PO-03

Using the time-closest non-missing value to replace the missing value

Yubo Gao, University of Iowa Hospitals and Clinics, Iowa City, Iowa

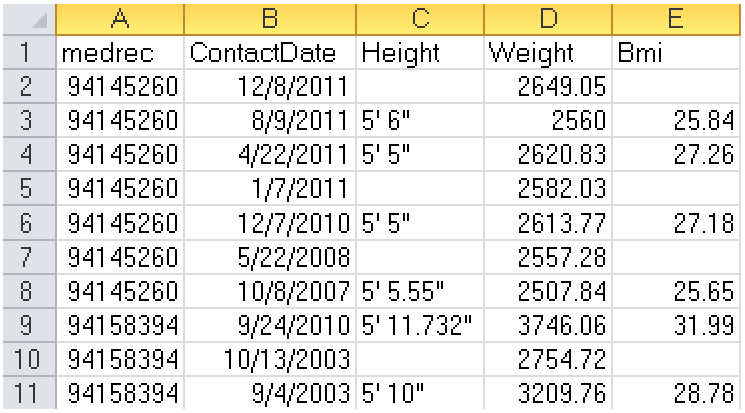
ABSTRACT

It is common that there are always some missing values for some variables in medical research project. The investigator usually chooses one of two options to deal with: ignoring the records having missing value, or imputing the missing values by some methods. Since ignoring them loses information and ultimately reduces power, some investigators like to impute them. There are several imputation methods available with varying complexity from single imputation to multiple imputation methods, performed by the procedure PROC MI. Single imputation is often used to replace the missing value of a variable in a dataset because this approach is both simple and efficient, and particularly useful when working with an extensive dataset containing millions of records and a large number of variables. In single imputation each missing value can be imputed with the variable mean (or minimum/maximum) of the complete cases or imputed by carrying forward or backward the non-missing value to an incomplete case with a single pass. There are other situations where it makes more sense to impute the missing values using the time-closest non-missing value. However, this imputation approach has received little attention in SAS community. This paper will illustrate it with a real example and present relevant SAS codes to accomplish that. SAS 9.3 was used to write the codes that were run under Microsoft Windows 7 Enterprise. The contents are appropriate for beginner and/or intermediate SAS users.

1. INITIAL DATA

Suppose we have a sample of ten records for two patients over the past several years, see below Table 1. Here, medrec is the unique medical record number for patient, Bmi represents Body Mass Index, and others are self-explanatory. In four occasions they had no height data. Therefore, the corresponding Bmi are missing. Now the principle investigator wanted to study their Bmi trends over time, and asked to impute the missing heights with time-closest non-missing heights.

Table 1



	A	B	C	D	E
1	medrec	ContactDate	Height	Weight	Bmi
2	94145260	12/8/2011		2649.05	
3	94145260	8/9/2011	5' 6"	2560	25.84
4	94145260	4/22/2011	5' 5"	2620.83	27.26
5	94145260	1/7/2011		2582.03	
6	94145260	12/7/2010	5' 5"	2613.77	27.18
7	94145260	5/22/2008		2557.28	
8	94145260	10/8/2007	5' 5.55"	2507.84	25.65
9	94158394	9/24/2010	5' 11.732"	3746.06	31.99
10	94158394	10/13/2003		2754.72	
11	94158394	9/4/2003	5' 10"	3209.76	28.78

To accomplish this, the program consists of three steps: Backward, Forward, and Comparison steps. Now these steps will be illustrated.

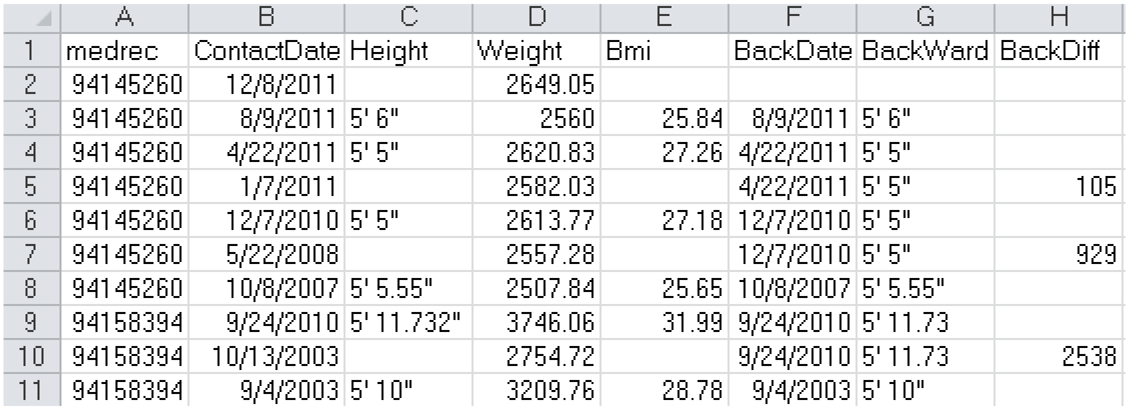
2. BACKWARD STEP

Based on the descending sorted data in Table 1 (called w5 in the code), apply the codes below, and produce the result in Table 2(called w510 in the code). The purpose is to record for observation beginning with the first non-missing observation the BackDate, BackWard value (height), and difference in days between current record and BackWard record. If its height is not missing, then its BackDate and BackWard value are its ContactDate and height. Otherwise, its BackDate and BackWard value will use retained BackDate and BackWard value.

```
DATA w510;
        set w5;
        by medrec DESCENDING contactdate;
        retain BackDate BackWard;
        if first.medrec then do; BackDate=.; BackWard=’      ‘; end;
        if height ne ‘ ‘ then do; BackDate=contactdate;BackWard=height;end;

        if (height=’ ‘ and BackDate ne .) then BackDiff=BackDate-contactdate;
        format BackDate date9.;
RUN;
```

Table 2



	A	B	C	D	E	F	G	H
1	medrec	ContactDate	Height	Weight	Bmi	BackDate	BackWard	BackDiff
2	94145260	12/8/2011		2649.05				
3	94145260	8/9/2011	5' 6"	2560	25.84	8/9/2011	5' 6"	
4	94145260	4/22/2011	5' 5"	2620.83	27.26	4/22/2011	5' 5"	
5	94145260	1/7/2011		2582.03		4/22/2011	5' 5"	105
6	94145260	12/7/2010	5' 5"	2613.77	27.18	12/7/2010	5' 5"	
7	94145260	5/22/2008		2557.28		12/7/2010	5' 5"	929
8	94145260	10/8/2007	5' 5.55"	2507.84	25.65	10/8/2007	5' 5.55"	
9	94158394	9/24/2010	5' 11.732"	3746.06	31.99	9/24/2010	5' 11.73	
10	94158394	10/13/2003		2754.72		9/24/2010	5' 11.73	2538
11	94158394	9/4/2003	5' 10"	3209.76	28.78	9/4/2003	5' 10"	

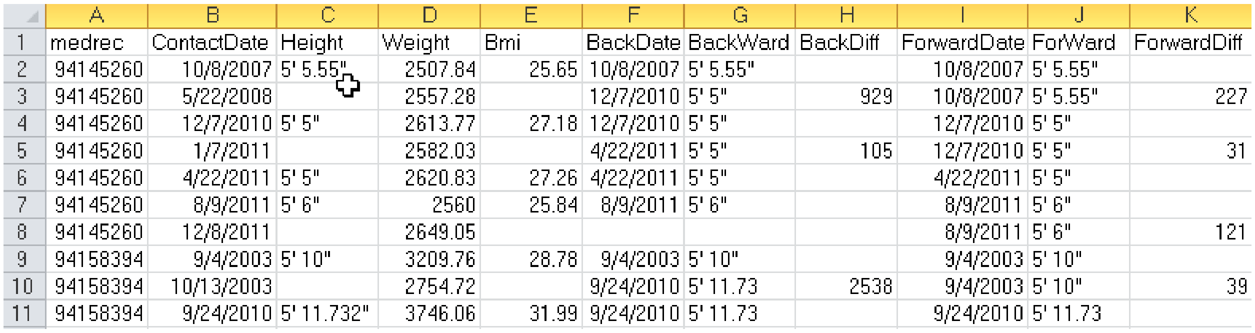
3. FORWARD STEP

Then, next do a forward step using codes below to produce Table 3. The goal is the same as in Step 2, except in other direction.

```
PROC SORT data=w510;
        by medrec contactdate;
RUN;

DATA w5121;
        set w510;
        by medrec contactdate;
        retain ForwardDate ForWard;
        if first.medrec then do; ForwardDate=.; ForWard='      '; end;
        if height ne ' ' then do; ForwardDate=contactdate; ForWard=height;end;
        if (height=' ' and ForwardDate ne .) then ForwardDiff=contactdate- ForwardDate;
        format ForwardDate date9.;
RUN;
```

Table 3



	A	B	C	D	E	F	G	H	I	J	K
1	medrec	ContactDate	Height	Weight	Bmi	BackDate	BackWard	BackDiff	ForwardDate	ForWard	ForwardDiff
2	94145260	10/8/2007	5' 5.55"	2507.84	25.65	10/8/2007	5' 5.55"		10/8/2007	5' 5.55"	
3	94145260	5/22/2008		2557.28		12/7/2010	5' 5"	929	10/8/2007	5' 5.55"	227
4	94145260	12/7/2010	5' 5"	2613.77	27.18	12/7/2010	5' 5"		12/7/2010	5' 5"	
5	94145260	1/7/2011		2582.03		4/22/2011	5' 5"	105	12/7/2010	5' 5"	31
6	94145260	4/22/2011	5' 5"	2620.83	27.26	4/22/2011	5' 5"		4/22/2011	5' 5"	
7	94145260	8/9/2011	5' 6"	2560	25.84	8/9/2011	5' 6"		8/9/2011	5' 6"	
8	94145260	12/8/2011		2649.05					8/9/2011	5' 6"	121
9	94158394	9/4/2003	5' 10"	3209.76	28.78	9/4/2003	5' 10"		9/4/2003	5' 10"	
10	94158394	10/13/2003		2754.72		9/24/2010	5' 11.73	2538	9/4/2003	5' 10"	39
11	94158394	9/24/2010	5' 11.732"	3746.06	31.99	9/24/2010	5' 11.73		9/24/2010	5' 11.73	

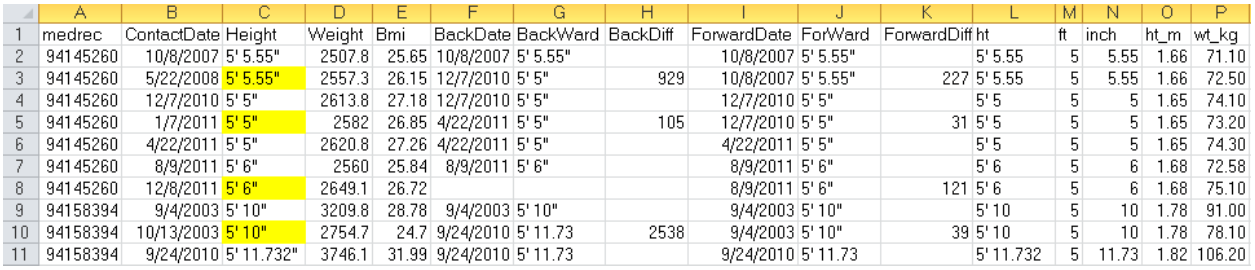
4. COMPARISON STEP

Finally, based on Table 3, for those records having missing heights, compare BackDiff and ForwardDiff, and select the imputed height from BackWard or ForWard, based on which has the smaller day difference. Use codes below to achieve these. Table 4 shows the final results.

```
DATA w51222;
        set w5121;
        if (height=' ' and ForWard ne ' ' and  BackWard ne ' ') then  /*middle*/
          do;
                if ForwardDiff<=BackDiff then height=ForWard;
                else height=BackWard;
          end;
        else if (height=' ' and ForWard ne ' ' and  BackWard = ' ') then height=ForWard;  /*top*/
        else if (height=' ' and ForWard = ' ' and  BackWard ne ' ') then height=BackWard; /*bottom*/

        /*bmi*/
        ht=compress(height,'"');
        ft=scan(ht,1,"'")+0;
        inch=scan(ht,2,"'")+0;
        ht_m=(30.48*ft+2.54*inch)/100; /*change inches into meters*/
        wt_kg=28.35*weight/1000;       /*change ounces into kilograms*/
        if bmi=. then bmi=round(wt_kg/(ht_m*ht_m),0.01);
RUN;
```

Table 4



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	medrec	ContactDate	Height	Weight	Bmi	BackDate	BackWard	BackDiff	ForwardDate	ForWard	ForwardDiff	ht	ft	inch	ht_m	wt_kg
2	94145260	10/8/2007	5' 5.55"	2507.8	25.65	10/8/2007	5' 5.55"		10/8/2007	5' 5.55"		5' 5.55	5	5.55	1.66	71.10
3	94145260	5/22/2008	5' 5.55"	2557.3	26.15	12/7/2010	5' 5"	929	10/8/2007	5' 5.55"	227	5' 5.55	5	5.55	1.66	72.50
4	94145260	12/7/2010	5' 5"	2613.8	27.18	12/7/2010	5' 5"		12/7/2010	5' 5"		5' 5	5	5	1.65	74.10
5	94145260	1/7/2011	5' 5"	2582	26.85	4/22/2011	5' 5"	105	12/7/2010	5' 5"	31	5' 5	5	5	1.65	73.20
6	94145260	4/22/2011	5' 5"	2620.8	27.26	4/22/2011	5' 5"		4/22/2011	5' 5"		5' 5	5	5	1.65	74.30
7	94145260	8/9/2011	5' 6"	2560	25.84	8/9/2011	5' 6"		8/9/2011	5' 6"		5' 6	5	6	1.68	72.58
8	94145260	12/8/2011	5' 6"	2649.1	26.72				8/9/2011	5' 6"	121	5' 6	5	6	1.68	75.10
9	94158394	9/4/2003	5' 10"	3209.8	28.78	9/4/2003	5' 10"		9/4/2003	5' 10"		5' 10	5	10	1.78	91.00
10	94158394	10/13/2003	5' 10"	2754.7	24.7	9/24/2010	5' 11.73	2538	9/4/2003	5' 10"	39	5' 10	5	10	1.78	78.10
11	94158394	9/24/2010	5' 11.732"	3746.1	31.99	9/24/2010	5' 11.73		9/24/2010	5' 11.73		5' 11.732	5	11.73	1.82	106.20

The lower part codes change height into meters, weight (in ounces) into kilograms, and calculate Bmi. Then we can do the trend analysis for Bmi.

5. CONCLUSION

In this paper, a procedure to impute the missing values using the time-closest non-missing value is presented. It can easily be extended to orther variables having missing values.

REFERENCES

Zhang,SP, Liao,J, and Zhu, XS. A SAS® Macro for Single Imputation. PharmaSUG 2008. June 1-4, 2008, Atlanta, Georgia.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the authors at:

Yubo Gao
University of Iowa Hospitals and Clinics (UIHC)
Orthopaedic Surgery, 01066 JPP
200 Hawkins Dr.
Iowa City, IA 52242
Phone: 319-356-1674
Email: yubo-gao@uiowa.edu