Missing Data and Multiple Imputation: An Unbiased Approach

M. Foy[1], M. VanBaalen[2], M. Wear[1], C. Mendez[3], S. Mason[3], V. Meyers[2], D. Alexander[2], J. Law[2]

[1]Wyle, [2] NASA Johnson Space Center, and [3]MEI Technologies

The default method of dealing with missing data in statistical analyses is to only use the complete observations (complete case analysis), which can lead to unexpected bias when data do not meet the assumption of missing completely at random (MCAR). For the assumption of MCAR to be met, missingness cannot be related to either the observed or unobserved variables. A less stringent assumption, missing at random (MAR), requires that missingness not be associated with the value of the missing variable itself, but can be associated with the other observed variables. When data are truly MAR as opposed to MCAR, the default complete case analysis method can lead to biased results.

There are statistical options available to adjust for data that are MAR, including multiple imputation (MI) which is consistent and efficient at estimating effects. Multiple imputation uses informing variables to determine statistical distributions for each piece of missing data. Then multiple datasets are created by randomly drawing on the distributions for each piece of missing data. Since MI is efficient, only a limited number, usually less than 20, of imputed datasets are required to get stable estimates. Each imputed dataset is analyzed using standard statistical techniques, and then results are combined to get overall estimates of effect. A simulation study will be demonstrated to show the results of using the default complete case analysis, and MI in a linear regression of MCAR and MAR simulated data.

Further, MI was successfully applied to the association study of CO2 levels and headaches when initial analysis showed there may be an underlying association between missing CO2 levels and reported headaches. Through MI, we were able to show that there is a strong association between average CO2 levels and the risk of headaches. Each unit increase in CO2 (mmHg) resulted in a doubling in the odds of reported headaches.