

Washington State Population Survey Imputation

Age, Race, Ethnicity, and Sex

Post-stratification of sample survey data requires that each record's control variables (age, race, ethnicity, sex) contain no missing data. Data are missing if the response to one or more questions is coded "Refused," "Don't know," or "Not ascertained" due to questionnaire design defects, interviewer errors, or data processing errors.

Two main imputation techniques are employed for age, sex, and race/ethnicity. The first, *deductive imputation*, involves inferring a response by checking the logic or context of questions before or after the question in which a response is missing. For example, if the response on sex is missing for a household member, but the relationship question indicates that this person is the daughter of the respondent, the person's sex is set to "Female".

The second method, *simple random imputation*, assigns a response at random within a specified range. This process involves the use of a random number function.

All three demographic variables underwent a combination of the two imputation methods. For the age variable, the respondent's age is imputed first. Other household members' ages and relationships to the respondent are checked to see if an age can be inferred for the respondent. If that fails, then a random age is selected from a uniform distribution of values between 18 and 75.

The age variable is then imputed for other household members. Similarly, the relationship of the other members to the respondent is checked to determine if a response can be inferred. If the relationship check fails, then the member's marital status, education attainment, and military service status are checked to determine whether this individual is a child or an adult. If the member is an adult, a random age from 18 to 75 is assigned. If the member is determined to be a child, a random age from 0 to 17 is assigned. If this process also fails, then a random age is selected from the range of 0 to 75.

The sex variable, as mentioned earlier, is checked for a member's relationship to the respondent. If the individual's sex cannot be determined by this process, then sex is randomly assigned.

The race/ethnicity variable is a composite variable constructed from two questionnaire variables: race and Hispanic origin. The imputation is performed on the two original variables and is carried out in two stages. In the survey, the question on Hispanic origin is asked first and then followed by the question on race. The race question has seven response categories: African American, Native American, Asian or Pacific Islander, White, Other, and Hispanic. Some in the "Other" category are of mixed races. Others are one of the five listed races

but choose to use the "Other" category. In addition to the "Other" category for race and the Hispanic variable, the "Refused", "Don't know," and "Not ascertained" responses are also imputed.

The "Other" responses are examined to determine whether a specific race or races can be ascertained. Also, if any description is found indicating Hispanic origin, that person is classified as Hispanic.

The remaining missing cases are imputed by *proportional random imputation*. It is assumed that these remaining cases have a distribution pattern similar to that of identified cases in the "Other" category. In the proportional random imputation, cases that contained missing race and Hispanic origin information are assigned a random number ranging from 0 to 1. Then the proportions of races and persons of Hispanic origin as found in the "Other" category are applied to this random number to determine the proportion of each race and people of Hispanic origin in the remaining missing cases.

Household income

If respondent responded to income range but not exact income, we took a random number within that range.

If there was no income range response, then income was imputed using coefficients from a regression (for those households that DID have income information) of income on region, household size, income, imputed income, respondent's and partner's: education, occupation, rent/own, social security, cash assistance, food stamps, marital status, age, sex, ethnicity, race, labor force status, full time work, industry, weeks worked, and job sector.

Household records with income that needs to be imputed may not have all the socioeconomic attributes for the regression. So the procedure employs 7 or 8 equations, each with a subsample of households with the same set of socioeconomic data (independent variables in the regression) available in the records.

Health Insurance, Health Status, and Labor Force Status

Health insurance source, health status, and labor force status are imputed for individuals using *hot deck methods*. This method randomly selects a response from the set of survey respondents who match the nonresponders on a set of other questions.

Health insurance responses were filled in by randomly choosing a response from respondents who match on various characteristics depending on type of insurance. The characteristics include: age category, sex, labor force status (individual and spouse), employment sector, size of employment, household income as percentage of poverty level, disability status, social security benefits,

military service (individual and spouse), relationship to respondent, and known health insurance information.

Health status responses were filled in by randomly choosing a response from respondents who match on age category and disability status.

Labor force status is imputed by randomly selecting responses from people who did respond and who match on household size, age category, sex, race, ethnicity, education, and household income as percentage of poverty level.