Multiple Imputation & fiml with xtdpdml - DRAFT

Last revised July 14, 2016
Richard Williams

Our experience to date suggests that it is almost always easier to use fiml (Full Information Maximum Likelihood) than Multiple Imputation to deal with missing data. Nonetheless, under at least some circumstances xtdpdml can be made to work with multiple imputation. How well and how often it works remain to be seen. I'll start by summarizing the results of some analyses done with the wages data and then show the code and output.

The following scenarios are considered:

0: No missing data. This is the baseline for assessing MI and fiml approaches, since you know what the results would be if data were not missing. In later scenarios every 10th value of union is set to missing, which causes 60% of the cases to be lost via listwise deletion if nothing is done to deal with missing data.

1: Data are in long format. Imputation needs to be done and an mi xtset long data set needs to be created. I use xtdpdml to facilitate this but an mi xtset long data set could be created other ways.

2: Data are already in mi xtset imputed long format.

3: Data are not imputed. fiml is used instead.

Key findings:

- The 2 multiple imputation approaches work identically (which they should, since the same imputation was done in each) and very well, even though the imputation model was probably sub-optimal. I may have gotten lucky in this case since union is highly correlated across time. The MI approaches shown here need to be tested in other situations, e.g. what happens when there is a lot more missing data?
- fiml also works extremely well in this case and was certainly far simpler and quicker than mi was.
- But, if fiml doesn't work well for some reason, maybe mi would be good. fiml has not been working well with unbalanced panels so maybe mi would work better. This is the main reason we are interested in checking out mi further.

Potential challenges:

- It is up to the user to figure out what the imputation model should be. This may be the toughest task if you want to use mi with xtdpdml. My guess is you should first reshape wide and the imputation model should include lagged and future values of the variable to be imputed. Indeed, maybe the imputation model should include every (wide) variable that appears somewhere in the model. Stata has suggestions for doing imputation with Panel Data at http://www.stata.com/support/faqs/statistics/clustering-and-mi-impute/.
- xtdpdml can be slow enough as it is, so God only knows how long it will take if you have 50 imputations. Then again, maybe having complete data could speed things up.

- xtdpdml wants t = 1, 2, 3, …, T. If T isn't numbered that way tfix will normally renumber t for you. I am not sure if that will work with imputed data. If not you should renumber t yourself before using xtdpdml.
- The highlights formatting gets zapped when you use mi, so there is a HUGE amount of data to wade through.

Here is the current code. You can copy and paste it if you want to run it yourself.

```
clear all
set more off
version 13.1

// Scenario 0: No missing data. Use this as a baseline for evaluating
// subsequent results.
use http://www3.nd.edu/~rwilliam/statafiles/wages, clear
xtdpdml wks L.lwage, inv(ed) pre(L.union)

*********************************************************

// Scenario 1: Data are in long format and need to be imputed.
// You don't have to use xtdpdml to create the imputed data
// but it may help.

// Step 1: Open the long, unimputed version of data.
use http://www3.nd.edu/~rwilliam/statafiles/wages, clear
// I am creating MD since there is none but normally you
// would not do this!
replace union = . if _n/10 == int(_n/10)

// Step 2: Do a dryrun, keeping the data in wide format
xtdpdml wks L.lwage, inv(ed) pre(L.union) dryrun staywide

// Step 3: Create the imputed data set. The user will have to figure
// out how best to do this. The mi impute command here is probably
// sub-optimal but works ok in this case.
mi set mlong
mi register imputed union1 union2 union3 union4 union5 union6 union7
mi xtset, clear
mi impute chained (logit) union1 union2 union3 union4 union5 union6 union7 ///
   , add(20) rseed(2232)
quietly mi reshape long wks union lwage, i(id) j(t)
mi xtset id t
// Note: data are now xtset in imputed long format

// step 4: run xtdpdml
mi estimate, dots cmdok: xtdpdml wks L.lwage, inv(ed) pre(L.union)

// Save the data for later use. You can change the directory if you want.
saveold wagesmilong, version(12) replace

*********************************************************

// Scenario 2: Data have already been imputed and are in wide format.

// Step 1: use the mi imputed long data. Tweak the sysuse command if needed.
sysuse wagesmilong, clear

// Step 2: Now run mi estimate with the imputed data set
mi estimate, dots cmdok: xtdpdml wks L.lwage, inv(ed) pre(L.union)
```

```
*********************************************************

// Scenario 3: Just use fiml instead

use http://www3.nd.edu/~rwilliam/statafiles/wages, clear
// I am creating MD since there is none but normally you
// would not do this!
replace union = . if _n/10 == int(_n/10)

// fiml not used -- 60% of cases lost, estimates are quite a bit off.
xtdpdml wks L.lwage, inv(ed) pre(L.union)

// fiml used -- works extremely well in this case
xtdpdml wks L.lwage, inv(ed) pre(L.union) fiml

*********************************************************
```

Selected output:

```
. // Scenario 0: No missing data. Use this as a baseline for evaluating
. // subsequent results.
. use http://www3.nd.edu/~rwilliam/statafiles/wages, clear
. xtdpdml wks L.lwage, inv(ed) pre(L.union)

Highlights parameterization:
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wks          |
         wks |
         L1. |   .1871266   .0201939     9.27   0.000     .1475473    .2267059
             |
       lwage |
         L1. |   .6417879   .4842305     1.33   0.185    -.3072865    1.590862
             |
       union |
         L1. |   -1.19136   .5168948    -2.30   0.021    -2.204455   -.1782652
             |
          ed |  -.1122268   .0559478    -2.01   0.045    -.2218824   -.0025712
------------------------------------------------------------------------------
# of units = 595. # of periods = 7. First dependent variable is from period 2.
LR test of model vs. saturated: chi2(71)  =     110.23, Prob > chi2 =  0.0020
Wald test of all coeff = 0: chi2(4) =      90.09, Prob > chi2 =  0.0000

. // Scenario 1: Data are in long format and need to be imputed.

-----------------------------------------------------------------------------------
                  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+----------------------------------------------------------------
Structural        |
  wks2 <-         |
             wks1 |   .1876171   .0202634     9.26   0.000     .1479016    .2273326
           lwage1 |   .6471068   .4846645     1.34   0.182    -.3028183    1.597032
           union1 |  -1.176531   .5414295    -2.17   0.030    -2.238533   -.1145291
               ed |  -.1117152   .0565429    -1.98   0.048    -.2225409   -.0008895
```

```
. // Scenario 2: Data have already been imputed and are in wide format.

-----------------------------------------------------------------------------------
                  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+----------------------------------------------------------------
Structural        |
  wks2 <-         |
             wks1 |   .1876171   .0202634     9.26   0.000     .1479016    .2273326
           lwage1 |   .6471068   .4846645     1.34   0.182    -.3028183    1.597032
           union1 |  -1.176531   .5414295    -2.17   0.030    -2.238533   -.1145291
               ed |  -.1117152   .0565429    -1.98   0.048    -.2225409   -.0008895

. // Scenario 3: Just use fiml instead
. // fiml not used -- 60% of cases lost, estimates are quite a bit off.
. xtdpdml wks L.lwage, inv(ed) pre(L.union)

Highlights parameterization:
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wks          |
         wks |
         L1. |   .2482778   .0328697     7.55   0.000     .1838544    .3127013
             |
       lwage |
         L1. |   1.612868   .8690274     1.86   0.063    -.0903948     3.31613
             |
       union |
         L1. |  -.3461451   .8655696    -0.40   0.689     -2.04263     1.35034
             |
          ed |  -.2082106   .0871731    -2.39   0.017    -.3790667   -.0373544
------------------------------------------------------------------------------
# of units = 238. # of periods = 7. First dependent variable is from period 2.
LR test of model vs. saturated: chi2(71)  =     167.22, Prob > chi2 =  0.0000
Wald test of all coeff = 0: chi2(4) =      62.63, Prob > chi2 =  0.0000

. // fiml used -- works extremely well in this case
. xtdpdml wks L.lwage, inv(ed) pre(L.union) fiml

Highlights parameterization:
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wks          |
         wks |
         L1. |   .1874704   .0202407     9.26   0.000     .1477993    .2271415
             |
       lwage |
         L1. |   .6512625   .4848332     1.34   0.179    -.2989931    1.601518
             |
       union |
         L1. |  -1.181285   .5566962    -2.12   0.034    -2.272389   -.0901801
             |
          ed |   -.112133    .056974    -1.97   0.049    -.2237999   -.0004661
------------------------------------------------------------------------------
# of units = 595. # of periods = 7. First dependent variable is from period 2.
LR test of model vs. saturated: chi2(71)  =     111.85, Prob > chi2 =  0.0014
Wald test of all coeff = 0: chi2(4) =      89.16, Prob > chi2 =  0.0000
```