Handling Missing Data

By
Dr. Muhammad Usman
Manager (Database)
PASTIC National Center, Islamabad.

Missing Data

Rarely does "real" data come to you without any missing values. And "missing" can take different forms:

1. System Missing (empty entries, denoted by S)
2. Out of Range (out of bounds values denoted by -,+ in SPSS)

Missing Data - Reasons

- Missing data can occur because of nonresponse: no information is provided for several items or no information is provided for a whole unit. Some items are more sensitive for nonresponse than others, for example items about private subjects such as income.
- Dropout is a type of missingness that occurs mostly when studying development over time. In this type of study the measurement is repeated after a certain period of time. Missingness occurs when participants drop out before the test ends and one or more measurements are missing.
- Sometimes missing values are caused by the researcher—for example, when data collection is done improperly or mistakes are made in data entry. Data often are missing in research in economics, sociology, and political science because governments choose not to, or fail to, report critical statistics.

Missing Data

With all of these issues, you also need to determine if the data is missing:

1) **Missing at random** – also called MAR.

   An example of a MAR mechanism would be that a laboratory sample is dropped, so the resulting observation is missing.

2) **Missing that depends upon latent variables**. For example, there could be a latent (unobserved) variable which is highly correlated with the missing values. A familiar example from medical studies is that if a particular treatment causes discomfort, a patient is more likely to drop out of the study. This "missingness" is not at random (unless "discomfort" is measured and observed for all patients).

Impact of Missing Data

Think about this…if you are missing only 1% of your data and you have 1,000,000 observations and 50 variables, you could lose as much as 395,000 observations when you go to model…

[total observations – (((1-percent missing)^variables)*total observations)] or

[1,000,000 – (((1-.01)^50)*1,000,000)] = 394,994

That is A LOT of valid data that you would lose!

And, it could bias your results.

Handling Missing Values

We need a way to replace those values – logically.

Many options for replacement exist. Here are four of the primary methods:

1. Mean based
2. Median based
3. Stratified
4. Regressed

Replacement of Missing Values

Strategies:

1) **Mean Based** – this process is the most simple. This involves replacing the missing values with the mean of the variable.

2) **Median Based** – this process is also very simple. This involves replacing the missing values with the median of the variable.

3) **Stratified** – this process is slightly more involved. This involves replacing the missing values with the mean or median of the variable but with consideration for similar strata of observations.

4) **Regressed** – This process involves actually predicting the value of the missing values using Regression. It works well if: the variables are related to each other and if you only have one or two variables with missing data