Sample Design and Weighting Procedures for Bulgaria Crisis Monitoring Survey (CMS) and the Bulgarian Longitudinal Inclusive Society Survey (BLISS)

David J. Megill
Statistical Consultant, World Bank
May 2014

1. Background

The Bulgaria Crisis Monitoring Survey (CMS) was a multi-topic panel survey conducted in February 2010, October 2010 and February 2011. A longitudinal analysis is used for tracking the impact of the economic crisis over time. The CMS collected information on various demographic and socioeconomic characteristics of the households and individual household members, including questions on labor market participation and earnings, access to and receipts from social protection programs, informal safety nets and remittances. The survey focused on how Bulgarian households are affected by and are coping with the economic crisis, and includes information on informal employment, reduced spending, postponement of investments, the sale of household assets, and the reliance on formal and informal credit. A national sample of 2,400 households was selected for the main CMS, based on a stratified multi-stage sample design. Given the special need to study the more vulnerable ethnic minority Roma population, an independent "booster sample" of 300 households was selected in settlements and neighborhoods identified as predominantly Roma.

The Bulgarian Longitudinal Inclusive Society Survey (BLISS) was conducted in March and April 2013, as a continuation of the longitudinal CMS, with an additional module about the skills of the adult population in Bulgaria. The main purpose of BLISS is to analyze the major barriers to activation (such as skills gaps and mismatches, informational asymmetries, and/or disincentives inherent in the tax-benefit schemes) for different groups (such as women, older workers, and Roma) through an understanding of the labor markets' behavior. The BLISS was conducted using the same panel of households as the CMS for the main survey and the "booster" (Roma) survey. Therefore the BLISS will be treated as an additional wave of the CMS in the longitudinal analysis, as well as a cross-sectional survey. The results from the CMS and BLISS data will be disaggregated by income distribution, and by ethnic majority versus minorities (including the Roma booster).

The purpose of this report is to document the CMS and BLISS sample design, and procedures used for calculating the weights for the longitudinal and cross-sectional tabulations and analysis. The weighting methodology was developed in collaboration with the World Bank team, consisting of Kristen Himelein, Alessandra Marini, Abla Safir and Sylvia Guallarartal. Kristen Himelein provided early guidelines for the terms of reference, and later provided valuable input and review for the final weights. The team also consulted regularly with Boyan of the Open Society Institute (OSI), Bulgaria. Their collaboration is highly appreciated.

2. Sample Design for CMS Main Survey

A stratified three-stage sample design was used for the CMS Main Survey. The primary sampling units (PSUs) selected at the first stage were the "settlements", or administrative units that conceptually cover all of the population of Bulgaria. The first stage sampling frame consisted of a database of all the settlements in Bulgaria, with information on the population in each settlement. The sampling frame was stratified by district (NUTS 2) and type of settlement. Bulgaria is divided into 28 administrative districts or provinces. Within each province, the settlements were further stratified by type. Three categories were defined for the type of settlement: rural, metropolitan (cities or towns with a population of 50,000 or more), and other urban (towns with a population less than 50,000). Therefore a total of 84 strata were defined for the sampling frame. The number of PSUs to be selected in each stratum was allocated proportionally to the population. The settlements within each stratum were selected with probability proportional to size (PPS), where the measure of size was based on the total population of the settlement from the sampling frame. Table 1 shows the distribution of the total population in the sampling frame by province and type of settlement. It should be noted that this sampling frame was developed for the CMS baseline survey prior to the 2011 Census. The population figures in Table 1 were based on population projections that were available in early 2010. The total population in this frame (7,996,282) is considerably higher than the corresponding total population enumerated in the January 2011 Bulgaria Census (7,364,570). This issue is discussed later in this report.

At the second sampling stage, voting stations (clusters) were selected in the sample settlements for the main survey, with PPS based on the number of registered voters. There is no information on the total population or number of households in each voting station, but these should be highly correlated with the number of voters. A total of 240 clusters were selected for the CMS Main Survey. In the case of large self-representing cities or settlements that were selected more than once at the first sampling stage, the number of clusters to be selected in each settlement was based on the number of sample "hits"; for example, Sofia was selected 40 times in the main sample and therefore has 40 sample clusters. This procedure results in a proportional allocation of the sample to the larger settlements selected with certainty at the first sampling stage. For the smaller non-self-representing settlements (not selected with certainty at the first stage), one voting station was selected in each settlement.

For the third sampling stage a random sample of addresses of individual voters was selected from the electoral database for each sample voting station. A sample of 10 addresses was selected in each sample voting station so that the corresponding households could be interviewed, and an additional 10 addresses were randomly selected as a reserve of replacements. When an original sample household could not be interviewed for any reason (including addresses with vacant houses), a replacement household from the reserve sample was interviewed. It should be pointed out that the addresses of households with more than one registered voter could appear multiple times in the database. Since the list of addresses for the sample voting stations was not unduplicated, this affects the probabilities of selection and results in a slight bias. The weighting procedures were adjusted to reduce this bias, as described in the section on the weighting procedures.

Table 1. Distribution of Total Population in Sampling Frame for CMS Baseline Survey, by Province, Urban and Rural Stratum

Province	Total Population in CMS Sampling Frame			
	Rural	Other Urban	Metropolitan	Total
Blagoevgrad	137,667	129,952	77,197	344,816
Burgas	122,116	119,628	204,612	446,356
Varna	84,964	63,402	351,809	500,175
Veliko Tarnovo	88,728	127,946	72,105	288,779
Vidin	41,768	-	57,320	99,088
Vratsa	84,689	56,572	68,417	209,678
Gabrovo	26,716	47,635	66,175	140,526
Dobrich	67,656	43,502	103,094	214,252
Kardzhali	99,581	22,631	51,000	173,212
Kyustendil	46,132	57,711	51,277	155,120
Lovech	58,186	100,849	-	159,035
Montana	57,517	105,697	-	163,214
Pazardzhik	110,416	111,423	79,654	301,493
Pernik	30,838	26,656	84,594	142,088
Pleven	101,667	84,495	122,487	308,649
Plovdiv	177,976	129,798	433,098	740,872
Razgrad	74,132	67,074	-	141,206
Ruse	61,768	34,349	168,018	264,135
Silistra	72,042	63,149	-	135,191
Sliven	71,951	44,372	103,918	220,241
Smolyan	57,802	71,454	-	129,256
Sofiya	55,305	-	1,266,746	1,322,051
Sofiyska	97,514	157,938	-	255,452
Stara Zagora	107,956	55,489	207,393	370,838
Targovishte	67,189	74,070	-	141,259
Haskovo	77,450	111,040	81,083	269,573
Shumen	78,696	39,064	95,035	212,795
Yambol	43,830	19,518	83,584	146,932
Total	2,202,252	1,965,414	3,828,616	7,996,282

In several sample rural settlements there were no street names, no household names and no other means for identifying the exact address of the residents. In these cases the sample addresses were randomly selected using a Global Positioning System (GPS) device. This procedure is described in the next section on the sample design for the Booster Survey.

3. Sample Design for CMS Booster Survey of Roma Communities

There is a special interest in studying the ethnic minority population that is predominantly Roma, given that this population group is generally poorer and has more challenges integrating into the

formal labor force. Since the proportional distribution of the main CMS sample would result in a relatively small sample of Roma households, it was decided to have a separate "booster" sample from a special frame of communities with a concentration of Roma households. The sampling frame was based on a list of communities or neighborhoods throughout Bulgaria with a predominantly Roma population, identified by experts who are knowledgeable about this minority population. The sampling frame database includes information on the approximate total population and Roma population in each neighborhood, as well as the corresponding geographic information and type of settlement. The frame includes a total of 889 Roma neighborhoods, with an estimated total population of 880,767 and a Roma population of about 729,498, so the population in this frame is estimated to be about 82.8% Roma.

A two-stage sample design was used for the Booster Survey. The PSUs or clusters were defined as the individual Roma neighborhoods identified in the sampling frame for all of Bulgaria. This frame was not stratified. A total of 30 sample neighborhoods were selected in 20 districts at the first sampling stage with PPS, where the measure of size was based on the estimated Roma population of each neighborhood in the frame.

For the second stage of selection, there was no frame of addresses available for the 30 sample Roma neighborhoods. Therefore it was necessary to use a GPS sampling method for selecting the households in each sample neighborhood at the second stage. The selection of households involved the following steps:

1. The geographical coordinates of the four framing points of the sample neighborhood were identified. These framing points are the most northerly, westerly, easterly and southerly points of the residential area. A rectangle surrounding the sample neighborhood is formed by connecting these four points.

2. A random sample of 20 geographic coordinates within the neighborhood rectangle is generated; each coordinate is determined by its longitude and latitude. If some of the random points selected in the rectangle are outside the boundaries of the neighborhood, they were removed from the list and new random coordinates were generated. For each selected coordinate, the nearest "door" (that is, dwelling unit) is identified. Details about the location and description of each selected dwelling unit were recorded so that the interviewer could find it in the field. A list of 20 sample dwelling units was selected in this way, including 10 for the original sample and 10 for the reserve sample households for replacement.

3. If there were apartment buildings in the neighborhood, people who live there would have a smaller chance of being selected than those who live in a separate house, since the probability of selection of each household is conceptually proportional to the distance to the nearest "door". For this reason the households in sample neighborhoods with both apartment buildings and individual houses were divided into two respective parts. A list was made of all the apartments with the estimated number of people, and it was also necessary to estimate the population living in individual houses within the sample neighborhood. The sample of 10 households and 10 reserve households was allocated proportionally to the two parts. Then two independent samples of households were

selected: one for the individual houses (GPS sample), and one for the apartments (simple random sample from the list of apartments).

It should be noted that using this type of GPS sampling procedure for selecting the households results in different probabilities by sample household depending on the space between each house and the next "door" (for example, based on the size of the yard). Since it is not possible to calculate these differential probabilities, it is necessary to calculate the weights based on an assumption that the households within each cluster are selected with equal probability, so the results will be affected by a corresponding small bias. In the case of the households living in apartments, this bias was reduced by listing these households and selecting this portion of the sample with equal probability.

4. Panel Households and New Sample Households for Cross-Sectional Data

In the case of the baseline CMS Main Survey and Booster Survey, the original sample households that could not be interviewed were replaced by households from the reserve sample. The combination of original sample and replacement households in the CMS1 baseline survey became the panel households that would be followed each subsequent wave. Beginning with the second wave, any panel household that could not be interviewed for any reason was replaced by a household from the reserve sample for that cluster following the baseline survey. After all of the 10 households in the reserve sample were used, then a new random sample of 10 households was selected from the voter registration list (or based on random GPS coordinates) to provide a new set of reserve sample households for replacement. Any replacement households included in the sample after the baseline survey are considered part of the cross-sectional sample for the corresponding wave, but are not part of the panel for the longitudinal analysis.

5. Weighting Procedures for Baseline CMS1 Main Survey

In order for the sample estimates from the CMS and BLISS to be representative of the population, it is necessary to multiply the data by a sampling weight, or expansion factor. Since the sample households for the baseline survey became the panel of households that was followed in each subsequent wave, the weights for the baseline CMS1 are the basis for calculating the weights for the subsequent waves, with an adjustment for the attrition in each wave of the panel survey. The cross-sectional survey data for each wave also have the same basic weights, which are adjusted taking into account the new replacement households. The cross-sectional weights were later adjusted based on population projections for the data collection period of the corresponding wave, as described later in this report.

The calculation of the weights depends on the different sampling stages. The basic weight for each sample household would be equal to the inverse of its probability of selection (calculated by multiplying the probabilities at each sampling stage). An Excel spreadsheet was used to maintain the information from the sampling frame for each sample cluster in the Main Survey and the Booster Survey, with formulas for calculating the sampling probabilities at each stage of selection and the corresponding overall weight.

Based on the stratified three-stage sample design described previously, the overall approximate probability of selection for the baseline sample households in the CMS1 Main Survey can be calculated as follows:

$$p_{hijk} = \frac{n_h \times P_{hi}}{P_h} \times \frac{n_{hi} \times V_{hij}}{V_{hi}} \times \frac{n_{hij}}{V_{hij}} \times m_{hijk(18+)} = \frac{n_h \times P_{hi}}{P_h} \times \frac{n_{hi}}{V_{hi}} \times n_{hij} \times m_{hijk(18+)},$$

where:

p_{hijk} = probability of selection of the k-th baseline sample household in the j-th sample voting station of the i-th sample settlement in stratum (province by settlement type) h

n_h = number of sample settlements selected in stratum h

P_{hi} = population in the frame for the i-th sample settlement in stratum h

P_h = population in the frame for stratum h

n_{hi} = number of sample voting stations selected in the i-th sample settlement in stratum h

V_{hij} = number of registered voters (measure of size) in the sampling frame for the j-th sample voting station of the i-th sample settlement in stratum h

V_{hi} = total number of registered voters in the sampling frame for the i-th sample settlement in stratum h

n_{hij} = number of sample households with completed baseline interviews in the j-th sample voting station of the i-th sample settlement in stratum h (generally equal to 10)

$m_{hijk(18+)}$ = number of persons 18 years and older in the k-th household in the j-th sample voting station of the i-th sample settlement in stratum h

The basic baseline weights for the Main Survey will be the inverse of this probability of selection, expressed as follows:

$$W_{hijk} = \frac{1}{p_{hij}} = \frac{P_h \times V_{hi}}{n_h \times P_{hi} \times n_{hi} \times n_{hij} \times m_{hijk(18+)}} = \frac{P_h \times V_{hi}}{n_h \times P_{hi} \times n_{hi} \times n_{hij}} \times \frac{1}{m_{hijk(18+)}},$$

where:

W_{hijk} = basic weight of the k-th baseline sample household in the j-th sample voting station of the i-th sample settlement of stratum h

As pointed out previously, the households with multiple voters may be duplicated in the register of voters, since the addresses are repeated for each voter. As a result, the probabilities vary by household. Since no information is available on the number of voters in each household, we used the number of household members 18 years and older to approximate the number of times the household appears in the list of addresses for registered voters. Apparently the staff selecting the addresses made sure that the same household was not selected twice in the sample, so the exact probabilities are not known. However, this weighting formula based on the number of household members 18 years and older is considered to be the least biased estimate of the approximate weight.

Since the basic weights for the baseline sample households within a cluster vary by the number of members 18 years and older in each household, the last expression for the weight has a separate "cluster weight component" that is the same for all sample households in the cluster, defined as follows:

$$CCW_{hij} = \frac{P_h \times V_{hi}}{n_h \times P_{hi} \times n_{hi} \times n_{hij}} = \text{ cluster component of the weight}$$

The baseline weight for each household was then calculated by dividing this cluster component of the weight by the number of household members 18 years and older in each sample household for CMS1. The baseline weights were calculated at the household level in this way. All the individuals in a sample household have the same baseline weight. The "cluster weight component" was also used for the calculation of the cross-sectional weights for each wave, since the same 240 sample clusters for the main CMS are used for all waves. The panel weights for each wave were adjusted separately to take into account attrition.

6. Weighting Procedures for "Booster" Baseline CMS1

Based on the sampling procedures for the Booster Survey described previously, the approximate overall probability of selection for the "booster" sample households within a particular sample neighborhood can be expressed as follows:

$$p_{Bi} \approx \frac{n_B \times P_{Ri}}{P_R} \times \frac{n_{Bi}}{M'_{Bi}},$$

where:

p_{Bi} = probability of selection of the sample households in the i-th selected neighborhood in the Roma "booster" sampling frame

n_B = number of sample neighborhoods selected from the sampling frame for the Booster Survey (that is, 30)

P_{Ri} = estimated total Roma population (measure of size) for the i-th sample

neighborhood from the "booster" sampling frame

P_R = estimated total Roma population in sampling frame for booster survey (729,498)

n_{Bi} = number of sample households with completed baseline interviews in the i-th sample neighborhood from the "booster" sampling frame (generally equal to 10)

M'_{Bi} = approximate total number of households in the i-th sample neighborhood in the Booster Survey

In the case of any sample neighborhood that has a Roma population greater than the sampling interval (729,498/30 = 24,317), the first stage probability would be equal to 1.

The weight for the booster sample households would be the inverse of this probability of selection, expressed as follows:

$$W_{Bi} \approx \frac{P_R \times M'_{Bi}}{n_B \times P_{Ri} \times n_{Bi}},$$

where:

W_{Bi} = basic design weight for the sample households in the i-th selected neighborhood in the Roma "booster" sampling frame

Since we do not have any information on the number of households in each of the 30 "booster" sample neighborhoods (M'_{Bi}), it was necessary to estimate these values from the information in the frame for the estimated population in each neighborhood. It should be pointed out that the estimate of total population for each neighborhood is different from the total Roma population used as the measure of size. In order to estimate the number of households in each neighborhood, it is necessary to divide the total population for that neighborhood by an average number of persons per household. After discussing the options for estimating the average household size, it was decided that it would be reasonable to use the average number of persons in the 10 selected households in each sample neighborhood. Although the estimate of average household size for each neighborhood is subject to sampling error, it is an approximately unbiased estimate of the actual value. The average household size varies by neighborhood, since some areas may have more families with larger households, and other areas may have mostly adult migrating workers with smaller households, for example. The distribution of the number of persons per sample household was examined for the 30 "booster" sample clusters. The largest household has 12 persons, and no extreme values were found in the data. The average household size varies by cluster from 1.9 to 6.0. In this case the value of M'_{Bi} can be estimated as follows:

$$M'_{Bi} \approx \frac{P_{Bi}}{\bar{n}_{Bi}},$$

where:

P_{Bi} = estimated total population for the i-th sample neighborhood from the "booster" sampling frame

$\bar{n}_{Bi}$ = average number of persons per household calculated from the sample households in the i-th sample neighborhood from the Booster Survey data

7. Adjustment of the Panel Individual Weights for Attrition

The longitudinal analysis involves using the baseline CMS data for individuals from panel households that are successfully interviewed again in the different survey waves. This will allow a study of the trends in household-level data over time. Some of the sample panel households could not be interviewed for a particular wave because they moved, were temporarily absent or refused to be interviewed again; this is referred to as attrition. In the case of entire households that moved, there was an attempt to track and interview them. In order to ensure that the survey results from the panel households for each wave are representative of the frame for the baseline CMS, it is necessary to adjust these weights for attrition. One alternative for adjusting the weights for nonresponse is to multiply the weights at the cluster level by the inverse of the response rate. This is based on the assumption that the characteristics of non-interview households are similar to those of households that are interviewed. However, if certain types of households or individuals have a lower probability of responding, this would result in a corresponding bias. In order to reduce this potential bias, it was decided to use a logistic regression analysis to determine the probabilities of individuals to respond based on their characteristics, by generating corresponding response probability scores. The logit regression model used for this analysis is similar to the methodology used for the Tanzania LSMS-ISA Panel Survey, described in the report "Weight Calculations for Panel Surveys with Sub-Sampling and Split-off Tracking" (Kristen Himelein, Policy Research Working Paper No. 6373, The World Bank, Development Research Group, Poverty and Inequality Team, February 2013). The baseline weights for the sample panel individuals were adjusted for attrition for the second and third waves of the CMS and for the BLISS, as well as for a data set for panel households included in all waves. The attrition analysis for calculating the individual panel weights for each wave involved using a data set that included the baseline data for all panel individuals.

The weight adjustment factors for attrition were calculated using a stepwise logistic regression based on a response propensity model, including the household and individual characteristics measured in the baseline as covariates. The following characteristics were included as independent variables in the model: region, gender, age, education, ethnicity, religion, household size, education of head of household, whether individual has a spouse, and income quartile. The Stata version 12 software was used for this analysis. The syntax for this analysis and the output reports are presented in Annexes A and B. The propensity score represents the probability that an individual with a particular combination of characteristics would be successfully interviewed, and is therefore a measure of the response rate at the micro level. By multiplying the basic weight by the inverse of this response rate, the weighted estimates reduce the bias due to some groups being under-represented in the sample due to nonresponse. The logistic response propensity scores were calculated separately for the data from the CMS2, CMS3 and BLISS panels of sample individuals for the Main Survey and the Booster Survey. Another attrition analysis was run for the panel households interviewed in all waves of each survey.

The original data file for individuals in the CMS1 baseline survey had missing values for some of the variables included in the logistic regression model. The World Bank team made an effort to impute many of the missing values, but the final data file used for the logistic regression still had missing values for some variables. As a result, there were some individuals with missing response probability scores in the output file from the logistic regression analysis. In this case it was necessary to impute the missing probability scores. When ps values were available for other members of the same household, the average value for all individuals in that household was used as the imputed value. Since several of the dependent variables in the logistic regression were household-level characteristics, there was not much variability in the ps values of household members. However, when ps values were missing for all household members, the average ps value for individuals at the cluster level was used. The spreadsheet used for this imputation of the missing ps values was provided to the World Bank team. Table 2 shows the number of missing (and imputed) ps values by wave.

Table 2. Number of Missing ps Values for Individuals Imputed by Wave Following the Logistic Regression Analysis

	Missing ps Values	
Wave	Main Survey	Booster Survey
CMS2	56	6
CMS3	60	4
BLISS	44	6
Combined	71	5

Given the relatively small number of missing ps values, this imputation should not have much effect on the results of the adjustment of the individual panel weights for attrition. The preliminary individual panel weights for each wave were calculated as the baseline panel weight defined previously divided by the corresponding ps value. These weights were later adjusted based on population estimates by province, urban and rural strata, as described later in this report.

8. Cross-Sectional Individual Weights for Each Wave

For the cross-sectional analysis of the data for each wave we will use the data from all households with completed interviews. The sample households in each cluster are divided into panel (CMS1 baseline) and non-panel (new) households. Since we would like take advantage of the adjustment of the panel weights for attrition in calculating the cross-sectional weights for each wave, it is necessary to have separate weighting procedures for the sample panel and non-panel households within the cross-sectional sample for each cluster.

The panel weights are designed to represent the full frame, while for the cross-sectional survey the panel households will only represent a certain proportion of the frame; the rest of the frame is represented by the new (non-panel) sample households. Therefore it is first necessary to determine the proportion of panel households in the cross-sectional sample for each sample

cluster for a particular wave. In the case of the Main Survey, this proportion is calculated as follows:

$$p_{whi} = \frac{n_{wphi}}{n_{whi}},$$

where:

p_{whi} = proportion of panel households among the cross-sectional households for the i-th sample cluster in stratum h for wave w

n_{wphi} = number of panel households with completed interviews in wave w for the i-th sample cluster in stratum h

n_{whi} = number of cross-sectional households with completed interviews in wave w for the i-th sample cluster in stratum h

This proportion would then be applied to all the panel individual weights in the sample cluster for the particular wave, as follows:

$$W'_{cwhijk} = W'_{whijk} \times p_{whi},$$

where:

W'_{cwhijk} = cross-sectional weight for the k-th panel individual in the j-th panel household in the i-th sample cluster of stratum h for wave w

W'_{whijk} = panel weight for the k-th individual in j-th panel household in the i-th sample cluster of stratum h for wave w

The cross-sectional weight for the individuals in the non-panel (new) sample households within a cluster would be calculated as the inverse of the overall probability of selection. This weight can be calculated by adjusting the baseline weight for the households in the sample cluster as follows:

$$W_{cwhij} = CCW_{hi} \times \frac{n_{bhi}}{n_{whi}} \times \frac{1}{m_{whij(18+)}},$$

where:

W_{cwhij} = cross-sectional weight for the individuals in the j-th non-panel (new) sample household in the i-th sample cluster of stratum h in wave w

CCW_{hi} = cluster component of the weight (defined previously) for the i-th sample cluster in stratum h

n_{bhi} = total number of CMS1 baseline sample households in the i-th sample cluster of stratum h (including those without data in wave w)

n_{whi} = number of cross-sectional households with completed interviews in wave w for the i-th sample cluster of stratum h

$m_{whij(18+)}$ = number of persons 18 years and older in the j-th household in the i-th sample cluster of stratum h from data for wave w

The second term in this weight is to compensate for the difference in the number of sample households in the cluster between the baseline survey and the cross-sectional sample for a particular wave.

A simple example will be useful to illustrate the calculation of the cross-sectional weights. Let us assume that we have a cluster of 100 households and we select 5. In order to simplify this example, let us assume that the number of household members 18 years and older is the same for each of these households, so the basic weight for each household would be 20 for the baseline survey. In this example two panel households drop out for a particular wave. We do the propensity score attrition adjustment for the panel weights, and we end up with the following adjustments of the weights for the three panel households interviewed that wave:

(20*1.2) + (20*1.8)+(20*2) = 100

Now let us assume that two new households were selected for this wave to replace the two panel households that dropped out, so we have a cross-sectional sample of 5 households. Let us also assume that the number of households members 18 years and older does not change from the baseline to this wave. Using the cross-sectional weighting procedures described above, we would multiply the panel weights adjusted for attrition by a factor of 3/5 (that is, 0.6), and the replacements would have a weight of 20. In this case, the weighted total would be:

0.6*(20*1.2)+0.6*(20*1.8)+0.6*(20*2)+(20*2) = 100

It can be seen in this simple example that these weighting procedures which use the panel weights to represent the corresponding proportion of the frame are unbiased.

9. Adjustment of the Cross-Sectional Weights to Account for Strata with No Sample

In reviewing the distribution of the households interviewed for the cross-sectional survey each wave, it was found that some waves do not have data for individual sample clusters, or for entire strata. For the panel weights of each wave this is not a problem, since the weight adjustment factors based on the response probability scores were based on the full panel from the CMS1 baseline. However, in the case of the cross-sectional weights, it was necessary to adjust the overall basic weights by province, urban/rural strata to take into account any baseline sample clusters without data for that wave. In the case of an entire stratum without data, the basic weights had to be adjusted at the higher regional, urban and rural level. There was also a case

where a sample cluster not included in the baseline was arbitrarily added to a particular wave (it was explained as a type of "replacement"), in which case the weights had to take into account the additional sample cluster for the corresponding stratum. The purpose of this section is to document each of these cases.

In the case of the CMS2 Main Survey cross-sectional sample, the following adjustments were made to the basic weights:

- Sample cluster 0703 had no data. This cluster belonged to the Gabrovo metropolitan stratum, which only had 2 sample clusters. Since one cluster was missing, the basic weight for the other cluster (0704) in this stratum was multiplied by 2.

- All 3 sample clusters for Vidin province in Region 1 (North West) had no data. Cluster 0501 was rural and clusters 0502 and 0503 were metropolitan. Since both strata in Vidin province had no data, it was necessary to adjust the basic weights at the regional level. The CMS1 baseline weights were used to tabulate the weighted population for Region 1 with and without these three missing clusters. At same time, cluster 0609, which was not in the CMS1 baseline survey, was arbitrarily added to the Region 1 sample for the cross-sectional survey for CMS2. In this case, all the cross-sectional basic weights for Region 1 were multiplied by the following factor:

$$A_{R1} = \frac{\hat{P}_{R1}}{\left(\hat{P}_{R1} - \hat{P}_{(0501,0502,0503)}\right) + P_{0609}} = \frac{992145}{867282 + 56572} = 1.073921,$$

where:

A_{R1} = adjustment factor for basic CMS2 main survey cross-sectional individual weights for Region 1

$\hat{P}_{R1}$ = estimated total population in Region 1 using CMS1 baseline weights

$\hat{P}_{(0501,0502,0503)}$ = weighted total population for clusters 0501, 0502 and 0503 using CMS1 baseline weights

P_{0609} = population for cluster 0609 based on sampling frame for Main Survey

- Sample cluster 1601 had no data. This cluster belonged to the Plovdiv (Province 16) rural stratum, which has five sample clusters. In this case the following adjustment factor was used for the basic cross-sectional weights of this stratum:

$$A_{P16r} = \frac{\hat{P}_{P16r}}{\hat{P}_{P16r} - \hat{P}_{(1601)}} = \frac{190893}{190893 - 43276} = 1.293167,$$

where:

A_{P16r} = adjustment factor for basic CMS2 main survey cross-sectional individual weights for the rural stratum of Plovdiv (Province 16)

$\hat{P}_{P16r}$ = estimated total population in the rural stratum of Plovdiv (Province 16) using CMS1 baseline weights

$\hat{P}_{(1601)}$ = weighted total population for cluster 1601 using CMS1 baseline weights

For the CMS3 Main Survey cross-section, the following adjustment was made to the basic weights:

- Cluster 0609 that was not in the CMS1 baseline survey was arbitrarily added to the Region 1 sample for the cross-sectional survey for CMS3. In this case, all the cross-sectional basic weights for Region 1 were multiplied by the following factor:

$$A_{R1} = \frac{\hat{P}_{R1}}{\hat{P}_{R1} + P_{0609}} = \frac{992145}{992145 + 56572} = 0.9460561,$$

where:

A_{R1} = adjustment factor for basic CMS3 main survey cross-sectional individual weights for Region 1

10. Adjustment of the Baseline, Panel and Cross-Sectional Individual Weights for the Main Survey Based on Population Projections

When the total weighted population was calculated based on each set of baseline, panel and cross-sectional weights for the Main Survey, the total population was generally higher than 8 million, compared to a total population of less than 7.4 million in the 2011 Census. The main reason for these higher estimates is that the pre-Census sampling frame used for selecting the sample settlements and clusters for the CMS1 baseline survey was based on higher population projections, as shown in Table 1. The estimated total population in that sampling frame was about 8 million, so the different sets of weights reflected that frame. The sampling frame appeared to have over-estimated the population across all the provinces and urban/rural strata, although the percent difference varied by stratum. In order to make all sets of individual weights

consistent with the population estimates based on the 2011 Census, it was necessary to calculate weight adjustment factors using population estimates based on the Census results. Kristen Himelein used the total urban and rural population by province from the 2001 and 2011 Censuses to estimate the monthly population change rate. This population change rate was then used to estimate the population by province, urban and rural stratum for each month between January 2010 and December 2013. In this way we obtained the estimated population distribution by stratum for the data collection month of the CMS1 baseline and each subsequent wave of CMS and BLISS. The February 2010 population estimates were used for the CMS1 baseline. Since the reference population for the panel surveys is the same as the CMS1 baseline, the population estimates for February 2010 were also used for adjusting the panel weights for each wave of the Main Survey.

In the case of the cross-sectional weights for the Main Survey, the weights were adjusted based on the population estimates for the month of the data collection. The population estimates for October 2010 were used for adjusting the CMS2 cross-sectional weights, and February 2011 for CMS3. In the case of BLISS, the data collection was conducted in March and April 2013, so the average population by stratum for those two months was used for adjusting the cross-sectional weights of that survey.

The weight adjustment factor by province, urban and rural stratum for each set of weights for the Main Survey is defined as follows:

$$A_h = \frac{\hat{P}_h}{\sum_{i \in h} \sum_j \sum_k W_{hijk}},$$

where:

A_h = adjustment factor for province, urban/rural stratum h

$\hat{P}_h$ = estimate of population for stratum h based on projections using census data; in the case of the CMS1 baseline and panel surveys, the population estimates are for October 2010, and for cross-sectional weights, population estimate for data collection month of each wave

$\sum_{i \in h} \sum_i \sum_j W_{hijk}$ = sum of weights for all sample individuals in stratum h from data for corresponding survey and wave

When these weight adjustment factors were first calculated, it was found that the rural and urban factors for the provinces of Sofia and Sofia District were very high or low compared to the other factors. Therefore it was necessary to examine the population distribution in the CMS sampling frame for these provinces, identified as Sofiya and Sofiyska; these are shown in Table 1. In comparing the rural and urban population in the 2011 Census for these provinces to the corresponding figures in the frame, the extreme differences indicated that apparently the

classification of geographic areas for these provinces in the frame were not consistent with those in the Census. However, when these two provinces were combined, the rural and urban distribution of the population was more consistent with the Census. For this reason Sofia and Sofia District were combined for the adjustment of the weights by urban and rural strata.

In the case of CMS/BLISS waves that did not have any data for a particular province, urban/rural stratum, it was necessary to combine the strata to the region, urban/rural level for the adjustment of the weights. In this case the formula for the adjustment factor is the same, except the subscript h refers to the region, urban/rural stratum. The World Bank team was provided with a spreadsheet showing the calculation of the adjustment factors for all sets of the CMS/BLISS weights. For each set of CMS/BLISS panel and cross-sectional weights, Table 3 shows a list of regions where it was necessary to adjust the weights at the regional, urban/rural level. For the remaining regions the weights were adjusted at the province, urban/rural stratum level (with the exception of Sofia and Sofia District, which were combined).

Table 3. CMS/BLISS Weights for Main Survey Adjusted by Region, Urban and Rural Strata

Wave, Main Survey	Region
CMS2 Panel Survey	Region 1 Region 4 Region 6
CMS3 Panel Survey	Region 4 Region 6
BLISS Panel Survey	Region 4 Region 6
Combined Wave Panel Survey	Region 1 Region 4 Region 6
CMS2 Cross Sectional Survey	Region 1

The final weight for each sample individual in the Main Survey for each set of baseline, panel and cross-sectional weights was equal to the corresponding preliminary individual weight multiplied by the corresponding weight adjustment factor for that stratum and survey.

11. Adjustment of the Baseline, Panel and Cross-Sectional Individual Weights for the Booster Survey Based on Population in Frame

In the case of the booster sample for the CMS, the sampling frame consists of a list of predominantly Roma neighborhoods throughout Bulgaria. There are no census population estimates available for the areas included in this frame, and the small sample of 30 clusters does not cover all the provinces. Therefore it was decided to adjust all the weights based on the total population figures in the frame, separately for the urban and rural strata. The total population in the frame was 306,625 for the rural stratum and 574,142 for the urban stratum, for a total population of 880,767. The CMS1 baseline weights and each set of panel and cross-sectional weights for the Booster Survey were adjusted using the following factor:

$$A_h = \frac{\hat{P}_h}{\sum_{i \in h} \sum_j \sum_k W_{hijk}},$$

where:

A_h = weight adjustment factor for booster sample in national urban/rural stratum h

$\hat{P}_h$ = total population in booster sampling frame of Roma neighborhoods for stratum h; this population is 306,625 for the rural stratum and 574,142 for the urban stratum

$\sum_{i \in h} \sum_i \sum_j W_{hijk}$ = sum of weights for all sample individuals in stratum h from data for the corresponding wave of the Booster Survey

The final weight for each sample individual in the Booster Survey for each set of baseline, panel and cross-sectional weights was equal to the corresponding preliminary individual weight multiplied by the corresponding weight adjustment factor for that stratum and survey.

12. Calculation of CMS/BLISS Household Weights

In order for the households weights for the CMS/BLISS baseline, panel and cross-sectional surveys to be consistent with the corresponding individual weights for these surveys, each household weight was calculated as the average of the corresponding final adjusted weights for the individual members in that household. In this way the household weights also benefit from the adjustment of the individual weights for attrition as well as consistency with the total population estimates by stratum. Given slight differences in the household size by wave, the weighted total number of households will vary slightly by survey accordingly, but this should not be a problem for the analysis. The projected total number of households per stratum for each wave was not available for adjusting the household weights further. The procedure for calculating the household weights was the same for the CMS/BLISS Main Survey and the Booster Survey.

In the case of the panel weights for households, it was found that 1 panel household in CMS2 and 4 households in CMS3 did not have any individuals in the panel, so it was not possible to calculate the panel weight for those households. Conceptually it seems that by definition a panel household should have at least one household member that was in the original baseline panel household. If all the members have changed, then this should probably be considered a new household. The World Bank team may want to examine these 5 panel households further, but if they are dropped from the sample for the panel analysis for the affected waves, it should not make much difference in the survey results.

13. Calculation of the Cross-Sectional and Panel Weights for the BLISS Skills Module

In the case of BLISS, a skills module was included for adults age 18 to 65 years. However, only one eligible person in this age range was randomly selected in each sample household to be interviewed. In this case the overall probability of selection would be equal to the probability of selection of the household times the probability of selecting one eligible person in the household. The weight would be the inverse of this probability of selection, which can be defined in general as follows:

$$W_{SMhij} = W'_{hij} \times m_{hij(18-65)},$$

where:

W_{SMhij} = weight for the individual with a completed BLISS skills module in the j-th sample household in the i-th sample cluster of stratum h (for cross-sectional or panel sample of the Main Survey or Booster Survey)

W'_{hij} = final BLISS cross-sectional (or panel) household weight for Main Survey (or Booster Survey) in the j-th sample household in the i-th sample cluster of stratum h

$m_{hij(18-65)}$ = number of eligible household members age 18 to 65 years in the j-th sample household in the i-th sample cluster of stratum h for BLISS Main Survey or Booster Survey

Since there are BLISS sample households with members age 18 to 65 years but without a completed skills module, it is necessary to adjust the basic skills module weight above for nonresponse. For the Main Survey, the weights were adjusted for nonresponse at the province, urban/rural stratum level. The nonresponse adjustment factor for each stratum was calculated as follows:

$$A_{NRh} = \frac{\hat{P}_{h(18-65)}}{\sum_{i \in h} \sum_{j} \sum_{k} W_{SMhij}},$$

where:

A_{NRh} = nonresponse weight adjustment factor for the BLISS skills module weights in stratum h (for cross-sectional or panel sample of the Main Survey)

$\hat{P}_{h(18-65)}$ = weighted estimate of total population age 18 to 65 years in stratum h based on the BLISS final cross-sectional (or panel) weights for the Main Survey

W_{SMhij} = basic weight for the individual with a completed BLISS skills module in the j-th sample household in the i-th sample cluster of stratum h

The numerator of this weight adjustment factor corresponds to the weighted total number of persons age 18 to 65 years in stratum h based on the final BLISS cross-sectional or panel weights (depending on the particular data set). The denominator is the sum of the basic skills module weights for all the completed interviews.

As in the case of the weight adjustment based on population estimates, it was necessary to calculate the adjustment factors at the regional, urban/rural level when there were no BLISS skills module data for a particular province, urban/rural stratum within the region. Specifically, the BLISS skills module panel weights were adjusted for Regions 4 and 6 at the urban and rural levels. In the remaining regions the weights were adjusted at the province, urban and rural levels.

For the Booster Survey the basic BLISS skills module weights for the cross-sectional and panel samples were adjusted at the national, urban and rural levels. The formula for the weight adjustment factor is the same as that specified above for the Main Survey, but the stratum h in this case refers to the national urban and rural strata.

The final skills module weights were calculated as the basic skills module weights defined previously times the nonresponse adjustment factor for the corresponding stratum.

14. Procedures Used for Calculating CMS/BLISS Individual and Household Weights

The different steps involved in producing the different sets of individual and households weights for the CMS/BLISS baseline, panel and cross-sectional surveys are presented in Annex A. The logistic regression for the attrition analysis was carried out using the Stata software. However, most of the tabulations, data aggregation and merging of variables from different files was done using the SPSS software, and some intermediary files were compiled in Excel spreadsheets. Annex A documents all of the different steps and files that were involved in the calculation of each set of weights, including examples of the Stata and SPSS syntax that were used.

Annex A. Steps Involved in Producing the Different Sets of CMS/BLISS Individual and Household Weights

The purpose of this Annex is to document all of the steps involved in producing the different sets of weights for each wave of the CMS/BLISS Main Survey and Booster Survey. Examples of the Stata or SPSS syntax used for some of these steps are also presented here. References are made to the different intermediary files that were submitted as part of the documentation of the weighting procedures.

1. The probabilities of selection of each sampling stage and the corresponding components of the CMS1 baseline weights for households and individuals were calculated using an Excel spreadsheet with information from the sampling frame for each sample cluster. These spreadsheets include the formulas used for calculating the probabilities and basic weights described in this report. The basic weights for the Main Survey were generated in the Excel spreadsheet "Calculation_weights_CMS_baseline_Main_Survey_final.xlsx", and the baseline weights for the Booster Survey were produced in the file "Calculation_weights_CMS_baseline_Booster_Survey_final.xlsx". Given that the methodology used for calculating the basic weights for the Main Survey involved calculating weights at the household level based on a "cluster weight component" and the number of household members 18 years and older, in this case the "cluster weight component" was calculated in the weighting spreadsheet for the Main Survey, and then merged in the data files for calculating the household-level weights.

2. The World Bank team provided two CMS/BLISS data files in a Stata format that were used in the process of generating the weights. The data file for individuals was named "individuals_weights_dataset_imputed_gv.dta", and the data file for households was named "household_weights_dataset_imputed_gv.dta". Each file included the variables from all the different waves for both the Main Survey and the Booster Survey. These files were used to extract a file for each survey that was used for the logistic regression analysis to adjust the panel weights for attrition, based on the procedures described in the report.

3. The Stata software was used for the logistic regression analysis of the panel data for each wave of the Main Survey and the Booster Survey. An example of the Stata syntax that was used is presented here. The first step involved extracting the CMS1 panel data for individuals from the data file "individuals_weights_dataset_imputed_gv.dta", using the following Stata syntax for the Main Survey:

   ```
   keep if CMS1==1 & sample_type==1
   ```

4. A separate panel data file was created for the logistic regression of each wave for the Main Survey and the Panel Survey. For each panel data set, a (0, 1) response variable was generated for the corresponding wave. For example, in the case of the attrition analysis for the CMS2 panel from the Main Survey, the syntax above was followed by:

   ```
   gen response = 0
   ```

```
replace response=1 if CMS2==1
```

This response variable was the dependent variable in the logistic regression for each wave of the Main Survey and the Booster Survey.

5. A stepwise logistic regression function of Stata was used to generate the propensity score corresponding to the probability of a response based on the individual and household characteristics in the model. The following syntax was used for the logistic regression:

```
xi: stepwise, pr(.2): logit response i.region i.settl_type age_1 i.gender_1
i.group_marital_status_1 i.group_education_1 i.ethnicity_1 i.group_religion_1
hhsize_1 income_quintile_1 i.group_head_education_1
i.group_spouse_education_1
predict ps
```

6. This analysis added various output variables to the individual panel data file for each wave, included in the corresponding files with the final individual panel weights for each wave, including the response probability score ps that was used for adjusting the individual panel weights for attrition. As indicated in the report, there were individual records with missing ps values because of missing data for particular independent variables included in the logistic regression model. The missing ps values were imputed based on the average ps value for other members of the same household, or with the average ps value for the cluster when there were no other individuals with data for that household. The imputation of the ps values was done using the Excel file "Impute_ps_CMS_BLISS.xlsx", which was provided to the World Bank team. The variable with the imputed values has the name ps2.

7. In order to calculate the baseline weight for individuals in the file with the panel weights for each wave of the Main Survey, the "cluster weight component" described in the report was merged in each weighting file (converted to an SPSS format), and divided by the number of persons 18 years and older in the CMS1 (num_hhm_18more_1), based on the formula described in the report. The following SPSS syntax was used:

```
COMPUTE  wt_CMS1_baseline_ms_ind= Weight_cluster_component_ms /
num_hhm_18more_1.
EXECUTE.
```

8. The preliminary panel weight for individuals was calculated by dividing the baseline weight for each individual record by the corresponding value of ps2. For example, in the case of the individual panel weights for the CMS2 Main Survey, the following SPSS syntax was used:

```
COMPUTE  wt_CMS2_ms_panel_ind= wt_CMS1_baseline_ms_ind / ps2.
EXECUTE.
```

9. As explained in the report, for calculating the individual cross-sectional weights for each survey and wave, it was necessary to determine the proportion of the cross-sectional sample households in each cluster that were in the panel. In the case of the panel households in the cross-sectional sample, the panel weights were multiplied by this proportion to represent the corresponding proportion of the frame. The weights were calculated separately for the individuals in the new sample households, using the cluster component of the weight divided by the number of persons 18 years and older in the household. However, it was also necessary to adjust the cluster component of the weight by the ratio of the number of sample CMS1 baseline households in the cluster and the number of cross-sectional households in the cluster. Therefore it was necessary to compile a cluster-level file with the relevant summary data from the CMS1 and cross-sectional household data to calculate the corresponding proportion of panel households and the ratio of the panel households to the cross-sectional households. The cluster-level information for each set of cross-sectional weights is documented in the Excel file "Compilation_cluster_factors.xlsx". For example, in the case of the CMS2 Main Survey cross-sectional weights, the following factors are included in the weighting file:

- wt_CMS2_ms_panel_ind: CMS2 Main Survey panel weight for individuals, before population adjustment
- Weight_cluster_component_ms: cluster component of baseline weight for CMS1 Main Survey
- p_panel_CMS2_ms: proportion of CMS2 households in panel within cluster of Main Survey
- num_hhm_18more_2: number of household members 18 years or older in CMS2
- Baseline_wt_adjustment_CMS2_ms = number of CMS1 baseline panel households divided by CMS2 Main Survey households in cluster
- wt_CMS2_ms_cs_ind = weight for CMS2 Main Survey cross-sectional survey, before adjustment based on population estimates

10. As described in the report, it was necessary to adjust the cross-sectional weights for certain regions or provinces to account for missing clusters or additional clusters. These adjustments were made directly to cluster weight components used for calculating the cross-sectional weights.

11. Continuing with the CMS2 Main Survey as an example, the following SPSS syntax was used to generate the cross-sectional weights for the panel and new individuals:

    For panel individuals:

```
IF  (cms1 = 1) wt_CMS2_ms_cs_ind = wt_CMS2_ms_panel_ind *
p_panel_CMS2_ms.
EXECUTE.
```

    For new individuals:

```
IF  (cms1  ~=  1) wt_CMS2_ms_cs_ind = Baseline_wt_adjustment_CMS2_ms *
Weight_cluster_component_ms / num_hhm_18more_2.
EXECUTE.
```

12. The last step in producing the final individual weights for each set of panel and cross-sectional weights for all waves was to adjust the preliminary weights generated in the preceding steps based on population projections for the corresponding data collection month, as described in the report. This involved using the 2001 and 2011 Bulgaria Census population counts by province, urban and rural stratum to estimate the monthly population change by stratum, which was used to project the total population by province each month. These population estimates are included in the spreadsheet "Census_2001_2011_including_regional_estimates.xlsx". The population estimates for the data collection month of each wave were used for adjusting the weights. Based on the formula for the adjustment factors described in the report, the weighted total population by province and settlement type (rural, metropolitan and other urban) was calculated for each set of panel and cross-sectional weights. These weighted estimates were used in the denominator of the adjustment factors, and the numerator consisted of the corresponding census projections. The adjustment factors were calculated in the spreadsheet "BLISS_CMS_population_estimates_adjustment_factors_final.xlsx". As explained in the report, in some cases it was necessary to aggregate the strata to the region, urban/rural level. The final weight adjustment factors were compiled in the spreadsheet "CMS_BLISS_wt_adj_matrix.xlsx" so that they could be merged with the individual weighting files. Each weighting file for the baseline, panel and cross-sectional surveys has the preliminary weights, the adjustment factors and the final weight, which is the product of the preliminary weight and the adjustment factor.

13. For the calculation of the skills module weights, the final BLISS cross-sectional and panel household weights (for both the Main Survey and the Booster Survey) were merged with corresponding new household-level files that had the following variables:

- psy_valid_B: the observation was valid for analysis
- num_eligible_skills_B: number of household members eligible for the skills module
- hh_skills_module_B: One member of this household participated in the skills module

For the sample households that had one member who participated in the skills module and the observation was valid for the analysis, the basic cross-sectional or panel skills module weight was calculated as the corresponding final weight times the variable num_eligible_skills_B. For example, in the case of the BLISS panel sample for the Main Survey, the following syntax was used in SPSS:

```
IF  (hh_skills_module_B = 1 & psy_valid_B = 1)
wt_BLISS_SM_ms_panel=wt_BLISS_ms_panel_hh_final *
num_eligible_skills_B.
EXECUTE.
```

14. The next step was to calculate the nonresponse adjustment factors for the skills module weights. This involved producing tables for the numerator and denominator variables of the adjustment factors. The numerator estimates came from a table on the weighted distribution of the population age 18 to 65 using the final BLISS cross-sectional (or

panel) weights. The denominator estimates came from a similar weighted distribution table corresponding to the sum of the basic skills module weights for the households that had one member who participated in the skills module and the observation was valid for the analysis. The ratios of the numerator and denominator estimates for all strata were calculated in the spreadsheet "Nonresponse_adjustment_factors_skills_module". These final adjustment factors were then added to the spreadsheet "CMS_BLISS_wt_adj_matrix.xlsx" so they could be merged with the corresponding files with the weights.

15. After the weight adjustment factors were merged with the household data files with the weights, the final skills module weight was calculated as the basic skills module weight times the nonresponse weight adjustment factor for the corresponding stratum. For example, in the case of the BLISS Main Survey skills module weight for the panel, the final weights were calculated using the following SPSS syntax:

```
COMPUTE wt_BLISS_SM_ms_panel_final=wt_BLISS_SM_ms_panel *
Adj_BLISS_ms_panel_SM.
EXECUTE.
```

16. The files with the weights for the different panel and cross-sectional data sets were submitted in a Stata format, as specified in the Terms of Reference. The intermediary Excel files described in this Annex were also submitted to the World Bank team as part of the metadata documentation.

Annex B. Output from Logistic Regression Analysis for Adjusting Panel Weights of each Wave for Attrition

CMS2, Main Survey

```
. do "C:\Users\David\AppData\Local\Temp\STD0200000
> 0.tmp"

. xi: stepwise, pr(.2): logit response i.region i.settl_type age_1 i.gender_1 i.group_marital_stat
> us_1 i.group_education_1 i.ethnicity_1 i.group_religion_1 hhsize_1 income_quintile_1
i.group_head_education_1 i.group_spouse_education_1
i.region          _Iregion_1-6       (naturally coded; _Iregion_1 omitted)
i.settl_type      _Isettl_typ_1-3    (naturally coded; _Isettl_typ_1 omitted)
i.gender_1        _Igender_1_0-1     (naturally coded; _Igender_1_0 omitted)
i.group_marit~1   _Igroup_mar_1-4    (naturally coded; _Igroup_mar_1 omitted)
i.group_educa~1   _Igroup_edu_1-4    (naturally coded; _Igroup_edu_1 omitted)
i.ethnicity_1     _Iethnicity_1-4    (naturally coded; _Iethnicity_1 omitted)
i.group_relig~1   _Igroup_rel_1-5    (naturally coded; _Igroup_rel_1 omitted)
i.group_head_~1   _Igroup_hea_1-4    (naturally coded; _Igroup_hea_1 omitted)
i.group_spous~1   _Igroup_spo_1-9999 (naturally coded; _Igroup_spo_1 omitted)

begin with full model
p = 0.9644 >= 0.2000  removing _Igroup_spo_4
p = 0.8846 >= 0.2000  removing _Igroup_hea_4
p = 0.8352 >= 0.2000  removing income_quintile_1
p = 0.7923 >= 0.2000  removing _Iethnicity_2
p = 0.7456 >= 0.2000  removing _Iethnicity_3
p = 0.6361 >= 0.2000  removing hhsize_1
p = 0.6176 >= 0.2000  removing _Iethnicity_4
p = 0.5413 >= 0.2000  removing _Igroup_mar_3
p = 0.3167 >= 0.2000  removing _Igroup_spo_2

Logistic regression
> Number of obs   =      6578
> LR chi2(22)     =    497.49
> Prob > chi2     =    0.0000
Log likelihood =  -3490.9246
> Pseudo R2       =    0.0665
```

> ------------------------------

response	Coef.	Std. Err.	z	P> z	[95% Conf.	Interval]
_Iregion_2	1.100734	.1424872	7.73	0.000	.8214643	1.380004
_Iregion_3	.5312869	.1306777	4.07	0. 000	.2751632	.7874105
_Iregion_4	.3305014	.1053131	3.14	0.002	.1240915	.5369113
_Iregion_5	-.272703	.1059601	-2.57	0.010	-.480381	-.065025
_Iregion_6	-.6985859	.1089753	-6.41	0.000	-.9121736	-.4849981
_Isettl_ty~2	-.1504093	.0797159	-1.89	0.059	-.3066496	.005831
_Isettl_ty~3	.2841214	.0726283	3.91	0.000	.1417726	.4264702
age_1	.0088838	.0018017	4.93	0.000	.0053526	.0124151
_Igender_1_1	-.0837689	.0595734	-1.41	0.160	-.2005306	.0329927
_Igroup_ma~2	.2535576	.0844446	3.00	0.003	.0880493	.4190659
_Igroup_sp~3	-.1626187	.0786317	-2.07	0.039	-.3167339	-.0085034
_Igroup_ma~4	-.260252	.1627859	-1.60	0.110	-.5793064	.0588024
_Igroup_ed~2	-.4049926	.1231559	-3.29	0.001	-.6463737	-.1636115
_Igroup_ed~3	-.5957029	.1262313	-4.72	0.000	-.8431117	-.348294
_Igroup_ed~4	-.6842259	.1478307	-4.63	0.000	-.9739688	-.3944831
_Igroup_he~3	-.2537297	.0947	-2.68	0.007	-.4393384	-.0681211
_Igroup~9999	.3560669	.101897	3.49	0.000	.1563523	.5557814
_Igroup_he~2	-.3868639	.1175522	-3.29	0.001	-.617262	-.1564657
_Igroup_re~2	.2902559	.1011722	2.87	0.004	.0919621	.4885497
_Igroup_re~3	1.062375	.2810896	3.78	0.000	.5114499	1.613301
_Igroup_re~4	1.47272	.5399203	2.73	0.006	.4144961	2.530945
_Igroup_re~5	.7171895	.1389254	5.16	0.000	.4449007	.9894783
_cons	1.088189	.1584306	6.87	0.000	.7776708	1.398707

> ------------------------------

. predict ps
(option pr assumed; Pr(response))
(56 missing values generated)

CMS2, Booster Survey

```
. * logistic regression
. xi: stepwise, pr(.2): logit response i.region i.
> settl_type age_1 i.gender_1 i.group_marital_stat
> us_1 i.group_education_1 i.ethnicity_1 i.group_r
> eligion_1 hhsize_1 income_quintile_1 i.group_hea
> d_education_1 i.group_spouse_education_1
i.region          _Iregion_1-6       (naturally coded; _Iregion_1 omitted)
i.settl_type      _Isettl_typ_1-3    (naturally coded; _Isettl_typ_1 omitted)
i.gender_1        _Igender_1_0-1     (naturally coded; _Igender_1_0 omitted)
i.group_marit~1   _Igroup_mar_1-4    (naturally coded; _Igroup_mar_1 omitted)
i.group_educa~1   _Igroup_edu_1-4    (naturally coded; _Igroup_edu_1 omitted)
i.ethnicity_1     _Iethnicity_1-4    (naturally coded; _Iethnicity_1 omitted)
i.group_relig~1   _Igroup_rel_1-5    (naturally coded; _Igroup_rel_1 omitted)
i.group_head_~1   _Igroup_hea_1-4    (naturally coded; _Igroup_hea_1 omitted)
i.group_spous~1   _Igroup_spo_1-9999 (naturally coded; _Igroup_spo_1 omitted)
note: _Iethnicity_4 dropped because of estimability
note: _Igroup_rel_4 dropped because of estimability
note: o._Iethnicity_4 dropped because of estimability
note: o._Igroup_rel_4 dropped because of estimability
note: 18 obs. dropped because of estimability
                begin with full model
p = 0.8608 >= 0.2000  removing _Igroup_hea_3
p = 0.7855 >= 0.2000  removing income_quintile_1
p = 0.7998 >= 0.2000  removing _Igroup_edu_4
p = 0.6694 >= 0.2000  removing _Iethnicity_3
p = 0.6709 >= 0.2000  removing _Igroup_spo_9999
p = 0.4573 >= 0.2000  removing _Igroup_edu_3
p = 0.3390 >= 0.2000  removing _Igroup_rel_2
p = 0.3780 >= 0.2000  removing _Iethnicity_2
p = 0.3288 >= 0.2000  removing _Igroup_rel_3
p = 0.3076 >= 0.2000  removing _Igender_1_1
p = 0.2276 >= 0.2000  removing _Igroup_hea_4

Logistic regression
> Number of obs      =      1075
> LR chi2(18)  =   174.44
> Prob > chi2     =     0.0000
Log likelihood =  -404.8323
> Pseudo R2   =     0.1773
```

response	Coef.	Std. Err.	z	P> |z|	[95% Conf. Interval]	
_Iregion_2	-2.589164	.5942749	-4.36	0.000	-3.753921	-1.424406
_Iregion_3	-2.413023	.6212505	-3.88	0. 000	-3.630652	-1.195394
_Iregion_4	-1.667104	.5747922	-2.90	0.004	-2.793676	-.540532
_Iregion_5	-2.299613	.5681598	-4.05	0.000	-3.413186	-1.186041
_Iregion_6	-2.387711	.5741891	-4.16	0.000	-3.513101	-1.262321
_Isettl_ty~2	-.6106434	.2675856	-2.28	0.022	-1.135102	-.0861852
_Isettl_ty~3	-.8383407	.2217683	-3.78	0.000	-1.272999	-.4036828
age_1	.0326597	.0075044	4.35	0.000	.0179514	.0473681
_Igroup_re~5	.5489469	.3462977	1.59	0.113	-.1297841	1.227678
_Igroup_ma~2	-.6514265	.2833279	-2.30	0.021	-1.206739	-.0961139
_Igroup_ma~3	-1.810125	.567655	-3.19	0.001	-2.922709	-.697542
_Igroup_ma~4	-.7710342	.5812255	-1.33	0.185	-1.910215	.3681468
_Igroup_ed~2	-.3873698	.2251937	-1.72	0.085	-.8287413	.0540017
hhsize_1	.1911629	.0551095	3.47	0.001	.0831503	.2991756
_Igroup_sp~3	-2.274672	.3641336	-6.25	0.000	-2.98836	-1.560983
_Igroup_sp~4	-1.528152	.6913962	-2.21	0.027	-2.883263	-.17304
_Igroup_sp~2	-1.465395	.3164671	-4.63	0.000	-2.085659	-.8451306
_Igroup_he~2	.435832	.3040256	1.43	0.152	-.1600471	1.031711
_cons	3.926479	.7294054	5.38	0. 000	2.496871	5.356087

. predict ps
(option pr assumed; Pr(response))
(6 missing values generated)

CMS3, Main Survey

```
. do "C:\Users\David\AppData\Local\Temp\STD0200000
> 0.tmp"

. xi: stepwise, pr(.2): logit response i.region i.
> settl_type age_1 i.gender_1 i.group_marital_status_1 i.group_education_1 i.ethnicity_1 i.group_r
> eligion_1 hhsize_1 income_quintile_1 i.group_head_education_1 i.group_spouse_education_1
i.region          _Iregion_1-6       (naturally coded; _Iregion_1 omitted)
i.settl_type      _Isettl_typ_1-3    (naturally coded; _Isettl_typ_1 omitted)
i.gender_1        _Igender_1_0-1     (naturally coded; _Igender_1_0 omitted)
i.group_marit~1   _Igroup_mar_1-4    (naturally coded; _Igroup_mar_1 omitted)
i.group_educa~1   _Igroup_edu_1-4    (naturally coded; _Igroup_edu_1 omitted)
i.ethnicity_1     _Iethnicity_1-4    (naturally coded; _Iethnicity_1 omitted)
i.group_relig~1   _Igroup_rel_1-5    (naturally coded; _Igroup_rel_1 omitted)
i.group_head_~1   _Igroup_hea_1-4    (naturally coded; _Igroup_hea_1 omitted)
i.group_spous~1   _Igroup_spo_1-9999 (naturally coded; _Igroup_spo_1 omitted)
begin with full model
p = 0.8381 >= 0.2000  removing _Iregion_2
p = 0.8285 >= 0.2000  removing _Igroup_spo_2
p = 0.7643 >= 0.2000  removing _Iethnicity_2
p = 0.8333 >= 0.2000  removing _Iethnicity_4
p = 0.7216 >= 0.2000  removing _Iethnicity_3
p = 0.7192 >= 0.2000  removing _Igender_1_1
p = 0.6865 >= 0.2000  removing _Igroup_spo_3
p = 0.6385 >= 0.2000  removing _Igroup_spo_4
p = 0.2865 >= 0.2000  removing _Igroup_hea_4
p = 0.8105 >= 0.2000  removing _Igroup_hea_3
p = 0.6819 >= 0.2000  removing _Igroup_hea_2
p = 0.2808 >= 0.2000  removing _Igroup_mar_4
p = 0.4299 >= 0.2000  removing _Igroup_mar_3

Logistic regression
> Number of obs    =      6578
> LR chi2(18)      =    420.42
> Prob > chi2      =    0.0000
Log likelihood = -3681.8577
> Pseudo R2     =    0.0540
```

```
> ------------------------------
response            Coef.       Std. Err.      z    P>z        [95% Conf. Interval]

> ------------------------------
_Igroup~9999     .2363213   .0907358     2.60   0.009     .0584825     .4141602
_Iregion_3      -.3460478   .1123815    -3.08   0.002    -.5663116    -.1257841
_Iregion_4      -.6242997   .0894026    -6.98   0.000    -.7995255    -.4490739
_Iregion_5      -.6549689   .0921925    -7.10   0.000    -.8356629    -.4742749
_Iregion_6      -1.347915   .0956464   -14.09   0.000    -1.535378    -1.160451
_Isettl_ty~2    -.2798567   .0769627    -3.64   0.000    -.4307008    -.1290126
_Isettl_ty~3     .3765608   .0709609     5.31   0.000     .2374799     .5156416
age_1            .0085732   .0018155     4.72   0.000     .0050149     .0121314
_Igroup_re~5     .2184205   .1221566     1.79   0.074    -.0210021     .457843
_Igroup_ma~2     .2250727   .0782934     2.87   0.004     .0716203     .378525
_Igroup_re~2     .3796546   .0963235     3.94   0.000     .1908641     .5684451
_Igroup_re~3     1.052644   .2668501     3.94   0.000      .529628     1.575661
_Igroup_ed~2    -.3465225   .1162498    -2.98   0.003    -.574368     -.1186771
_Igroup_ed~3    -.5955724   .1164477    -5.11   0.000    -.8238056    -.3673392
_Igroup_ed~4    -.5919745   .1328361    -4.46   0.000    -.8523286    -.3316205
_Igroup_re~4     .7470255    .457288     1.63   0.102    -.1492426     1.643294
hhsize_1        -.1102563   .0225346    -4.89   0.000    -.1544233    -.0660893
income_qui~1     .0523955   .0255815     2.05   0.041     .0022567     .1025343
_cons             1.52471   .1681942     9.07   0.000     1.195056     1.854365
> ------------------------------

. predict ps
(option pr assumed; Pr(response))
(60 missing values generated)
```

CMS3, Booster Survey

```
. xi: stepwise, pr(.2): logit response i.region i.
> settl_type age_1 i.gender_1 i.group_marital_status_1 i.group_education_1 i.ethnicity_1 i.group_r
> eligion_1 hhsize_1 income_quintile_1 i.group_head_education_1 i.group_spouse_education_1
i.region          _Iregion_1-6       (naturally coded; _Iregion_1 omitted)
i.settl_type      _Isettl_typ_1-3    (naturally coded; _Isettl_typ_1 omitted)
i.gender_1        _Igender_1_0-1     (naturally coded; _Igender_1_0 omitted)
i.group_marit~1   _Igroup_mar_1-4    (naturally coded; _Igroup_mar_1 omitted)
i.group_educa~1   _Igroup_edu_1-4    (naturally coded; _Igroup_edu_1 omitted)
i.ethnicity_1     _Iethnicity_1-4    (naturally coded; _Iethnicity_1 omitted)
i.group_relig~1   _Igroup_rel_1-5    (naturally coded; _Igroup_rel_1 omitted)
i.group_head_~1   _Igroup_hea_1-4    (naturally coded; _Igroup_hea_1 omitted)
i.group_spous~1   _Igroup_spo_1-9999 (naturally coded; _Igroup_spo_1 omitted)
note: _Iethnicity_4 dropped because of estimability
note: o._Iethnicity_4 dropped because of estimability
note: 7 obs. dropped because of estimability
begin with full model
p = 0.9797 >= 0.2000  removing _Iethnicity_2
p = 0.9433 >= 0.2000  removing _Igroup_edu_4
p = 0.7697 >= 0.2000  removing _Igroup_mar_3
p = 0.7661 >= 0.2000  removing _Igroup_mar_4
p = 0.7642 >= 0.2000  removing _Igroup_rel_3
p = 0.6695 >= 0.2000  removing _Igender_1_1
p = 0.6227 >= 0.2000  removing _Iethnicity_3
p = 0.6191 >= 0.2000  removing _Igroup_edu_3
p = 0.4926 >= 0.2000  removing _Igroup_spo_3
p = 0.4372 >= 0.2000  removing _Igroup_rel_4
p = 0.4396 >= 0.2000  removing _Igroup_mar_2
p = 0.3625 >= 0.2000  removing _Igroup_edu_2
p = 0.2651 >= 0.2000  removing income_quintile_1

Logistic regression
> Number of obs      =      1086
> LR chi2(17)        =    123.90
> Prob > chi2        =    0.0000
Log likelihood =    -520.32512
> Pseudo R2          =    0.1064
```

> ----------

response	Coef.	Std. Err.	z	P> z	[95% Conf.	Interval]
> ----------						
_Iregion_2	-1.59218	.4848203	-3.28	0.001	-2.54241	-.6419497
_Iregion_3	-2.783157	.4725329	-5.89	0.000	-3.709305	-1.85701
_Iregion_4	-1.344841	.4162512	-3.23	0.001	-2.160679	-.5290041
_Iregion_5	-2.278209	.4128145	-5.52	0.000	-3.087311	-1.469108
_Iregion_6	-2.131069	.4235051	-5.03	0.000	-2.961123	-1.301014
_Isettl_ty~2	-1.181814	.2536585	-4.66	0.000	-1.678976	-.6846529
_Isettl_ty~3	-1.035286	.2091843	-4.95	0. 000	-1.44528	-.6252928
age_1	.0059437	.0039871	1.49	0.136	-.0018709	.0137583
_Igroup_he~3	-.8527962	.382003	-2.23	0.026	-1.601508	-.104084
_Igroup_re~5	.9065572	.3342369	2.71	0.007	.251465	1.561649
hhsize_1	-.0590981	.0410147	-1.44	0.150	-.1394854	.0212893
_Igroup_sp~2	.5217429	.2514771	2.07	0.038	.0288567	1.014629
_Igroup_he~4	-1.680762	.7065664	-2.38	0.017	-3.065607	-.2959173
_Igroup_re~2	.5491733	.2367104	2.32	0.020	.0852293	1.013117
_Igroup_sp~4	1.098121	.6505781	1.69	0.091	-.1769887	2.373231
_Igroup~9999	.5063047	.2643733	1.92	0.055	-.0118574	1.024467
_Igroup_he~2	-.5510302	.3308846	-1.67	0.096	-1.199552	.0974917
_cons	3.931257	.5562457	7.07	0.000	2.841036	5.021479

> ----------

. predict ps
(option pr assumed; Pr(response))
(4 missing values generated)

BLISS, Main Survey

```
. xi: stepwise, pr(.2): logit response i.region i.
> settl_type age_1 i.gender_1 i.group_marital_status_1 i.group_education_1 i.ethnicity_1 i.group_r
> eligion_1 hhsize_1 income_quintile_1 i.group_head_education_1 i.group_spouse_education_1
i.region          _Iregion_1-6       (naturally coded; _Iregion_1 omitted)
i.settl_type      _Isettl_typ_1-3    (naturally coded; _Isettl_typ_1 omitted)
i.gender_1        _Igender_1_0-1     (naturally coded; _Igender_1_0 omitted)
i.group_marit~1   _Igroup_mar_1-4    (naturally coded; _Igroup_mar_1 omitted)
i.group_educa~1   _Igroup_edu_1-4    (naturally coded; _Igroup_edu_1 omitted)
i.ethnicity_1     _Iethnicity_1-4    (naturally coded; _Iethnicity_1 omitted)
i.group_relig~1   _Igroup_rel_1-5    (naturally coded; _Igroup_rel_1 omitted)
i.group_head_~1   _Igroup_hea_1-4    (naturally coded; _Igroup_hea_1 omitted)
i.group_spous~1   _Igroup_spo_1-9999 (naturally coded; _Igroup_spo_1 omitted)
begin with full model
p = 0.9804 >= 0.2000  removing _Igroup_rel_4
p = 0.8601 >= 0.2000  removing _Igroup_edu_4
p = 0.8217 >= 0.2000  removing _Iethnicity_4
p = 0.4669 >= 0.2000  removing _Igroup_edu_3
p = 0.4580 >= 0.2000  removing _Igroup_hea_3
p = 0.3958 >= 0.2000  removing _Igroup_edu_2

Logistic regression
> Number of obs     =      6578
> LR chi2(25)       =    786.38
> Prob > chi2       =    0.0000
Log likelihood = -3858.6738
> Pseudo R2         =    0.0925
```

> ------------------------------

response	Coef.	Std. Err.	z	P> z	[95% Conf.	Interval]
> ------------------------------						
_Iregion_2	.7953882	.151875	5.24	0.000	.4977187	1.093058
_Iregion_3	-.9769731	.1246291	-7.84	0.000	-1.221242	-.7327046
_Iregion_4	-1.330733	.1096244	-12.14	0. 000	-1.545593	-1.115873
_Iregion_5	-.7500582	.1141706	-6.57	0.000	-.9738284	-.5262879
_Iregion_6	-.8070358	.1188967	-6.79	0.000	-1.040069	-.5740024
_Isettl_ty~2	-.3369266	.0790262	-4.26	0.000	-.4918151	-.1820382
_Isettl_ty~3	-.16217	.0692539	-2.34	0.019	-.2979051	-.0264349
age_1	.0156602	.0022103	7.08	0.000	.011328	.0199923
_Igender_1_1	-.0750082	.0565658	-1.33	0.185	-.1858752	.0358587
_Igroup_ma~2	-.2887661	.0914416	-3.16	0.002	-.4679884	-.1095439
_Igroup_ma~3	-.7901736	.1546604	-5.11	0.000	-1.093302	-.4870447
_Igroup_ma~4	-.4599813	.1661976	-2.77	0.006	-.7857227	-.1342399
_Igroup_he~4	-.2179619	.0780943	-2.79	0.005	-.3710239	-.0648999
_Igroup_sp~2	-1.210969	.4856398	-2.49	0.013	-2.162806	-.2591324
_Igroup_sp~4	-1.681027	.4916234	-3.42	0.001	-2.644591	-.717463
_Iethnicit~2	-.9750163	.1685243	-5.79	0.000	-1.305318	-.6447147
_Iethnicit~3	-.6671698	.1495563	-4.46	0.000	-.9602946	-.3740449
_Igroup_sp~3	-1.614922	.4869554	-3.32	0.001	-2.569337	-.6605066
_Igroup_re~2	.7257553	.1521965	4.77	0.000	.4274556	1.024055
_Igroup_re~3	.9204771	.2437745	3.78	0.000	.4426879	1.398266
_Igroup~9999	-1.259816	.491233	-2.56	0.010	-2.222615	-.2970166
_Igroup_re~5	.5292418	.1266956	4.18	0.000	.280923	.7775606
hhsize_1	.2154234	.0253603	8.49	0.000	.1657182	.2651286
income_qui~1	-.1087604	.0270685	-4.02	0.000	-.1618137	-.0557072
_Igroup_he~2	-.3918628	.094353	-4.15	0.000	-.5767912	-.2069344
_cons	2.448791	.5155266	4.75	0.000	1.438378	3.459205

> ------------------------------

. predict ps
(option pr assumed; Pr(response))
(44 missing values generated)

BLISS, Booster Survey

```
. xi: stepwise, pr(.2): logit response i.region i.
> settl_type age_1 i.gender_1 i.group_marital_status_1 i.group_education_1 i.ethnicity_1 i.group_r
> eligion_1 hhsize_1 income_quintile_1 i.group_head_education_1 i.group_spouse_education_1
i.region          _Iregion_1-6       (naturally coded; _Iregion_1 omitted)
i.settl_type      _Isettl_typ_1-3    (naturally coded; _Isettl_typ_1 omitted)
i.gender_1        _Igender_1_0-1     (naturally coded; _Igender_1_0 omitted)
i.group_marit~1   _Igroup_mar_1-4    (naturally coded; _Igroup_mar_1 omitted)
i.group_educa~1   _Igroup_edu_1-4    (naturally coded; _Igroup_edu_1 omitted)
i.ethnicity_1     _Iethnicity_1-4    (naturally coded; _Iethnicity_1 omitted)
i.group_relig~1   _Igroup_rel_1-5    (naturally coded; _Igroup_rel_1 omitted)
i.group_head_~1   _Igroup_hea_1-4    (naturally coded; _Igroup_hea_1 omitted)
i.group_spous~1   _Igroup_spo_1-9999 (naturally coded; _Igroup_spo_1 omitted)
note: _Iethnicity_4 dropped because of estimability
note: o._Iethnicity_4 dropped because of estimability
note: 7 obs. dropped because of estimability
begin with full model
p = 0.9917 >= 0.2000  removing _Igroup_edu_3
p = 0.9883 >= 0.2000  removing _Igroup_mar_3
p = 0.9851 >= 0.2000  removing _Igroup_edu_4
p = 0.9090 >= 0.2000  removing _Igroup_rel_3
p = 0.9010 >= 0.2000  removing _Iethnicity_3
p = 0.7000 >= 0.2000  removing _Igroup_rel_4
p = 0.4973 >= 0.2000  removing _Igroup_mar_4
p = 0.3512 >= 0.2000  removing _Igroup_hea_4
p = 0.3855 >= 0.2000  removing _Igroup_spo_4
p = 0.3221 >= 0.2000  removing _Igender_1_1
p = 0.2642 >= 0.2000  removing _Igroup_hea_3
p = 0.2627 >= 0.2000  removing _Iethnicity_2

Logistic regression
> Number of obs      =      1086
> LR chi2(18)   =    207.84
> Prob > chi2       =    0.0000
Log likelihood = -477.12762
> Pseudo R2     =    0.1789
```

> ----------

response	Coef.	Std. Err.	z	P> z	[95% Conf.	Interval]
> ----------						
_Iregion_2	-1.451114	.4535722	-3.20	0.001	-2.340099	-.5621287
_Iregion_3	-3.01214	.4409289	-6.83	0.000	-3.876345	-2.147935
_Iregion_4	-.6528194	.37739	-1.73	0.084	-1.39249	.0868514
_Iregion_5	-1.337877	.3719091	-3.60	0.000	-2.066806	-.6089485
_Iregion_6	-1.784679	.3909069	-4.57	0.000	-2.550842	-1.018515
_Isettl_ty~2	-1.042125	.2648761	-3.93	0.000	-1.561272	-.522977
_Isettl_ty~3	-1.376973	.2188301	-6.29	0.000	-1.805872	-.9480734
age_1	.0147379	.0052518	2.81	0.005	.0044445	.0250312
_Igroup_re~5	2.322123	.4667697	4.97	0.000	1.407271	3.236975
_Igroup_ma~2	-.3729972	.2135726	-1.75	0.081	-.7915918	.0455974
hhsize_1	.2837979	.0499721	5.68	0.000	.1858544	.3817415
_Igroup_he~2	.6797888	.2541576	2.67	0.007	.181649	1.177929
_Igroup_ed~2	-.3752976	.2052546	-1.83	0.067	-.7775894	.0269941
_Igroup~9999	.5838476	.3252632	1.80	0.073	-.0536565	1.221352
_Igroup_sp~3	1.264992	.33388	3.79	0.000	.6105992	1.919385
_Igroup_sp~2	1.057728	.3109454	3.40	0.001	.4482862	1.66717
income_qui~1	.1173789	.0812874	1.44	0.149	-.0419415	.2766992
_Igroup_re~2	.4928334	.2423011	2.03	0.042	.017932	.9677347
_cons	.3538556	.4865322	0.73	0.467	-.59973	1.307441

> ----------

. predict ps
(option pr assumed; Pr(response))
(6 missing values generated)

Combined panel, Main Survey

```
. * logistic regression
. xi: stepwise, pr(.2): logit response i.region i.
> settl_type age_1 i.gender_1 i.group_marital_status_1 i.group_education_1 i.ethnicity_1 i.group_r
> eligion_1 hhsize_1 income_quintile_1 i.group_head_education_1 i.group_spouse_education_1
i.region          _Iregion_1-6       (naturally coded; _Iregion_1 omitted)
i.settl_type      _Isettl_typ_1-3    (naturally coded; _Isettl_typ_1 omitted)
i.gender_1        _Igender_1_0-1     (naturally coded; _Igender_1_0 omitted)
i.group_marit~1   _Igroup_mar_1-4    (naturally coded; _Igroup_mar_1 omitted)
i.group_educa~1   _Igroup_edu_1-4    (naturally coded; _Igroup_edu_1 omitted)
i.ethnicity_1     _Iethnicity_1-4    (naturally coded; _Iethnicity_1 omitted)
i.group_relig~1   _Igroup_rel_1-5    (naturally coded; _Igroup_rel_1 omitted)
i.group_head_~1   _Igroup_hea_1-4    (naturally coded; _Igroup_hea_1 omitted)
i.group_spous~1   _Igroup_spo_1-9999 (naturally coded; _Igroup_spo_1 omitted)
                  begin with full model
p = 0.9433 >= 0.2000  removing _Igroup_spo_2
p = 0.3428 >= 0.2000  removing _Iethnicity_4
p = 0.3108 >= 0.2000  removing _Igroup_rel_4
p = 0.2678 >= 0.2000  removing hhsize_1
p = 0.2675 >= 0.2000  removing _Igender_1_1

Logistic regression
> Number of obs      =      6578
> LR chi2(26)        =    465.28
> Prob > chi2        =    0.0000
Log likelihood =  -4316.917
> Pseudo R2         =    0.0511
```

```
------------------------------------------------------------------------------
    response |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iregion_2 |    .7206572   .1114392     6.47   0.000     .5022403    .9390741
  _Iregion_3 |   -.3055954   .1093989    -2.79   0.005    -.5200133   -.0911776
  _Iregion_4 |   -.6124702   .0944536    -6.48   0.000    -.7975958   -.4273445
  _Iregion_5 |   -.2773072   .0965893    -2.87   0.004    -.4666188   -.0879957
  _Iregion_6 |   -.3832497   .1028458    -3.73   0.000    -.5848237   -.1816758
_Isettl_ty~2 |   -.4155772   .0715902    -5.80   0.000    -.5558914   -.2752629
_Isettl_ty~3 |   -.2196983   .0635234    -3.46   0.001    -.3442018   -.0951948
       age_1 |    .0081926   .0020159     4.06   0.000     .0042415    .0121436
_Igroup_he~4 |   -.6933502   .3428718    -2.02   0.043    -1.365367   -.0213338
_Igroup_ma~2 |    .1999798   .0886553     2.26   0.024     .0262186    .373741
_Igroup_ma~3 |   -.2901031    .139603    -2.08   0.038     -.56372   -.0164862
_Igroup_ma~4 |    -.263739   .1561253    -1.69   0.091     -.569739    .042261
_Igroup_ed~2 |    -.220766   .1098251    -2.01   0.044    -.4360192   -.0055129
_Igroup_ed~3 |   -.3162242   .1141771    -2.77   0.006    -.5400072   -.0924413
_Igroup_ed~4 |   -.3850111   .1354014    -2.84   0.004    -.6503929   -.1196293
_Iethnicit~2 |   -.8356628   .1483017    -5.63   0.000    -1.126329   -.5449969
_Iethnicit~3 |   -.2892772   .1301588    -2.22   0.026    -.5443839   -.0341706
_Igroup_sp~4 |   -.2565706    .115669    -2.22   0.027    -.4832776   -.0298636
_Igroup_re~2 |    .4189353    .131343     3.19   0.001     .1615077    .6763629
_Igroup_re~3 |    .8436189   .2070016     4.08   0.000     .4379033    1.249335
_Igroup_sp~3 |   -.1912606   .0979859    -1.95   0.051    -.3833093    .0007882
_Igroup_re~5 |    .3602555   .1139843     3.16   0.002     .1368504    .5836606
_Igroup~9999 |    .1869726   .1095071     1.71   0.088    -.0276574    .4016025
income_qui~1 |   -.0827494   .0236364    -3.50   0.000    -.1290758    -.036423
_Igroup_he~2 |   -.9295959   .3270525    -2.84   0.004    -1.570607   -.2885848
_Igroup_he~3 |   -.5607891   .3351202    -1.67   0.094    -1.217613    .0960344
       _cons |    1.245332   .3513923     3.54   0.000     .5566156    1.934048
------------------------------------------------------------------------------

. predict ps
(option pr assumed; Pr(response))
(71 missing values generated)
```

Combined panel, Booster Survey

```
. xi: stepwise, pr(.2): logit response i.region i.settl_type age_1 i.gender_1 i.group_marital_stat
> us_1 i.group_education_1 i.ethnicity_1 i.group_religion_1 hhsize_1 income_quintile_1 i.group_hea
> d_education_1 i.group_spouse_education_1
i.region          _Iregion_1-6       (naturally coded; _Iregion_1 omitted)
i.settl_type      _Isettl_typ_1-3    (naturally coded; _Isettl_typ_1 omitted)
i.gender_1        _Igender_1_0-1     (naturally coded; _Igender_1_0 omitted)
i.group_marit~1   _Igroup_mar_1-4    (naturally coded; _Igroup_mar_1 omitted)
i.group_educa~1   _Igroup_edu_1-4    (naturally coded; _Igroup_edu_1 omitted)
i.ethnicity_1     _Iethnicity_1-4    (naturally coded; _Iethnicity_1 omitted)
i.group_relig~1   _Igroup_rel_1-5    (naturally coded; _Igroup_rel_1 omitted)
i.group_head_~1   _Igroup_hea_1-4    (naturally coded; _Igroup_hea_1 omitted)
i.group_spous~1   _Igroup_spo_1-9999 (naturally coded; _Igroup_spo_1 omitted)
                  begin with full model
p = 0.8802 >= 0.2000  removing _Igroup_hea_3
p = 0.8545 >= 0.2000  removing _Igender_1_1
p = 0.8219 >= 0.2000  removing _Igroup_mar_4
p = 0.8056 >= 0.2000  removing _Igroup_spo_4
p = 0.7265 >= 0.2000  removing _Iethnicity_4
p = 0.7194 >= 0.2000  removing _Igroup_mar_3
p = 0.6732 >= 0.2000  removing _Igroup_edu_4
p = 0.6536 >= 0.2000  removing _Igroup_rel_3
p = 0.6008 >= 0.2000  removing _Igroup_spo_3
p = 0.4728 >= 0.2000  removing _Iethnicity_3
p = 0.4337 >= 0.2000  removing _Igroup_edu_3
p = 0.4033 >= 0.2000  removing _Igroup_mar_2
p = 0.3668 >= 0.2000  removing _Igroup_hea_2
p = 0.2876 >= 0.2000  removing _Igroup_rel_4
p = 0.2216 >= 0.2000  removing income_quintile_1
p = 0.2621 >= 0.2000  removing _Iregion_4

Logistic regression
> Number of obs      =      1093
> LR chi2(15)        =    167.07
> Prob > chi2        =    0.0000
Log likelihood = -664.16154
> Pseudo R2          =    0.1117
```

```
------------------------------------------------------------------------------
    response |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iregion_2 |   -.5927956   .2702862    -2.19   0.028    -1.122547   -.0630445
  _Iregion_3 |   -2.187327   .2853698    -7.66   0.000    -2.746641   -1.628012
_Iethnicit~2 |   -.5538502   .1979498    -2.80   0.005    -.9418248   -.1658756
  _Iregion_5 |   -1.091889    .18805    -5.81   0.000     -1.46046   -.7233178
  _Iregion_6 |   -.7178772   .1959617    -3.66   0.000    -1.101955   -.3337993
_Isettl_ty~2 |   -.9290532   .1997312    -4.65   0.000    -1.320519   -.5375873
_Isettl_ty~3 |   -1.252944   .1709861    -7.33   0.000    -1.588071   -.9178179
      age_1  |    .010937   .0036153     3.03   0.002     .0038511    .018023
_Igroup_sp~2 |   .4618083    .184795     2.50   0.012     .0996167    .8239998
_Igroup_re~5 |   .7411629   .2616839     2.83   0.005     .2282718    1.254054
_Igroup~9999 |   .4420719   .2082556     2.12   0.034     .0338985    .8502453
    hhsize_1 |    .0836751   .0347795     2.41   0.016     .015508     .1518418
_Igroup_ed~2 |   -.2504085   .1563624    -1.60   0.109    -.5568731    .0560562
_Igroup_he~4 |   -.7161659   .5145558    -1.39   0.164    -1.724677    .2923449
_Igroup_re~2 |    .3920318   .2064554     1.90   0.058    -.0126135    .796677
       _cons |    .7337503   .3063962     2.39   0.017     .1332247    1.334276
------------------------------------------------------------------------------

. predict ps
(option pr assumed; Pr(response))
(5 missing values generated)
```