

Imputation SHADAC's Case Study with Colorado

SHADAC staff worked closely with the state of Colorado to collect survey data on health insurance from 10,000 Coloradoans. A particular example from the Colorado survey illustrates the level of detail and expertise the SHADAC teams offers states. In Colorado's survey roughly 17.5% of the population refused to answer the household income question and the Colorado team was initially resistant to working with those cases. Colorado thought it would be better to leave those 17.5% of cases out of analyses that involved income and health insurance. However, the SHADAC staff were able to convince the Colorado team that leaving them out would bias estimates from the survey and imputing income data was a more preferable option. The following argument was presented to the Colorado research team:

To impute or not to impute? In survey research there is a substantial amount of missing data for certain types of items (e.g., income) because survey respondent's refuse to answer them for some reason. If the organization collecting the data decides to not impute missing values, they have made an assumption that the respondents with missing data are no different from the people with reported data. This assumption does not hold up under examination. For example, on average in the Colorado Household Survey the respondents with missing data on income had higher levels of education than those without missing income data. Higher levels of education are related to higher levels of income. Thus, the assumption that the respondents with reported data is incorrect and the estimates deriving from this assumption will be biased.

The work of statisticians Donald B. Rubin and Roderick A. Little present strong arguments for explicit model based imputation. For example, the model that a respondent with missing data is similar to others respondents with shared demographic and geographic variables is preferable to the implicit assumption that the respondents with missing data do not differ systematically from those with reported data. The bottom line is that high quality model based imputation results in superior estimates than simply throwing out the cases with missing values.