.2 Models Based on Truth Deck Data

The three models were then estimated using the truth deck dataset. If the model did not fit the truth deck dataset well then the model was dropped.

The first model, which included the predictor variables building type, tabulation date, and urban/rural area did not fit the data well. The third region had a p-value of 0.008 for the Pearson Goodness-Of-Fit. Also in the Hosmer-Lemeshow Statistics the third region had less than five groups even though the p-value was equal to 0.99. This means that there is a possible over-fit.

The second model, which included the predictor variables building type, mode, and tabulation date, also did not fit the data well. Both urban/rural areas had less than five groups for the Hosmer –Lemeshow Statistics indicating a possible over fit. The p-values from the Pearson Goodness-Of-Fit Statistics ranged from 0.89 to 0.99 but since the Hosmer Lemeshow Statistics indicated a possible over fit this model was chosen not to go through the weighting process.

The third model included the predictor variables building type and date. The p-values from the Hosmer-Lemeshow Statistics ranged from 0.92 to 1.00. All of the groups were greater than five. The p-values from the Pearson Goodness-Of-Fit Statistics ranged from 0.12 to 0.96. The R-Square values ranged from 0.09 to 0.21. Therefore the third model was the only model to continue through the weighting process. The building type variable is kept in all of the Census divisions in the model.

The models for all of the divisions included the five highest date variables. This means that there were 27 days and longer from the time the ACS form was first mailed. This is expected since CATI starts the second month of a panel. In terms of interactions, all of the divisions had the two highest date variables with building type. This means that there were 53 days and longer from the time the ACS form was first mailed. This is not surprising since CAPI operations start the last month of the panel.

4.3 Evaluation of Alternative Methods

The distributions of the non-interview adjustment factors from using the alternative methods were produced. If the non-interview adjustments were high this would indicate that the variances would be high. This was not the case though. Table 2 shows the distributions of the non-interview adjustment factors. For the production method the factor was the overall non-interview adjustment factor since there were multiple non-interview adjustment factors. The distributions of the non-interview adjustment factors for the alternative methods using the propensity scores were similar to each other. The production method had higher non-interview adjustment factors for the upper quantiles and lower non-interview adjustment factors for the lower quantiles. The low non-interview adjustment factors in the production method are partly due to the MBF adjustment. The response propensity cannot be greater than one for the methods using the propensity scores, so these adjustment factors cannot be less than one.

The MALPE and MAPE were computed across geography and table by type for the pre-controls and post-controls at the national, state and county levels. The MALPE and MAPE estimates and CV's for the state and county levels are shown in Tables 3 and 4.

For the pre-controls the MALPE and MAPE estimates generally were the smallest for the production method for the total, summary, and dimension lines at the national, state, and county levels.

The MALPE and MAPE CV's were typically the smallest for the production method for the total, summary, and dimension lines at the national, state, and county levels. The method using the inverse of the propensity score had the highest MALPE CV's and MAPE CV's at the state and county levels. The method using the inverse of the propensity score and method using propensity scores to form the cells had the highest MALPE CV's and MAPE CV's at the national level.

For the post-controls the MALPE and MAPE estimates were generally the smallest for the production method for the total and summary lines at the national, state, and county levels. However, the MALPE and MAPE estimates were typically similar to one another for the dimension lines at the national, state, and county levels.

The MALPE and MAPE CV's were typically the smallest for the production method for the total and summary lines at the national, state, and county levels. However, the MALPE and MAPE estimates were typically similar to one another for the dimension lines at the national, state, and county levels.

The method using the inverse of the propensity score had the highest MALPE CV's and MAPE CV's at the county levels. The method using the inverse of the propensity score and method using propensity scores to form the cells had the highest MALPE CV's and MAPE CV's at the national and state levels.

5. Conclusion

The current non-interview adjustment methodology seems to perform better than the three non-interview adjustment methods using propensity scores in terms of the MALPE, and MAPE estimates. Also generally the MALPE, and MAPE CV's were higher in the method using the inverse propensity score than the other two alternative non-interview adjustment methods. The non-interview adjustment method using the mean inverse of the propensity score and method using the propensity score to form weighting adjustment cells perform better than the method using the inverse of the propensity score. There needs to be additional research to determine if there is an alternative non-interview adjustment method that performs better than the current non-interview adjustment methodology.

6. Limitations

There were limitations for the four alternative non-interview adjustment methods. One limitation is that the results are dependent on how the truth deck was constructed. Another limitation is that the number of non-interviews in the ACS is typically low so that the number of variable that can be used in any model is small.

7. Further Research

There are five additional alternative non-interview adjustment methods being considered. The five methods are variations of the current non-interview adjustment process. The methods are using only NIF1 only, NIF2, NIF1 and MBF, NIF2 and MBF, and only NIF1 and NIF2. There may be additional base tables produced for all nine of the methods. There needs to be additional research to determine if there is an alternative non-interview adjustment that could replace the current non-interview production process.

8. Acknowledgements

I wish to thank Mark Asiala and Lynn Weidman for their useful suggestions and comments.

9. References

Aertker, Leela Marie, and Kalsbeek, William D.,
(2006), "Estimated Response Propensities as a Means to Evaluate Error Effects Due to Nonresponse," Proceedings of the Annual Meeting of the American Statistical Association.

Carlson, B.L., and Williams, S. (2001), "A Comparison of Two Methods to Adjust for Non-Response: Propensity Modeling and Weighting Class Adjustments," Proceedings of the Annual Meeting of the American Statistical Association.

Diaz-Tena, N., Potter, F., Sinclair, M., and Williams, S. (2002), "Logistic Propensity Models to Adjust For Nonresponse in Physician Surveys," Proceedings of the Annual Meeting of the American Statistical Association.

Diaz-Tena, N., Potter, F., Ezzati-Rice, M., Greenblatt, Janet., and Wun, Lap-Ming. (2005), "Evaluation of Alternative Propensity Models for Adjusting Weights to Compensate for Dwelling Unit Nonresponse in the Medial Expenditure Panel Survey (MEPS)," Proceedings of the Annual Meeting of the American Statistical Association.

Joshipura, Megha (2008), "2005 ACS Respondent Characteristics Evaluation" U.S. Census Bureau.

U.S. Census Bureau (2006). "Design and Methodology: American Community Survey." *Technical Paper 67*. Issued May 2006; http://www.census.gov/acs/www/Downloads/tp67.pdf

Weidman, Lynn (2006), "Overview of Results from a Study of Four Simplified ACS Weighting Procedures" U.S. Census Bureau.

Table 1. Description of Tables and Number of Estimates

Table ID	Table Description	Total Lines [2]	Table ID	Table Description	Total Lines
B01001	Sex by Age	35	B16002	HH Language by Linguistic Isolation	14
B02001	Race	10	B17001	Poverty Status by Sex by Age	31
B03001	Hispanic or Latino Origin	10	B19001	HH Income in the Past 12 Months	17
B05003	Sex by Age by Citizenship Status	23	B25002	Occupancy Status	3
B07001	Residence 1 Year Ago by Age	36	B25003	Tenure	3
B08202	HH Size by Number of Workers in HH	22	B25004	Vacancy Status	8
B11001	HH Type (Including Living Alone)	9	B25007	Tenure by Age of Householder	21
B11011	HH Type by Units in Structure	19	B25020	Tenure by Rooms	21
B11016	HH Type by HH Size	16	B25024	Units in Structure	11
B12001	Sex by Marital Status	19	B25032	Tenure by Units in Structure	23
B14001	School Enrollment by Level of School	10	B25075	Value	25

Table 2. Distribution of Non-Interview Adjustment Factors

Method	Max	P99	P95	P90	Q3	Q2	Q1	P10	P5	P1	Min
Inverse of the Propensity Score	1.25	1.19	1.13	1.09	1.02	1.01	1.01	1.01	1.01	1.00	1.00
Mean Inverse of the Propensity Score	1.17	1.10	1.07	1.05	1.03	1.03	1.02	1.02	1.02	1.01	1.01
Propensity Scores to Form Cells	1.76	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Production Method	2.68	1.41	1.23	1.16	1.08	1.02	0.99	0.98	0.97	0.94	0.58

[2] This is at the national level. For the county level multiply by 3219.

Table 3. MALPE and MAPE Estimates (Pre – Controlled)[3]

Method	Type	State				County			
		MALPE Estimate	MAPE Estimate	MALPE CV	MAPE CV	MALPE Estimate	MAPE Estimate	MALPE CV	MAPE CV
Inverse Propensity Score	T	-0.75	1.17	9.82	9.93	-0.18	2.63	6.32	6.60
Mean Inverse Propensity Score	T	-1.82	1.86	9.81	9.90	-1.17	2.49	4.83	5.29
Propensity to Form Cells	T	-4.53	4.53	9.74	9.83	-3.59	3.61	4.82	5.28
Production Method	T	-0.13	0.15	0.63	1.06	-0.08	0.58	0.56	1.19
Inverse Propensity Score	S	-0.95	1.74	5.19	5.58	-0.17	4.38	3.62	4.00
Mean Inverse Propensity Score	S	-2.57	2.74	3.44	4.24	-1.42	4.01	2.29	2.85
Propensity to Form Cells	S	-5.54	5.54	3.38	4.20	-4.04	4.13	2.29	2.84
Production Method	S	-0.38	0.95	2.25	3.18	-0.01	3.35	1.58	2.57
Inverse Propensity Score	D	-0.86	2.27	4.00	4.49	-0.14	4.86	2.65	2.94
Mean Inverse Propensity Score	D	-2.25	2.89	2.38	3.39	-1.09	4.46	1.70	2.11
Propensity to Form Cells	D	-5.07	5.09	2.35	3.36	-3.59	3.71	1.70	2.09
Production Method	D	-0.18	1.84	2.02	3.12	0.24	4.65	1.40	2.27

Table 4. MALPE and MAPE Estimates (Post – Controlled)

Method	Type	State				County			
		MALPE Estimate	MAPE Estimate	MALPE CV	MAPE CV	MALPE Estimate	MAPE Estimate	MALPE CV	MAPE CV
Inverse Propensity Score	T	0.01	0.15	2.44	4.34	-0.08	1.35	2.27	3.63
Mean Inverse Propensity Score	T	-0.01	0.20	3.20	5.12	-0.12	1.30	1.59	3.10
Propensity to Form Cells	T	-0.12	0.51	3.75	5.96	-0.29	1.43	2.15	3.41
Production Method	T	-0.02	0.04	1.82	3.97	-0.03	0.65	0.77	2.19
Inverse Propensity Score	S	-0.07	0.70	3.46	4.26	0.03	3.79	8.64	9.99
Mean Inverse Propensity Score	S	-0.23	0.85	2.03	3.45	0.02	3.59	7.71	9.41
Propensity to Form Cells	S	-0.52	1.09	2.15	3.54	-0.31	3.59	7.87	9.47
Production Method	S	0.06	0.71	2.73	3.96	0.32	3.57	4.12	5.92
Inverse Propensity Score	D	-0.02	1.67	3.20	3.95	0.10	5.09	2.48	3.72
Mean Inverse Propensity Score	D	-0.13	1.79	1.89	3.19	0.26	4.83	1.81	3.32
Propensity to Form Cells	D	-0.32	2.10	2.00	3.29	0.03	4.81	1.95	3.39
Production Method	D	0.08	1.69	2.31	3.53	0.50	5.12	1.89	3.53

[3] MALPE and MAPE estimates and CV's were computed across geography and table by type.