An informal tutorial on the `ice` command for chained equations imputation in Stata

Patrick Royston, MRC Clinical Trials Unit, London

Nordic Stata Users' Meeting, September 7th, 2007

Multiple, multivariate imputation with `ice`: the basic algorithm

MRC Clinical Trials Unit

- The imputation model is x_1, …, x_k
- Some observations are assumed to be missing at random (MAR)
- Initialise – fill in missing values at random
- Apply `uvis` to x_1 regressing on x_2, …, x_k
- Replace missing values in x_1
- Repeat for x_2 , …, x_k on the other x's (cycle 1)
- Repeat for about 10 cycles
- Do m times to give m imputed datasets with complete observations

We might have started like this:

```
ice varlist [if exp] [in range] [weight],
  [ saving(filename [, replace]) m(#)
  cmd(cmdlist) cycles(#) boot[(varlist)]
  draw[(varlist)] seed(#) dryrun
  eq(eqlist) passive(passivelist)
  noshoweq substitute(sublist)
  interval(intlist) replace
  genmiss(string) dropmissing
  other_options ]
```

Instead, we will start with a real-life problem:

MRC Clinical Trials Unit

Variable	Obs	Mean	Std. Dev.	Min	Max
survtime	347	395.44	431.53	2.00	2268.00
censdead	347	0.93	0.26	0.00	1.00
age	347	58.62	10.12	28.92	79.89
sex	347	0.32	0.47	0.00	1.00
who	347	2.97	0.72	2.00	4.00
rem	347	1.57	0.50	1.00	2.00
mets	346	1.84	0.37	1.00	2.00
lung	346	0.45	0.50	0.00	1.00
res	256	4.28	0.64	2.00	5.00
grade	224	3.49	0.87	2.00	5.00
t_mt	346	129.42	421.11	0.00	6405.00
wcc	324	8.72	4.14	3.10	55.20
haem	324	12.27	1.91	7.60	17.50
scal	306	2.46	0.21	2.10	3.74
wt	268	73.16	13.45	29.80	132.00
maxdtu	204	9.70	3.57	2.50	22.00
esr	169	49.61	35.06	1.00	154.00
visc	50	1.90	0.26	1.50	2.50

Prognostic factors in advanced kidney cancer (MRC RE01 trial)

- 350 patients, of whom 347 had follow-up information and 322 died
- 16 potential prognostic factors
  - 12 with some missing data
  - 4 binary
  - 3 ordered categorical
  - 0 nominal
  - 9 continuous
- Survival time, censoring indicator

Steps of an `ice` analysis

- Discard observations with missing response
- Summarise the variables
- Weed out variables with "too much" missingness
  - !!Subjective
  - Some use rule of thumb that >50% missing is unacceptable
- Distinguish types of variable
  - binary, ordinal, nominal, continuous, time-to-event with censoring
  - Treated differently
- Construct the `ice` command and run it.

MRC Clinical Trials Unit

Binary variables

- Imputed using logistic regression
- Consider discarding sparse variables
  - Hardly any 1's (or 0's)
  - Useless predictors in many analyses
  - Can give problems in imputation
- Otherwise generally trouble-free
  - Be aware of the "perfect prediction" problem (Ian White)
  - Fixed automatically for logistic regression in ice
  - Could affect ordinal logistic regression

Ordinal variables

- **Good choice of regression command is often `ologit`**
  - **note: *not* the default – that is `mlogit`**
- **(May) need to deal with dummy variables**
  - **Include in main *varlist***
  - **May need `passive()` and `substitute()` options**
- **Generally fairly trouble-free**

MRC Clinical Trials Unit

Nominal (unordered categorical) variables

- **Default and only sensible choice of regression command is `mlogit`**
- **Need to deal with dummy variables correctly**
  - **Include in main *varlist***
  - **Use `passive()` and `substitute()` options**
- **May need to combine sparse categories**

Continuous variables

- **Normality assumed – must be checked**
- **If necessary, transform variable to approximate normality**
- **Stata's `lnskew0` command is useful here**
  - **Must back-transform after imputation**
  - **Check range of imputed values is valid**
- **If can't find suitable transformation, use the `match()` option**
  - **no longer assumes normality**

Time-to-event with censoring

- To reduce bias, essential to include the outcome variable in the chained-equation imputation models
- What functional form for _t in model?
- Include _d?
- Van Buuren et al (1999): use ln(_t) & _d
  - No theoretical underpinning
- White (unpub): use H0(_t) and _d if have single binary variable in imputation model (for use with later Cox PH model)

Some tips

- **Use the `dryrun` option and carefully check the equations `ice` has created**
- **Sometimes you want to "tailor" specific equations according to subject-matter knowledge or detailed investigation**
  - **Use the `eq()` option to do this**
- **Compare the distribution of imputed and observed values – they should generally be roughly similar**
  - **Use the `genmiss()` option to mark the missing observations**

Now we'll do an example in Stata. The finished product:

```
ice age sex who rem mets lung grade
 g2 g3 g4 res r1 r2 r4 ln_t_mt ln_wcc
 haem ln_scal ln_wt ln_maxdtu esr ln_visc
 H0 _d, saving(re01i) m(5) match(esr)
 cmd(grade:mlogit, res:ologit) genmiss(M)
 seed(1001) substitute(grade:g2 g3 g4,
  res:r1 r2 r4)
 passive(g2:grade==2 \g3:grade==3
 \g4:grade==4 \r1:res==1 \r2:res==2
 \r4:res==4)
```