CODERS' CORNER

CENTRE FOR THE STUDY OF AFRICAN ECONOMIES

UNIVERSITY OF OXFORD

DEALING WITH MISSING OBSERVATIONS: MULTIPLE IMPUTATION

Missing observations or attrition are a common issue in empirical research. This post briefly discusses the method multiple imputation to deal with missing observations[1]. Multiple imputation replaces every missing value of the variable with a list of simulated values. We can then run the required specification to estimate the parameter of interest using the imputed dataset.

To illustrate the method, I use data from a randomized controlled trial (RCT) conducted by Blattman et al. (2011, 2014). The data can be downloaded from Harvard Dataverse[2].

Step 1: Setting up the dataset

Suppose we are interested in investigating the impact of randomly assigned cash transfers (given at baseline) on migration (measured at endline). We can run a logistic regression since the outcome variable "migrate_e" is a binary variable that refers to whether the respondent has changed parish at endline, since baseline. However, after checking the data, we notice that the variable "migrate_e" has 161 missing observations.

Now, suppose we wish to impute those 161 observations for the "migrate_e" variable and then fit a logistic regression model with the complete dataset. To do so, we can use multiple imputation "mi" commands, which are readily available to use on Stata[3]. To use the commands, the dataset in memory should be set as an "mi" dataset, as shown in the second command below.

```
. summ

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
     treated |      5,354    .4407919    .4965284          0          1
       urban |      5,354    .2099365    .4073011          0          1
         age |      5,354    24.95106    5.190868         14         59
       age_2 |      5,354    649.4957    292.4608        196       3481
       age_3 |      5,354    17697.34    13689.43       2744     205379
-------------+---------------------------------------------------------
voc_training |      5,354    .0765783    .2659459          0          1
   migrate_e |      5,193    .3672251     .482095          0          1
   education |      5,354    7.859171    2.938558          0         14
 wealthindex |      5,354    -.119519    .9905206  -2.322873   5.178377
risk_avers~n |      5,354   -7.38e-09           1  -3.394701   1.391094
```

[1] Multiple imputation was first discussed in Rubin (1987).
[2] I am using the dataset "yop2_yop4_deid.dta".
[3] The Stata version used here is Stata 17.0.

```
. mi misstable summarize, all
                                                              Obs<.
                                                +------------------------------
               |                                | Unique
      Variable |     Obs=.     Obs>.     Obs<.  | values          Min         Max
  -------------+--------------------------------+------------------------------
       treated |                         5,354  |      2            0           1
         urban |                         5,354  |      2            0           1
           age |                         5,354  |     40           14          59
         age_2 |                         5,354  |     40          196        3481
         age_3 |                         5,354  |     40         2744      205379
  voc_training |                         5,354  |      2            0           1
     migrate_e |       161               5,193  |      2            0           1
     education |                         5,354  |     15            0          14
   wealthindex |                         5,354  |   >500    -2.322873    5.178377
  risk_avers~n |                         5,354  |    173    -3.394701    1.391094
  -----------------------------------------------------------------------------
```

The misstable summarize command is particularly useful when you have multiple variables with missing values.

Step 2: Imputing missing values

I impute the values for "migrate_e" using a logit regression since it is a binary variable. Refer to table 1 below for which commands to use after "mi impute" depending on the type of variable.

Table 1: Commands used for the different types of imputation variables.

Type of Variable	**Commands for the relevant imputation method**
Continuous	regress, pmm, truncreg, intreg
Binary	logit
Categorical	ologit, mlogit
Count	Poisson, nbreg

```
. mi impute regress migrate_e treated $controls, add(20) rseed(1234)

Univariate imputation                       Imputations =       20
Linear regression                                 added =       20
Imputed: m=1 through m=20                       updated =        0

--------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
         migrate_e |       5193          161       161 |      5354
--------------------------------------------------------------------
(Complete + Incomplete = Total; Imputed is the minimum across m
 of the number of filled-in observations.)
```

I select the number of imputations to be equal to 20.[4] For the results to be reproducible, select the random number seed. I selected 1234. Other control variables being used are "age age_2 age_3 urban risk_aversion education wealthindex voc_training."[5]

Now, we are ready to estimate our logistic regression as shown below. Remember to look at the imputed values of the "migrate_e" and make sure they make sense. Notice that the number of observations is now 5,354.

Step 3: Estimating the model using imputed datasets

The mi estimate command estimates the desired model with each of the 20 imputed datasets (where 20 is the number of imputations decided by the researcher) and attains 20 respective coefficients with their corresponding standard errors. Stata combines these 20 estimates to attain one coefficient, standard error, and set of inferential statistics.

```
. mi estimate : logistic migrate_e treated $controls

Multiple-imputation estimates                   Imputations       =         20
Logistic regression                             Number of obs     =      5,354
                                                Average RVI       =     0.0000
                                                Largest FMI       =     0.0000
DF adjustment:   Large sample                   DF:     min       =          .
                                                        avg       =          .
                                                        max       =          .
Model F test:       Equal FMI                   F(   9,      .)   =      16.19
Within VCE type:          OIM                   Prob > F          =     0.0000

------------------------------------------------------------------------------
   migrate_e | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     treated |   .0662572   .0575711     1.15   0.250      -.04658    .1790944
         age |   .3068913   .1318661     2.33   0.020     .0484385    .5653441
       age_2 |  -.0112538   .0043636    -2.58   0.010    -.0198063   -.0027013
       age_3 |   .0001161   .0000455     2.55   0.011     .0000268    .0002053
       urban |   .6880293   .0696607     9.88   0.000     .5514967    .8245619
risk_aversion|   .0352542   .0289156     1.22   0.223    -.0214194    .0919277
   education |  -.0084242   .0101639    -0.83   0.407    -.0283451    .0114966
 wealthindex |   .0625688   .0297909     2.10   0.036     .0041797    .1209579
 voc_training|  -.1548266   .1099523    -1.41   0.159    -.3703291    .0606758
       _cons |   -2.96503   1.261716    -2.35   0.019    -5.437948   -.4921122
------------------------------------------------------------------------------
```

To sum up the steps that were implemented above:

1) Select the data you want to use, load it, and set it for use with "mi".
2) Check which variables contain missing observations and thus those that need to be imputed.
3) Select the imputation method depending on the type of the variable to be imputed. For example, use standard OLS for a continuous variable.
4) Select the number of imputations. T. Von Hippel (2020) provide details on how many imputations should be chosen.
5) Select the random number seed for the results to be reproducible.
6) Impute the missing values of the variable(s) using mi impute.
7) Estimate the desired model using mi estimate.

[4] This is selected arbitrarily here, but this number could be given more thought. Check, for example, T. Von Hippel (2020) for more details on how many imputations you need.
[5] All control variables are baseline variables. "age" is the age of the respondent, "age_2" is is age to the power 2, "age_3" is age to the power 3, "urban" is a binary variable for whether one lives in an urban or rural area, "risk_aversion" is an index that indicated the risk aversion of an individual (all indices are calculated based on a set of survey questions), "education" demonstrated the highest level of education reached at school, "wealthindex" is an index that indicated the respondent's wealth, "voc_training" is a binary variable indicating whether the respondent has had any vocational training.

References:

Blattman, Christopher; Fiala, Nathan; Martinez, Sebastian, 2014, "Northern Uganda Social Action Fund - Youth Opportunities Program", https://doi.org/10.7910/DVN/27898, Harvard Dataverse, V1.

Blattman, Christopher; Fiala, Nathan; Martinez, Sebastian, 2019, "The long term impacts of grants on poverty: 9-year evidence from Uganda's Youth Opportunities Program", https://doi.org/10.7910/DVN/V0N0HA, Harvard Dataverse, V1.

Stata. 2021. "Multiple-Imputation Reference Manual," 388.

Razan Amine, Research Assistant in Economics

03 November 2021