Post-stratification and Response Bias in Survey Data with Applications in Political Science

Iffigenia Barboza and Rohan Williams
Department of Statistics and Political Science, Michigan State University*

April 8, 2005

Abstract

Post-stratification is a common technique implemented to obtain more precise estimates of sample statistics in survey data. If used correctly, this technique increases the representativeness of the sample so we have greater confidence in the validity of our inferences about population parameters of interest. Methodologically speaking, however, using post-stratification to correct for the effects of differential non-response in the poststrata for the purpose of increasing the non-reflectiveness of the population is problematic. Moreover, when proportional allocation is used to design the sample, it is not necessarily the case that the number of units in the simple random sample for stratum h is proportional to the total number of sampling units in each stratum, or $n_h \propto N_h$. In this paper, we illustrate the pros and cons of using post-stratification as a method to increase representation due to non-response by simulation, using race as our post-stratifying variable. The population information is drawn from the Census' Current Population Survey March 2004 Supplement. Our intent is to design criteria that promote careful use of post-stratification when the survey instrument may not be reflective of certain segments of the population.

KEYWORDS: POST-STRATIFICATION, DIFFERENTIAL NONRESPONSE, NONIGNORABLE NONRESPONSE

1 Introduction

Stratifying sample data on known variables is often fruitless for several reasons, the most common reason being the difficulty involved in compiling a sufficient sampling frame. Post-stratification is sometimes used when a good stratification criterion is known but it is not feasible to sample from the strata or it is simply too costly to do so. Moreover, when the sample is not representative of the population, post-stratification is employed to mimic the stratification process. The main difference is that post-stratification methods rely on a process of stratification that occurs only after the sample is taken. This statistical methodology must be carefully implemented to avoid the dangers and pitfalls that occur with incorrect use.

In order to use post-stratification effectively the ratio of the candidate stratum proportions, i.e., $\frac{N_h}{N}$, must be known. In many social and political surveys it may be difficult to stratify such factors as income and age by race. The reason is that the stratum proportions are *not* known beforehand. Methodologically however we can get post-stratification proportions for example by turning to Census data via the following process: take a simple random sample of size n, split the data into the H strata and proceed as if we had originally stratified the random sample. Note, however, that under the stratification procedure, the n_h are fixed whereas when we post-stratify the n_h are random. This may cause difficulty for estimating population parameters.

A primary use of post-stratification is to reduce nonresponse bias in surveys. (Bethlehem 1988; Smith 1991; Zhang 1999). With respect to nonresponse, post-stratification estimation is most useful when the respondents are not reflective of the population. If the nonreflectiveness of the respondents

*Acknowledgements: The authors are grateful to Drs. Jim Stapleton ("Lao-Shi") and Connie Page for their support and guidance.

depends on the variable of interest, the sample mean estimator will be biased. If this is true, the post-stratification estimator, which incorporates known information about the size of components of the population, gives "better" results. In fact, post-stratification will completely remove nonresponse bias if the nonresponse is conditionally independent of the variable of interest within each post-stratum. (Zhang 1999). For example, it is common in political surveys for response rates to be lower among various ethnic subpopulations. If we are asking a simple random sample of individuals about political activities that are more common among whites and we use the basic sample mean estimator then a higher response rate for whites will lead to a biased over-estimate of the political behaviors of the population.

A major drawback of post-stratification, however, is that the nonresponse is rarely conditionally independent of the variable of interest *within* post-strata, and this is especially true for survey data in political science. This happens if the nonresponse is more severe among one subsample than the other in which case the response is not strictly independent of the post-strata conditional on the variable of interest. (Zhang 1999). To illustrate, consider estimating the proportion of registered voters who participated in the last primary election. Assuming that the nonresponse is conditionally independent of the variable of interest within each post-stratum would require assuming that the nonresponse does not depend on whether or not the individual voted given the individual's race. A more reasonable assumption would be that, for example, certain minority groups are less likely to respond if they did not vote in the last election, in which case the nonresponse is more severe among one subsample. In this paper, we first discuss the relative effect of post-stratification when the variable of interest depends on the response variable. Next, we show the relative reduction in bias and efficiency given that the dependency between the response and the variable of interest is moderate. Through simulation, we are able to evaluate two primary assumptions made by Zhang (1999), who showed that the bias-reduction due to post-stratification can be estimated from the respondents alone, with respect to variables of interest in political science. In this way we hope not only to extend Zhang's results, but also to make this method available to political scientists.

1.1 Sources of Selection Bias in Sample Survey Data

Selection bias occurs when the sampled population is not representative of the target population. There are several sources of selection bias in sample survey data, only one of which we are concerned with here: survey non-response.

1.2 Nonresponse Bias

Nonresponse is the failure to obtain responses from the sampling units ("SU") during the data collection process. Suppose that we take a simple random sample without replacement ("SRSWOR") of size n from N SUs. A non-response problem occurs if only n_r of the n units actually respond such that $n_r < n$. The response rate for the survey is simply $\frac{n_r}{n}$, or in percentage terms, $100 \times \frac{n_r}{n}$. Similarly, the nonresponse rate is $\frac{n_M}{n}$. Nonresponse occurs only after the sample is selected and can distort the results even of surveys designed to minimize other sources of selection bias. (Lohr 1999). There are two types of nonresponse: unit nonresponse and item nonresponse. Unit nonresponse is when data for an entire observation unit ("OU") is missing whereas item nonresponse occurs when only a subset of the SU's data is missing. Practically speaking, this means that one or more, but not all, of the questions were unanswered by the respondent.

2 Numerical Illustrations

2.1 Post-stratification to Increase Precision

Suppose we wish to estimate average voter turnout in East Lansing, Michigan. The target population is registered voters living in East Lansing. We believe that ethnicity would be a beneficial stratification criteria, but voter registration lists do not include information about, inter alia, race or ethnicity. Therefore, in order to incorporate the auxiliary information into our estimates of population means

and variances, we can use Census data to acquire the distribution of individuals by race and post-stratify.

Table 1: 2000 Census data on Racial Composition in East Lansing

Racial Composition	Number of Individuals	Percent
White	77766	65%
African American	26095	22%
American Indian	953	.8%
Asian	3367	3%
Pacific Islander	62	.05%
Other	5410	5.0%

h	$100 \times \frac{N_h}{N}$	n_h	$\frac{n_{hR}}{n_h}$	n_{hR}
White	65%	1200	88%	1056
African American	22%	400	45%	180
American Indian	.8%	49	36%	43
Asian	3%	220	22%	120
Pacific Islander	.05%	60	10%	6
Other	5.0%	60	16%	10
Total	100			

Then we observe the following post-stratum sample means and variances:

h	$\frac{N_h}{N}$	$\hat{p}_{hR}$	s_h^2
1	.65	.55	.10
2	.22	.45	.12
3	.008	.41	.08
4	.03	.48	.14
5	.005	.43	.09
6	.05	.42	.14

The post-stratified estimate of the population proportion of individuals in East Lansing who voted is given by:

$$\hat{p}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_{hR} = .49735 \tag{1}$$

and the post-stratum sample variance is

$$\hat{Var}[\hat{p}_{post}] \simeq \sum_{h=1}^{H} (\frac{N_h}{N})^2 (1 - \frac{n_{hR}}{N_h}) \frac{s_h^2}{n_{hR}} = .0001 \tag{2}$$

Under *simple random sampling* however, the estimate is

$$\hat{p}_{srs} = \sum_{i=1}^{n_R} \frac{\hat{p}_i}{n_R} = .45667 \tag{3}$$

and the variance is

$$\sum_{i=1}^{n_R} (1 - \frac{n_R}{N}) \frac{\hat{p}(1-\hat{p})}{n-1} = .0002 \tag{4}$$

which yields a 50% reduction in variance by using post-stratification.

2.2 Post-stratification and Weighting

Consider the target population of registered voters in East Lansing, Michigan. Our sampling design first stratifies the city into four geographic regions consisting of 200, 400, 600 and 800 blocks each. A simple random sample of 40, 40, 60, 80 blocks is taken from each region, respectively. For each selected block, a simple random sample of 10 households is taken and one adult is selected at random for an interview. Note that region number one has been oversampled since under proportional allocation, $\frac{N_h}{N} \times n = n_h = 20$. In this example, the primary sampling unit is the block, the secondary sampling unit is the household and the tertiary sampling unit is the individual. A hypothetical subset of the observations is given below, along with the sampling weight for each of the listed observation units.

Region	Block	# HHs	HH Label	Adults in HH	Adult Label
1	35	120	95	3	2
2	297	56	06	4	3
3	488	52	20	1	1
4	789	98	67	6	5

The sampling weights are given by:

$$\begin{array}{c} w_{ijk} = w_i \times w_{j|i} \times w_{k|i,j} \\ \hline 5 \times 12 \times 3 = 180 \\ 10 \times 5.6 \times 4 = 224 \\ 10 \times 5.2 \times 1 = 52 \\ 10 \times 9.8 \times 6 = 588 \end{array}$$

Now suppose we wanted to post-stratify by household type, which effectively subdivides the secondary sampling units into single parent and dual parent households. Suppose further that household 95 is a single parent household and that the number of single parent households in Region 1 is 45. Finally, our sample of households contained 5 single parent households. Accordingly, we can give a new sampling weight for the selected adult in that household (Region #1, Block #35, Household #95, Adult #2 in this example) with post-stratification.

$$\frac{200}{40} \times \frac{45}{5} \times \frac{3}{1} = 135$$

Note that the post-stratification occurred at the secondary sampling unit (household) level.

3 *Zhang's* General Framework

We follow the general framework of Zhang and let $U = \{1, ..., N\}$ denote the population and $s = \{1, ..., n\}$ the sample. Let $\overline{y}_U$ be the population mean of a binary random variable and let X be the post-stratifying variable. Here, X provides auxiliary information and it, too, is dichotomous. To continue, let R be the response variable such that $R = 1$ implies a respondent whereas $R = 0$ implies a nonrespondent. Denote the population proportion of $(X, Y) = (i, j)$ for $(i, j) = (0, 1)$ by $q_{ij} = N_{ij}/N$ and r_{ij} the response rate within the population group $(X, Y) = (i, j)$. According to Zhang, the population and expected sample have the following distribution:

The observed sample mean is given by

$$\overline{y}_{s_r} = \frac{[n_r(1,1) + n_r(0,1)]}{n_r} \tag{5}$$

	$Y=1$		$Y=0$	
	$R=1$	$R=0$	$R=1$	$R=0$
$X=1$	$q_{11}(1-r_{11})$	$q_{11}r_{11}$	$q_{10}(1-r_{10})$	$q_{10}r_{10}$
$X=0$	$q_{01}(1-r_{01})$	$q_{01}r_{01}$	$q_{00}(1-r_{00})$	$q_{00}r_{00}$

Table 2: Distribution of Population and Sample

where $n_r(i,j)$ denotes the size of the subsample $(X,Y)=(i,j)$ within the response group s_r. Finally, Zhang derives a formula for the bias of the simple mean estimator that is given by

$$E(\overline{y}_{s_r} - p|n_r) = \frac{n\{p[q_{00}r_{00} + q_{10}r_{10}] - q[q_{01}r_{01} + q_{11}r_{11}]\}}{E(n_r)} \tag{6}$$

Since post-stratification is applied only after the sample is taken it acts to further subdivide the sample of respondents into $s_r = (s_{r,1}, s_{r,2})$ and therefore also "stratifies" the sample of n_r respondents into mutually exclusive subgroups with sizes $n_{r,1}$ and $n_{r,0}$. The post-stratified mean is given by

$$\overline{y}_{post} = \frac{qn_r(1,1)}{n_{r,1}} + \frac{(1-q)n_r(0,1)}{n_{r,0}} \tag{7}$$

and the bias is

$$E(\overline{y}_{post} - p|n_{r,1}, n_{r,0}) = \frac{q_{11}q_{10}(r_{10} - r_{11})}{E(n_{r,1})/n} + \frac{q_{01}q_{00}(r_{00} - r_{01})}{E(n_{r,0})/n} \tag{8}$$

Zhang claims that the ratio of the biases, denoted by $\gamma = b_{post}/b_{srs}$ can be estimated from the response group alone. Finally, the bias correcting estimator that Zhang suggests be applied is $\overline{y}_{adj} = -\overline{y}\hat{\gamma}/(1-\gamma) + \overline{y}_{post}/(1-\hat{\gamma})$. If it is the case that q is unknown then it can be estimated from the data by $\hat{q}^* = n_1/n$ where n_1 is the size of the sample post-stratum $X=1$. (Zhang 1999). Non-response is ignorable if R is independent of Y given X whereas it is nonignorable if R is dependent on Y given X. According to Zhang, the nonresponse mechanism involves assuming that R is independent of X given Y, which implies that $r_{i0} = r_0$ and $r_{i1} = r_1$ for $(i = 0, 1)$. To continue, Zhang states that this assumption can be checked by considering the sample as having been generated under the model where $P(X,Y) = (i,j)) = q_{ij}$ and $P(R=0|(X,Y)=(i,j)) = r_{ij}$ and obtaining the likelihood function proportional to $P((X,Y,R))$. This can be done by calculating the maximum likelihood estimates and computing the scaled deviance which is twice the difference between the maximum attainable log-likelihood and the fitted log-likelihood. Whereas a good fit is not enough to establish the nonignorable nonresponse assumption and bad fit is, according to Zhang, certainly evidence against it. Finally, Zhang presents a formula originally derived by Thomsen (1978) for calculating the variances of the simple random and post-stratified estimators. The are:

$$Var(\overline{y}_{srs}) = E(n_r(1,1) + n_r(0,1)) \cdot E(n_r(1,0) + n_r(0,0))/E(n_r)^3 \tag{9}$$

$$Var(\overline{y}_{post} = q^2 E(n_r(1,1) \cdot E(n_r(1,0))/E(n_{r,1})^3 + (1-q)E(n_r(0,1)) \cdot E(n_r(0,0))/E(n_{r,0})^3 \tag{10}$$

The simple nonresponse assumption was not met in Zhang's case since "it is probable that the nonresponse is indeed more severe among the subsample $(X,Y)=(0,0)$ than among $(X,Y)=(1,0)$, in which case R is not strictly independent of X conditional on Y," therefore $\overline{y}_{adj}$ did not correct for the bias in his sample. In addition, the correlation between the dependent and independent variables Zhang uses is extremely high, much higher than we would normally see among variables in political science. Therefore, we set out to see whether Zhang's results (1) can be improved if the simple nonresponse assumption holds and (2) whether his results apply if the initial correlation is reasonable for social science data ($\rho \leq .50$).

4 An Example: The Youth Circle Survey

In this section, we apply the results of Zhang (1999), who derived an expression for the bias of both the observed sample mean and the post-stratified estimator assuming simple random sampling. We further apply Zhang's methodology to estimate the bias of the post-stratified estimator under the simple nonignorable nonresponse assumption.

Proceeding under the assumption that the nonresponse (R) is independent of ethnicity (X) conditional on registration status (Y) but that registration status varies by Hispanic origin, we apply Zhang's model to data from Youth Circle Political Participation Survey. We follow Zhang's general notation changing it only for clarity and consistency. In our example, $U = \{1, ..., N\} = \{1, ..., 1600\}$ denotes the population and $s = \{1, ..., n\} = \{1, ..., 640\}$ the sample. Let $\overline{y}_U$ be the population mean the decision for whether to register to vote among 18-24 year olds. Let $X = 1$ if the individual is of Hispanic origin and $X = 0$ if the individual is Non-Hispanic. Note that we have chosen to let ethnic origin provide us with auxiliary information and that it, too, is dichotomous. In accordance with Zhang, we let R be the response variable such that $R = 1$ implies a respondent whereas $R = 0$ implies a nonrespondent. Table 2 shows the population distribution whereas Table 3 shows the distribution of the sample. Notice that the resulting probabilities are very similar.

	$Y=1$		$Y=0$	
	$R=1$	$R=0$	$R=1$	$R=0$
$X=1$	.11	.04	.075	.04
$X=0$	.36	.11	.184	.10

Table 3: Distribution of Population in Youth Circle Data

	$Y=1$		$Y=0$	
	$R=1$	$R=0$	$R=1$	$R=0$
$X=1$	.10	.042	.07	.044
$X=0$	.36	.10	.19	.094

Table 4: Distribution of Sample in Youth Circle Data

The population mean, $\overline{y}_U$ is given by $p = q_{11} + q_{01} = .605$.[1] Furthermore, the marginal proportion of $X = 1$ is given by $q = q_{11} + q_{10}$ which in our case is .25625.[2]

We used standard statistical procedures to generate a simple random sample with nonresponse from the data. We denote $s_r = \{1, ..., n_r\}$ and $s_{nr} = \{1, ..., n - n_r\}$ where s_r is the sample of respondents and s_{nr} the sample of nonrespondents.

$X=1$			$X=0$		
$(Y,R)=(1,1)$	$(Y,R)=(0,1)$	$R=0$	$(Y,R)=(1,1)$	$(Y,R)=(0,1)$	$R=0$
64	43	54	239	116	124

Table 5: Data from the Youth Circle Sample

The ratio of the estimated variances gave us $\hat{\eta} \simeq 1.00$, in other words post-stratification according to ethnicity had no effect on the variance. In addition, the correlation coefficient between X and Y among the respondents, $\hat{\rho}$ was -0.0752 and $1 - \hat{\rho} = .99434496 \doteq \eta$, as Zhang suggests is should be. Applying the above formulas results in $(\overline{y}, \overline{y}_{post}, \overline{y}_{adj}) = (.660, .653, .643)$ with the known proportion

[1] This was obtained by calculating the following expression: $\frac{230+738}{1600} = .605$ and then verified using a statistical package that calculated the population mean directly from the data: sum regvote Variable Obs Mean Std. Dev.

[2] This was obtained by calculating the following expression:‘ $\frac{230+180}{1600} = .25625$ and then verified using a statistical package that calculated the marginal proportion of $X = 1$ directly from the data (see appendix for output).

registered to vote in the population being .605. This gave a ratio of the estimated biases under the simple nonignorable model of $\hat{\gamma} = .873$. This means that we gain an estimated $(1-.873)\times 100 = 12.7\%$ reduction in nonresponse bias by using the post-stratification estimator instead of the simple random sample estimator. Notice that here $\hat{\gamma} \neq 1 - \hat{\rho}$ which implies that the assumption is not good in this example. To check this we followed Zhang's advice and evaluated the nonresponse assumption $r_{i0} = r_0$ and $r_{i0} = r_1$ for $i = 0, 1$ from a model perspective which gave us $\chi^2_{LR} = 10.76252$, $p = 0.0046$. By doing this we are assured that the nonresponse assumption is invalid. A test of the null hypothesis of no difference between respondents and nonrespondents within each post-stratum with respect to registration status resulted in rejection of the null hypothesis and hence a violation of the basic assumption.

Post-stratum	**Hypothesis**	***t*-value**	$\boldsymbol{P < t}$
$X = 1$	$\mu_R - \mu_{NR} = 0$	-1.4074	.0806
$X = 0$	$\mu_R - \mu_{NR} = 0$	-2.3453	.0097

Table 6: t-test of Means of registration status by respondents

In the above example the correlation between registration status and ethnicity was low ($\rho = -.06$). We would like to examine the effects of post-stratification when the correlation between X and Y is higher. Therefore, we expect that post-stratification will result in a reduction in the variance of the estimators if we post-stratify on whether the respondent believes his or her vote matters. In addition, we do not have to assume as does Zhang that voting registration is lower among nonrespondents, even when conditional on X since we have this information available to us. Therefore, the nonresponse is nonignorable.

	$Y = 1$		$Y = 0$	
	$R = 1$	$R = 0$	$R = 1$	$R = 0$
$X = 1$	.383	.114	.145	.07
$X = 0$	.082	.026	.112	.069

Table 7: Distribution of Population in Youth Circle Data

	$Y = 1$		$Y = 0$	
	$R = 1$	$R = 0$	$R = 1$	$R = 0$
$X = 1$	.378	.119	.143	.071
$X = 0$	.083	.025	.116	.065

Table 8: Distribution of Sample in Youth Circle Data

Recall the population mean is .605 but now the marginal proportion of $X = 1$, still given by $q = q_{11} + q_{10}$, is .7113. Applying the above formulas gives $(\overline{y}, \overline{y}_{post}, \overline{y}_{adj}) = (.6600, .6534, .6050)$. Notice that the adjusted estimator results in zero bias, which is a nice result[3].

Calculation of the estimated variances gives $\hat{\eta} = .9424$ which yields an estimated 6% reduction in the variance due to post-stratification according to whether the individual feels his vote matters. Also, $\hat{\eta} \doteq 1 - \hat{\rho}^2 = .931$. In this case, we are assured that the simple nonresponse assumption is valid since the nonresponse is not more severe among the subsample $(X, Y) = (0, 0)$ than among $(X, Y) = (1, 0)$, or among individuals who do not believe their vote matters and did not register than among individuals who do believe their vote matters and did not register. For both of these cases, approximately 1/3 of the individuals did not respond. This explains why the adjusted estimate corrects for the bias in the simple random sample estimator. In this example, too, post-stratification

[3]The ratio of the biases was calculated as $\frac{.0484}{.0550} = .88$. Applying the adjustment yields $-.660 \times \frac{.88}{(1-.88)} + \frac{.6534}{(1-.88)} = -4.84 + 5.445 = .605$

$X=1$			$X=0$		
$(Y,R)=(1,1)$	$(Y,R)=(0,1)$	$R=0$	$(Y,R)=(1,1)$	$(Y,R)=(0,1)$	$R=0$
243	90	122	60	69	56

Table 9: Data from the Youth Circle Sample with Vote Matters as the Post-stratifying variable

results in a reduction of both the bias and variance caused by the nonresponse although the reduction in the bias is extremely small. The data also suggests that the nonignorable nonresponse assumption is invalid ($\chi^2_{LR} = 8.1, p = .0178$). Importantly, the bias-correcting estimator given by $\overline{y}_{adj}$ does eliminate the bias resulting from the nonignorable nonresponse. In addition, we find that the variance is reduced by using post-stratification presumably because of the higher correlation between X and Y in our sample ($\rho = .26$). Since Zhang did not have data for the nonrespondents, he was unsure whether the simple nonresponse assumption held. We, on the other hand, were certain that it did hold. When this is the case, the bias which results from differential nonresponse seems to be completely eliminated by post-stratifying and applying Zhang's bias-correction factor to the resulting estimates. Further, we found too that the bias-reduction due to post-stratification can indeed be estimated from the respondents alone, as Zhang suggests, even when the nonrespondents differ from the respondents in the sample. We cannot be certain that this was a fluke or it will work consistently until more work in this area is done.

5 Example: Post-stratification and Binary Regression Analysis in Youth Vote's National Youth Survey

Data on the voting behavior of various segments of the population is widely available and of general interest among political scientists. Here we analyze four subsamples of data derived from the Youth Vote Coalition's National Youth Survey (2002). The Youth Vote national survey was conducted by telephone using random digit dialing. The survey reached a total of 1,600 young people between the ages of 18 to 24 nationwide, including 300 Hispanics and 300 African Americans. The four subsamples we analyze are: a simple random sample, a stratified independent sample, a one-stage cluster sample and a two stage cluster sample. For each subsample, we compared the mean overall registered voting percentage across models to the true population estimates and fit regression models to compare model coefficients and standard errors. To do the analysis, we relied on the R statistical programming language. The survey library in R allows us to build a `survey.design` object that contained the variables given in Table 10.

Table 10: Variables from the Circle Youth Survey used in the analysis

Data Variable	Description
age	Respondent's Age (18 - 24)
posths	=1 if R had post-high school education
votmat	=1 if R believes his vote matters
female	=1 if female
democrat	=1 if Democrat
discpar	=1 if R discusses politics with parents
attn	=1 if R pays attention to politics
relig	=1 if R regularly attends religious services
voteimp	=1 if R believes his vote is important)

Next, to increase efficiency we used a Taylor series linearization estimator and then converted the design object to a jackknife replicate weight design. We were then able to post-stratify the data using the appropriate statistical functions that are packaged with the survey library.

	θ	θ_{srs}	θ_{strat}	θ_{clus_1}	θ_{clus_2}	$\theta_{jkk_{strat}}$	θ_{jkk_2}	θ_{post_1}	θ_{post_2}
Estimate	.1516	.0855	.1765	.1253	.1363	.1765	.1363	.1383	.1571
Std. Error	.0262	.0411	.0430	.0239	.0461	.0440	.0486	.0462	.0307
t value	5.78	2.08	4.11	5.24	2.95	4.01	2.81	2.99	5.11
$Pr(>\|t\|)$	0.000	0.038	0.000	0.000	0.004	0.001	0.006	0.046	0.031

Table 11: Estimates, standard errors and probability values for the survey design objects in the example

	θ	θ_{srs}	θ_{strat}	θ_{clus_1}	θ_{clus_2}	$\theta_{jkk_{strat}}$	θ_{jkk_2}	θ_{post_1}	θ_{post_2}
Estimate	.2275	.2358	.2260	.2475	.2394	.2260	.2394	.2367	.2352
Std. Error	.0242	.0383	.0386	.0342	.0329	.0395	.0347	.0335	.0341
t value	9.42	6.16	5.86	7.23	7.27	5.73	6.89	7.07	6.90
$Pr(>\|t\|)$	0.000	5.86	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 12: Estimates, standard errors and probability values for the survey design objects in the example

6 Discussion

We are interested in two variables that impact the decision of whether or not to register to vote among 18-24 year olds, namely, whether the individual believes his vote matters and whether the individual believes his vote is important. The results of the logistic regression indicate that both variables have a very statistically significant effect on the decision to register. In addition, the probability of registering to vote is greater if the individual believes his vote matters, and the same holds if the individual believes his vote is important. For the variable *VOTEMAT*, post-stratification resulted in practically all of the bias removed and had the second smallest standard error (second only to the one-stage cluster design). We also see improvement for the variable *VOTEIMP*. With respect to this variable, post-stratification resulted in a reduction of the variance over all other survey design objects except the two-stage cluster. Furthermore, it also resulted in more of the bias removed than any other estimator except the stratified random independent sample one (including the jackknife). Importantly, post-stratification gave better results than our simple random sample estimator in both cases. Researchers can make up their own mind but based on our results, we recommend that at the very least, researchers must account for the original survey design, which would require use of a capable statistical package. In this manner, they can be certain that their inferences are more valid than those resulting on a reliance of the simple random sample estimators.

Table 13: Estimates of the mean and standard error of REGVOTE by different type of survey design

Survey Design Object	$\theta_j - \hat{\theta}_j$	$se[\hat{\theta}]_j$
Simple Random	-0.0661	-0.0149
Stratified Independent Random Sample	-0.0249	-0.0168
One-stage Cluster	0.0263	0.0023
Two-stage Cluster	0.0153	-0.0199
Jackknife for stratified independent sample	-0.0249	-0.0178
Jackknife for two-stage cluster sample	0.0153	-0.0224
Post-stratification of cluster sample	0.0133	-0.02
Post-stratification using two variables	-0.0055	-0.0045

Table 14: Estimates of the mean and standard error of REGVOTE by different type of survey design

Survey Design Object	$\theta_j - \hat{\theta}_j$	$se[\hat{\theta}]_j$
Simple Random	-0.0083	-0.0141
Stratified Independent Random Sample	0.0015	-0.0144
One-stage Cluster	-0.02	-0.01
Two-stage Cluster	-0.0119	-0.0087
Jackknife for stratified independent sample	0.0015	-0.0153
Jackknife for two-stage cluster sample	0.0119	-0.0105
Post-stratification of cluster sample	-0.0092	-0.0093
Post-stratification using two variables	0.0077	0.0099

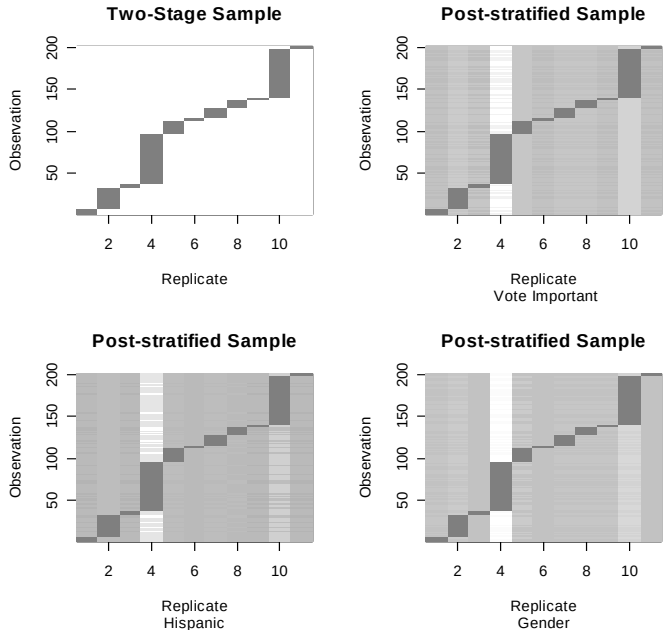

Figure 1: Replicate Weights for Two-Stage Cluster Sample With and Without Post-stratification

References

[1] Li-Chun Zhang. A Note on Post-Stratification When Analyzing Binary Survey Data Subject to Nonresponse. In **Journal of Official Statistics**, Volume 15, No. 2 1999, pp 329-334.

[2] Bethlehem, J.G. (1988). Reduction of Nonresponse Bias through Regression Estimation. Journal of Official Statistics, 4, 251 - 260.

[3] Smith, T.M.F. (1991). Post-stratification. The Statistician, 40, 315 - 323.

[4] Lohr, Sharon (1999). Sampling: Design and Analysis.

[5] Lumley, Thomas. (2004). Analysis of Complex Survey Samples. Avaliable online at http://www.jstatsoft.org/v09/i08/paper.pdf.