A Cautionary Note on Post-stratification Adjustment

Hee-Choon Shin

National Center for Health Statistics,[1] 3311 Toledo Rd., Hyattsville, MD 20782

Abstract

Adjustment factors for non-response and coverage error are applied to the initial weights to produce the final analytic weights. Frequently the adjustment is applied using post-stratification. As Cochran suggested, the effectiveness of the post-stratification adjustment does completely depend on available true population parameters, or the control totals. In practice, however, the population parameters are frequently estimated directly or indirectly through modeling. This paper demonstrates potential adverse effects of using incorrect population parameters in post-stratification adjustment. In particular, caution should be used when blindly applying raking ratio adjustment – in general, the last step for well-designed surveys but frequently the only step for not-so-well-designed surveys. Additional necessary conditions for implementing post-stratification adjustment for surveys are discussed.

Key Words: Weighting, Post-stratification, Non-response adjustment, Coverage, Log-linear model.

1. Introduction

Frequently, adjustment factors for non-response and coverage error are applied to the initial weights to produce the final analytic weights using post-stratification method. Post-stratification is to reconcile the known differences between sample and population and has been researched and used (Cochran, 1977; Neyman, 1934; Stephan, 1941). The effectiveness of the post-stratification adjustment depends completely on available true population parameters, or the control totals. In practice, however, the population parameters are frequently estimated directly or indirectly through modeling (Dever & Valliant, 2010). Let W_h be the true stratum proportion or control total and w_h be an estimate of the h^{th} stratum (h = 1, 2, …, H). The sample estimate, $\bar{y}_{st}$, of the true population mean, $\bar{Y}$, is $\sum_{h=1}^{H} w_h\,\bar{y}_h$. The mean squared error (MSE) of $\bar{y}_{st}$ is

$$MSE(\bar{y}_{st}) = \sum_{h=1}^{H} \frac{w_h^2 S_h^2}{n_h}(1-f_h) + \left[\sum_{h=1}^{H}(w_h - W_h)\bar{Y}_h\right]^2, \qquad (1)$$

where n_h and f_h are the sample size and the sampling rate in the h^{th} stratum (Cochran, 1977; Stephan, 1941). Based on equation (1), Cochran listed three consequences of using weights that are in error:

1. The sample estimate is biased.

[1] The findings and conclusions stated in this manuscript are solely those of the author. They do not necessarily reflect the views of the National Center for Health Statistics or the Centers for Disease Control and Prevention.

2. The bias remains constant as the sample size increases.
3. The usual variance under-estimates the true error since it does not reflect the contribution of the bias to the error.

From a sampling design perspective, Cochran's second item is quite relevant. Notice that the second term of the right-hand side of (1) is independent of the sample size. Wrong control totals or the weights in error could negate the gain from increased sample size. Cochran (1977, P. 118) concludes that "accurate estimation of the W_h is particularly important when stratification is highly effective or when the sample size is large." In his example, stratification becomes inferior to simple random sampling when *n* is more than 300. In the following, we will discuss post-stratification adjustment from a modeling perspective. Then the potential problem of using incorrect stratum totals will be demonstrated by analyzing cross-tabulations of sex, race/ethnicity, and age from the 2010 U.S. Census. Our focus is on the true stratum sizes or control totals. Consequently, we do not discuss various methods for post-stratification to estimated control totals in this paper. A good discussion on various estimation methods and modeling approaches can be found in other research papers (Dever & Valliant, 2010; Fuller, 1966; Gelman, 2007; Holt & Smith, 1979; Little R. J., 1993; Valliant, 1993; Yung & Rao, 2000).

2. Post-stratification adjustment from a modeling perspective

Consider classification variables A, B, C with categories $i = 1,\ldots, I$, $j = 1,\ldots, J$, and $k = 1,\ldots, K$, respectively. We will discuss our approach with a three-way cross-tabulation but extending our approach to higher-dimensional tables is straight-forward. Now consider the following model, a "saturated" multiplicative model with zero degrees of freedom:

$$\pi_{ijk} = \eta \tau_i^A \tau_j^B \tau_k^C \tau_{ij}^{AB} \tau_{ik}^{AC} \tau_{jk}^{BC} \tau_{ijk}^{ABC}, \quad (2)$$

where π_{ijk} is the unconditional cell probability. Assuming $\pi_{ijk} > 0$ for all *i*, *j*, and *k*, and letting $\nu_{ijk} = \log \pi_{ijk}$, the τ parameters can be written as

$$\tau_i^A = \exp(\bar{\nu}_{i..} - \bar{\nu}_{...}) \quad (2.1)$$

$$\tau_{ij}^{AB} = \exp\left(\bar{\nu}_{ij.} - \bar{\nu}_{i..} - \bar{\nu}_{.j.} + \bar{\nu}_{...}\right) \quad (2.2)$$

$$\tau_{ijk}^{ABC} = \exp\left(\nu_{ijk} - \bar{\nu}_{ij.} - \bar{\nu}_{i.k} - \bar{\nu}_{.jk} + \bar{\nu}_{i..} + \bar{\nu}_{.i.} + \bar{\nu}_{..k} - \bar{\nu}_{...}\right) \quad (2.3)$$

with similar formulas for τ_j^B, τ_k^C, τ_{ik}^{AC}, and τ_{jk}^{BC}. The dot subscript denotes summation with respect to the subscript it replaces and the bar denotes average. The parameter η is a scale factor ensuring the sum $\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \pi_{ijk} = 1$. Details on estimation of the parameters can be found in earlier works (Bishop, Fienberg, & Holland, 1975; Fienberg, 1980; Goodman, 1970).

From a modeling perspective, the correct approach for post-stratification adjustment utilizes the saturated model to re-calibrate the analytic weights. To a degree, the parsimonious best model can be utilized to estimate the cell probabilities as long as the model fits the data well. In the absence of significant two-way and three-way interaction effects, for example, we could estimate the cell probabilities using only main effects or marginal distributions by omitting interaction effects from (2):

$$\hat{\pi}_{ijk} = \eta\tau_i^A\tau_j^B\tau_k^C \quad (3)$$

In reality, we tend to estimate the cell probabilities or unconditional distributions with a simple but frequently unwarranted assumption of mutual independence among the classification variables when the unconditional distributions are not available. Estimates of the unconditional probabilities based on the model in equation (3) can be obtained by the raking method which estimates cell probabilities with given marginal distributions of classification variables and initial weights (Deming & Stephan, 1940; Deville, Sarndal, & Sautory, 1993; Ireland & Kullback, 1968; Oh & Scheuren, 1983). When we apply post-stratification adjustment to the analytic weights using estimated cell probabilities, however, we regard the control totals as the population parameters or the truth. In the following, we fit various models to the cross-tabulation of sex, race/ethnicity, and age from the 2010 U.S. Census, and demonstrate potential adverse effects of using estimated counts for post-stratification adjustment. We discuss additional necessary conditions for implementing post-stratification adjustment for surveys.

3. Analysis

Table 1 shows the distribution of the adult (18 and over) U. S. population in the 2010 Census by sex, race/ethnicity and age. Among 234,564,071 adults in the U.S., for example, 2.36% were 18-29 year-old men of Hispanic origin. Table 1 also shows the marginal distributions by sex, race/ethnicity and age, respectively. In 2010, 51.47% of U.S. adults were women, 14.22% were of Hispanic origin, and 7.91% were 75 and older. We fit 8 models to the population distribution and Table 2 shows model specifications and goodness of fit statistics. Model I is a “main effects” model and Model VIII is a “no three-way interaction” model. The abbreviations (S, R and A) stand for sex, race/ethnicity and age, respectively. RA, for example, indicates the interaction term between race/ethnicity and age. Model VIII does not fit the data, implying significance of the three-way interaction. Another noteworthy fact is that the interaction effect (RA) between race/ethnicity and age is very significant in the models. Note the reduction in likelihood ratio in Models IV, VI, VII and VIII which contain the RA term. By adding the RA interaction to the main effects model, as seen in Model IV, we improve the fit by 89.8%.

Based on the Models shown in Table 2, we estimated the cell probabilities and corresponding frequencies. Table 3 shows the 2010 Census and estimated counts in ten thousands for each model. According to the 2010 Census, there were 5.53 million 18-29 year old men of Hispanic origin in the U.S and 450 thousand 75 and older men of Hispanic origin. Under Model I (mutual independence model), the corresponding estimated numbers were 3.57 million and 1.28 million, indicating an underestimation of 18-29 year olds by 1.96 million and an overestimation of 75 and older men by 830,000. Under Model VIII (no three-way interaction model), the estimated numbers were 5.45 million and 460,000 which were closer to the true values but still incurring an underestimation of 18-29 year olds by 80,000 and overestimation of 75 and older by 10,000.

To evaluate the impact of each model, we calculate a relative error rate (%) for each sub-group under each model as

$$Relative\ Error\ Rate\ (RER) = \frac{Estimated\ Count - Census\ Count}{Census\ Count} \times 100. \qquad (4)$$

A positive *RER* indicates an overestimation and a negative *RER* indicates an underestimation. Figure 1 shows the impact of estimation under each model graphically. The horizontal line indicates all the sub-groups by sex, race/ethnicity, and age. The line based on the saturated model would coincide with the horizontal axis, indicating a perfect fit. A line close to the horizontal axis indicates a good fit of the model to the data. The line for Model VIII does almost coincide with the horizontal line but still has small peaks and valleys for older age categories, implying significance of the three-way interactions. The line for Model I (mutual independence model) shows extreme level of peaks and valleys, indicating extreme over- and under-estimations for particular groups. The lines for Models II and III show similar patterns. The line for Model IV does not show the extreme peaks and valleys as compared to the first three Models', implying relatively a better fit.

4. Concluding Remarks

Post-stratification adjustment to analytic weights is very popular in the survey community and among data producers and users, alike. However, a blind application can result in adverse consequences, as Cochran cautioned us 35 years ago. In particular, caution is advised when blindly applying a raking ratio adjustment – in general the last step for well-designed surveys but frequently the only step for not-so-well-designed surveys – since it is largely based on Model I shown in equation (3) above. In their seminal paper, Deming and Stephan (1940) emphasized the imperfectness of their approach by stating that any or all of the adjusted counts in any table of their analysis were not necessarily "closer to the truth" than the corresponding sampling frequencies, even under ideal conditions**.**

Additional necessary conditions. We focused on correct stratum sizes or control totals in the above. To be able to post-stratify sample data for post-stratification adjustment, however, sample surveys need to meet four additional conditions: (a) The classification variables should be measured in the same way as was done in the control totals; (b) The classification variables should be measured without any missing item values; (c) There should be an adequate number of sample elements in each cell; and (d) The sample elements should have been selected at random within each cell. Condition (a) is a necessary condition for classification variables in sample data. In our example, we used four-category race/ethnicity allowing Hispanic origin and multiple races. Did a survey measure the race/ethnicity in the same way? If the answer is negative, in a strict sense, we cannot use the four-category control totals for race/ethnicity to post-stratify the sample data. Condition (b) is to classify responding sample data into strata. For classification variables with missing values, we could regard sample elements having any missing values for one of the classification variables as a missing unit and drop them from the analysis file. Or we could impute missing values of each classification variable in the proportion to observed distributions to retain valuable data as much as possible. The imputation would be another source of uncertainty for post-stratification adjustment. Condition (c) prevents extreme adjustment factors. Condition (d) makes statistical inference meaningful and, to a degree, possible. We cannot expect unbiased and efficient estimates from a non-probability sample, even if we post-stratify the sample.

Lastly, we should treat the sample as a stratified sample after post-stratification, and use correct estimators. The correct variance estimator for a simple random sample after post-stratification is the first term in the right-hand side of equation (1). Let us enjoy the fruit of our labor!

5. Acknowledgements

We thank Meena Khare, Don Malec and Iris Shimizu for helpful comments and suggestions.

References

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: Wiley.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics, 11*(4), 427-444.

Dever, J. A., & Valliant, R. (2010). A comparision of variance estimators for poststratification to estimated control totals. *Survey Methodology, 36*(1), 45-56.

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association, 88*(423), 1013-1020.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data.* Cambridge, MA: MIT Press.

Fuller, W. A. (1966). Estimation Employing Post Strata. *Journal of the American Statistical Association, 61*(316), 1172-1183.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science, 22*(2), 153-164.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association, 65*(329), 226-256.

Holt, D., & Smith, T. M. (1979). Post stratification. *Journal of the Royal Statistical Society. Series A (General), 142*(1), 33-46.

Ireland, C. T., & Kullback, C. (1968). Contingency tables with given marginals. *Biometrika, 1*, 179-188.

Little, R. J. A. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association, 88*(423), 1001-1012.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society, 97*(4), 558-625.

Oh, H. L., & Scheuren, F. J. (1983). Weighting adjustment for unit nonresponse. In W. G. Madow, I. Olkin, & D. B. Rubin (Eds.), *Incomplete Data in Sample Surveys* (Vol. 2 Theory and Bibliographies, pp. 143-184). New York: Academic Press.

Stephan, F. F. (1941). Stratification in representative sampling. *The Journal of Marketing, 6*(1), 38-46.

Valliant, R. (1993). Poststratification and Conditional Variance Estimation. *Journal of the American Statistical Association, 88*(421), 89-96.

Yung, W., & Rao, J. N. K. (2000). Jackknife Variance Estimation under Imputation for Estimators Using Poststratification. *Journal of the American Statistical Association, 95*(451), 903-915.

Table 1. Distribution (%) of Adult (18 and Over) U.S. Population in 2010 by Sex, Race/Ethnicity, and Age.

Sex	Race/Ethnicity	Age					
		18-29	30-44	45-59	60-74	75+	All Age
Men	Hispanic	2.36	2.51	1.55	0.59	0.19	7.19
	Non-Hispanic White	6.43	7.87	9.61	6.07	2.57	32.56
	Non-Hispanic Black	1.48	1.54	1.51	0.68	0.21	5.41
	Non-Hispanic Other including Multiple Races	0.94	1.06	0.84	0.41	0.13	3.37
	All Men	11.21	12.97	13.50	7.75	3.10	48.53
Women	Hispanic	2.10	2.36	1.57	0.70	0.29	7.03
	Non-Hispanic White	6.26	7.78	9.83	6.62	3.94	34.42
	Non-Hispanic Black	1.53	1.73	1.72	0.87	0.39	6.24
	Non-Hispanic Other including Multiple Races	0.97	1.18	0.96	0.49	0.19	3.79
	All Women	10.86	13.05	14.07	8.68	4.81	51.47
ALL	All Men and Women	22.07	26.02	27.57	16.43	7.91	100.00

Data: 2010 U.S. Census Summary File 1.

Table 2. Goodness of Fit Statistics of the Models.

Model	Effects [1)]	Degree of Freedom	Likelihood Ratio Chi-square	P-Value
I	S, R, A	31	9,744,376	<.0001
II	S, R, A, SR	28	9,627,215	<.0001
III	S, R, A, SA	27	8,903,249	<.0001
IV	S, R, A, RA	19	993,353	<.0001
V	S, R, A, SR, SA	24	8,786,088	<.0001
VI	S, R, A, SR, RA	16	876,192	<.0001
VII	S, R, A, SA, RA	15	152,225	<.0001
VIII	S, R, A, SR, SA, RA	12	28,180	<.0001

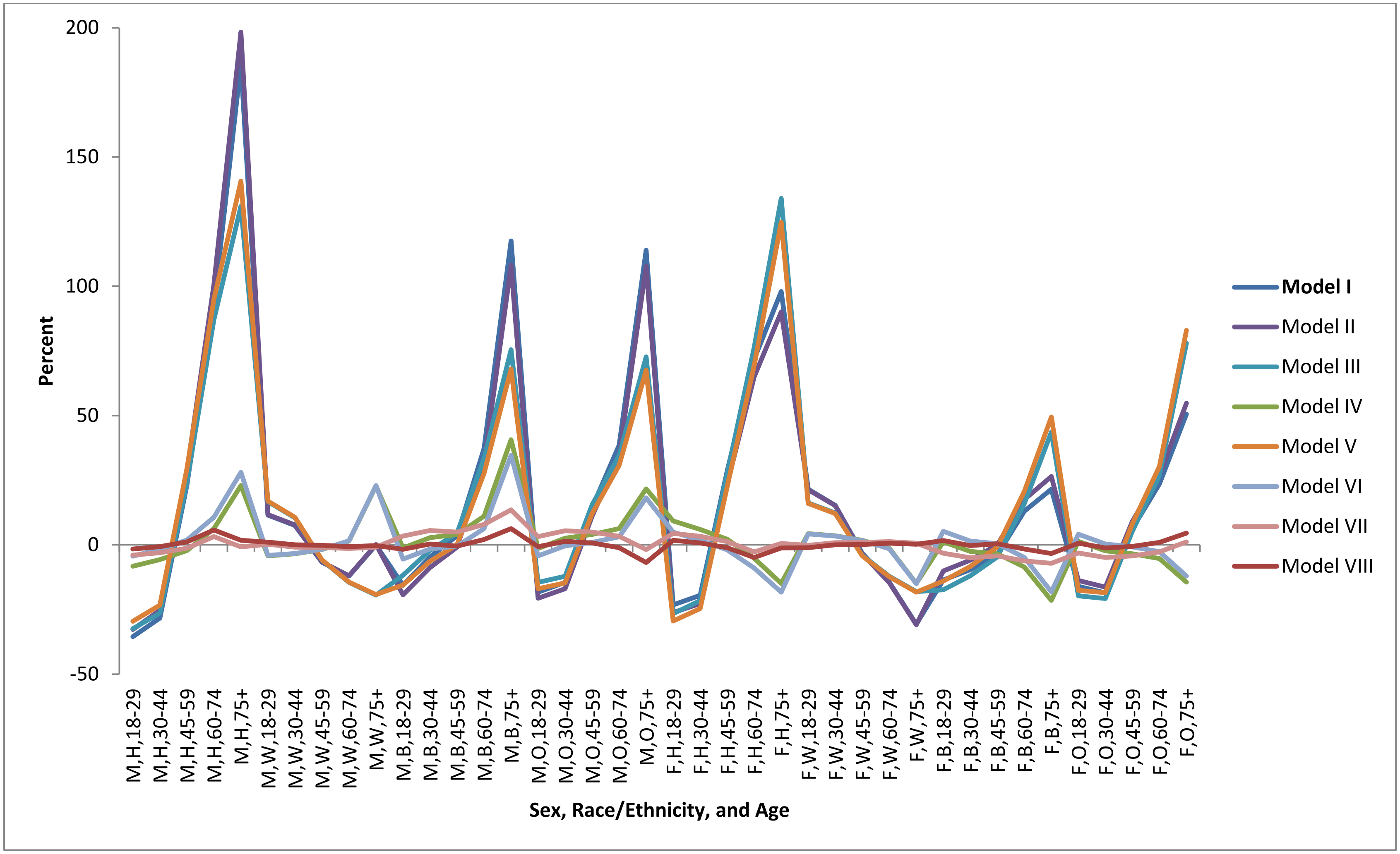
Note: [1)] S: Sex; R: Race/Ethnicity, A: Age

Table 3. 2010 Census Counts in Ten Thousands of Adult (18 and Over) U.S. Population and Estimated Counts under Each Model. [1)]

Sex	Race/Ethnicity	Age	2010 Census	Estimated I	II	III	IV	V	VI	VII	VIII
Men	Hispanic	18-29	553	357	372	374	508	390	529	532	545
		30-44	588	421	439	433	554	451	578	570	584
		45-59	363	446	465	450	355	469	369	358	367
		60-74	138	266	277	258	147	269	153	142	146
		75+	45	128	133	103	55	108	57	44	46
	Non-Hispanic White	18-29	1,509	1,683	1,686	1,761	1,445	1,764	1,447	1,512	1,525
		30-44	1,845	1,984	1,987	2,038	1,781	2,042	1,784	1,830	1,845
		45-59	2,255	2,102	2,106	2,121	2,213	2,124	2,217	2,233	2,251
		60-74	1,424	1,252	1,255	1,217	1,444	1,219	1,447	1,404	1,414
		75+	604	603	604	487	741	487	742	598	602
	Non-Hispanic Black	18-29	347	293	280	306	343	293	328	359	341
		30-44	362	345	330	354	372	339	356	382	363
		45-59	353	366	350	369	367	353	351	370	351
		60-74	159	218	208	212	176	203	169	171	162
		75+	48	105	100	85	68	81	65	55	51
	Non-Hispanic Other including Multiple Races	18-29	220	180	175	188	217	183	211	227	219
		30-44	248	212	206	218	254	211	247	261	251
		45-59	196	225	218	227	204	220	198	206	198
		60-74	97	134	130	130	103	126	100	100	96
		75+	30	64	63	52	37	51	36	30	28
Women	Hispanic	18-29	493	379	364	362	539	348	517	515	502
		30-44	555	447	429	435	588	418	565	573	559
		45-59	368	473	454	469	376	451	361	373	364
		60-74	164	282	271	289	155	278	149	160	156
		75+	69	136	130	160	58	154	56	69	68
	Non-Hispanic White	18-29	1,468	1,785	1,782	1,706	1,532	1,704	1,530	1,465	1,452
		30-44	1,824	2,104	2,101	2,049	1,889	2,046	1,886	1,840	1,825
		45-59	2,305	2,229	2,226	2,211	2,347	2,207	2,343	2,327	2,309
		60-74	1,552	1,328	1,326	1,363	1,532	1,361	1,529	1,573	1,562
		75+	923	640	639	756	786	755	785	929	925
	Non-Hispanic Black	18-29	359	310	323	297	363	309	378	347	365
		30-44	405	366	381	356	395	371	411	385	404
		45-59	403	388	403	384	389	400	405	386	405
		60-74	204	231	240	237	187	247	194	192	201
		75+	92	111	116	131	72	137	75	85	89
	Non-Hispanic Other including Multiple Races	18-29	227	191	196	182	230	187	237	220	229
		30-44	276	225	231	219	270	225	277	263	273
		45-59	224	238	245	236	216	243	222	215	223
		60-74	115	142	146	146	109	150	112	112	116
		75+	45	68	70	81	39	83	40	46	47

Note: See Table 2 for specification of models.

Figure 1. Relative Error Rates (RERs) of Models for a Cross-Tabulation of Sex, Race/Ethnicity and Age from the 2010 U.S. Census.



Note: M="Men"; F="Women"; H="Hispanic", W="Non-Hispanic White"; B="Non-Hispanic Black"; O="Non-Hispanic Other including multiple races".