Subject: Re: Imputing missing DOBs for children
Posted by Trevor-DHS on Wed, 23 Sep 2020 18:30:03 GMT
View Forum Message <> Reply to Message

A simple solution for dates of birth where you have the year of birth but not the month is to use a random imputation of the month using something like
gen newmonth = oldmonth
replace newmonth = runiformint(1,12) if !inrange(oldmonth,1,12)
but this is crude and only works if you have at least a year of birth for all dates of birth.

if you are also missing the year of birth, then you need to do something more elaborate to constrain dates of birth, and in that case it is better to work with century month codes. For each birth you would create minimum and maximum cmcs for the dates of birth (that's the tricky part), and then impute randomly within these constrained ranges
gen finalcmc = runiformint(cmcmin,cmcmax)
If you have month and year reported, cmcmin and cmcmax will be the same and the result will be the same.
If only year is reported then cmcmin and cmcmax will be the cmc of January and December of that year.
If you don't have the year of birth, then the cmcmin will be some cmc based on the date of birth of the respondent plus, say, 10, 12, 15 years (whatever you decide to use as a minimum age at which women (girls) can give birth [while 10 years is theoretically possible, consider the likelihood in your dataset, and recognize that if you use 10 years, some will be randomly be imputed at 10 years of age]), and a maximum set to the cmc of the date of interview. With this you will have a very wide range of possible values (possibly as wide as 40 years in the worst cases), so you need to constrain these ranges with other information. You can then use information about age of a child if that is provided to narrow the ranges. You can follow this with using other births in the birth history to further narrow ranges, and assume, say, 9 months between births.

See the linked paper for more details of how DHS does the imputation process.

The basic steps are 1) to create initial logical ranges for each date, 2) apply isolated constraints (such as age) to the ranges, 3) apply neighboring constraints (such as a minimum interval between births), 4) avoid overlapping ranges for dates, 5) randomly impute within the final logical ranges.

Note that you will likely get cases where the cmcmin is greater than the cmcmax for a date of birth, in which case you have an inconsistency in the data (or your rules are poorly specified)