PT11: Transposing a SAS® dataset from 'long' to 'wide‘

MWSUG Columbus OHIO
September 22-24, 2013

Peter Batra
College of Pharmacy
University of Michigan

Abstract/Summary

- This presentation will demonstrate how to transpose a ‘long’ dataset into a ‘wide’ dataset. In the example, we start with a dataset that contains a list of up to 10 drugs assigned to a unique ID. In this dataset each record contains the data by drug. The end result is a dataset where each record (observation) contains the combined information from all drugs for each unique ID.
- Additionally, I show how to add information on the original ‘long’ file for each drug and then combine the ‘wide’ dataset with an existing ID level dataset. The end result is a very large dataset that contains all of the data at the unique ID level.
- The presentation will show all of the SAS® software code used to achieve this conversion from a 'long' to a 'wide' file. The following PROC’s and the Data step were used to achieve this goal: CONTENTS, SORT, SQL, and TRANSPOSE.
- The SAS® code used in this presentation was run on SAS® for Windows (9.2) on a 64 bit platform.
- This material in this presentation is suitable for someone who is familiar with SAS® coding and wants to add to their knowledge base with the steps involved to transpose a dataset from ‘long’ to ‘wide’.

Introduction

- We start with a ‘long’ dataset that contains our list of unique ID’s (HHID + PN) and each drug for this ID.
- Each person (ID) can have from 1 to 10 drugs listed, and each is assigned a sequential number (position) in the data set.
- Note that the each record is unique when you combine the person (ID) and the drug position number.

Pds05e_m

	A	B	C	D
1	HHID	PN	POSITION	P1DRUGNAME
2	1	10	1	ATORVASTATIN
3	1	10	2	HYDROCHLOROTHIAZIDE-LOSARTAN
4	1	10	3	EZETIMIBE
5	1	10	4	ESTRADIOL
6	1	10	5	ALENDRONATE
7	2	20	1	ATORVASTATIN
8	2	20	2	FUROSEMIDE
9	2	20	3	VALSARTAN
10	2	20	4	MOMETASONE NASAL
11	2	20	5	FLUTICASONE-SALMETEROL
12	2	20	6	CARVEDILOL
13	2	20	7	EZETIMIBE
14	2	20	8	LISINOPRIL
15	3	10	1	IRBESARTAN
16	3	10	2	DILTIAZEM
17	3	10	3	SIMVASTATIN
18	3	10	4	METFORMIN
19	4	20	1	AMLODIPINE
20	4	20	2	METOPROLOL
21	4	20	3	FUROSEMIDE
22	5	10	1	CAPTOPRIL
23	5	10	2	FLUVASTATIN
24	5	10	3	PROPRANOLOL
25	5	10	4	LATANOPROST OPHTHALMIC

Step 1: Proc Contents

- Next we look at the dataset to understand the number of variables, variable names, and number of observations.

Variables in Creation Order

#	Variable	Type	Len	Label
1	HHID	Char	6	HOUSEHOLD IDENTIFIER
2	PN	Char	3	PERSON NUMBER
3	POSITION	Num	8	MEDICATION: LIST NUMBER
5	P1DRUGNAME	Char	50	MEDICATION: DRUG NAME (EDIT FROM E1A)
6	P1BRANDNAME	Char	50	MEDICATION: BRAND NAME (EDIT FROM E1A)
7	P1PRODSTRGTH	Char	40	MEDICATION: PRODUCT STRENGTH (EDIT FROM
8	P1E1B	Char	65	MEDICATION: DATE FILLED
9	P1E1C	Char	71	MEDICATION: DOSAGE INSTRUCTIONS
10	P1E2	Char	65	MEDICATION: HOW LONG TAKING
11	P1E3	Num	8	MEDICATION: OUT-OF-POCKET COST
12	P1E5A	Num	8	MEDICATION (DIS)AGREE: IMPORTANT
13	P1E5B	Num	8	MEDICATION (DIS)AGREE: SIDE EFFECTS
14	P1E5C	Num	8	MEDICATION (DIS)AGREE: TOO EXPENSIVE
15	P1E5D	Num	8	MEDICATION (DIS)AGREE: BEST AVAILABLE
16	P1E6	Num	8	MEDICATION: MISSED DOSE IN LAST YEAR
17	P1E7A	Num	8	MEDICATION MISSED DOSE: COST
18	P1E7B	Num	8	MEDICATION MISSED DOSE: SIDE EFFECTS
19	P1E7C	Num	8	MEDICATION MISSED DOSE: AWAY FROM HOME
20	P1E7D	Num	8	MEDICATION MISSED DOSE: FORGET

Step 2: Proc SQL

- In this step we want to add columns based on each drug listed. The data to add comes from another table that has a unique record ID for each drug.

```
proc sql;
   create table PDS05inc_classid as
      select a.*, b.*
         from hrsdata.Pds05e_m a  left join hrsdata.fivemultumdrugdatauniques  b
    on upcase(Pds05e_m.p1drugname)  = upcase(fivemultumdrugdatauniques.p1_drug_name)
   order by drug  id;
      quit;
```

Pds05e_m

	A	B	C	D
1	HHID	PN	POSITION	P1DRUGNAME
2	1	10	1	ATORVASTATIN
3	1	10	2	HYDROCHLOROTHIAZIDE-LOSARTAN
4	1	10	3	EZETIMIBE
5	1	10	4	ESTRADIOL
6	1	10	5	ALENDRONATE
7	2	20	1	ATORVASTATIN
8	2	20	2	FUROSEMIDE
9	2	20	3	VALSARTAN
10	2	20	4	MOMETASONE NASAL
11	2	20	5	FLUTICASONE-SALMETEROL
12	2	20	6	CARVEDILOL
13	2	20	7	EZETIMIBE
14	2	20	8	LISINOPRIL

fivemultumdrugdatauniques

	drug_id	P1_drug_name	FiveBR1_Tclass_id_1	FiveBR1_Tclass_nm_1
1	d00212	diphenhydrAMINE	122	respiratory agents
2	d00060	methotrexate	20	antineoplastics
3	d00215	DOBUTamine	40	cardiovascular agents
4	d07761	vandetanib	20	antineoplastics
5	d00216	DOPamine	40	cardiovascular agents
6	d00174	atropine	87	gastrointestinal agents
7	d00149	lorazepam	57	central nervous system agents
8	d00468	magnesium sulfate	87	gastrointestinal agents
9	d00284	medroxyPROGESTERone	20	antineoplastics

Step 3: Proc SORT

Prior to transposing the data, we need to ensure the dataset is sorted by unique records, or observations. In this case it's by the unique id (HHID + PN) and the drug position.

PDS05inc_classid

A	B	C	D	E	F	G	H
HHID	PN	POSITION	P1DRUGNAME	FiveBR1_T	FiveBR1_Tclass_nm_1	FiveBR1_T	FiveBR1_T
101	10	1	ATORVASTATIN	358	metabolic agents	19	antihyper
101	10	2	WARFARIN	81	coagulation modifiers	82	anticoagu
101	10	3	COLESTIPOL	358	metabolic agents	19	antihyper
101	10	4	DILTIAZEM	40	cardiovascular agents	48	calcium ch
101	10	5	DOXAZOSIN	40	cardiovascular agents	43	antiadren
101	10	6	GEMFIBROZIL	358	metabolic agents	19	antihyper
102	10	1	LISINOPRIL	40	cardiovascular agents	42	angiotens
102	10	2	GLIPIZIDE	358	metabolic agents	99	antidiabet
103	30	1	ATENOLOL	40	cardiovascular agents	47	beta-adre
103	30	2	AMLODIPINE	40	cardiovascular agents	48	calcium ch
104	10	1	URSODIOL	87	gastrointestinal agents	92	gallstone
104	10	2	LANSOPRAZOLE	87	gastrointestinal agents	272	proton pu
104	10	3	ISOSORBIDE DINITRATE	40	cardiovascular agents	45	antiangina
104	10	4	DOCUSATE	87	gastrointestinal agents	95	laxatives
104	10	5	PAROXETINE	242	psychotherapeutic agents	249	antidepre
104	10	6	FERROUS SULFATE	115	nutritional products	116	iron produ
104	10	7	CALCIUM-VITAMIN D	115	nutritional products	120	vitamin ar
104	10	8	LOSARTAN	40	cardiovascular agents	56	angiotens
104	10	9	MECLIZINE	57	central nervous system ag	65	antiemeti
104	10	10	CLOPIDOGREL	81	coagulation modifiers	83	antiplatel

Step 4: Proc Transpose

- For each variable in our ‘long’ dataset we need to have a separate Proc Transpose step. (43 variables in my example.)
- The created file will be wide and contain variables for the drug position number 1 to 10, where available.

```
proc transpose data=pds05inc_classid
               out=fivewidep1e1a
               prefix=p1e1a;
 by hhid pn;
 id position;
 var p1e1a;
run;
```

Output from Proc SQL step

Wide created dataset

1,2,3...10

	HOUSEHOLD IDENTIFIER	PERSON NUMBER	NAME OF FORMER VARIABLE	LABEL OF FORMER VARIABLE	p1e1a1	p1e1a2	p1e1a3
1	000003	010	P1E1A	MEDICATION: NAME AND DOSAGE	Lipitor 80mg tablets	Warfarin 5mg tablets	Colestid 1 gr
2	010063	010	P1E1A	MEDICATION: NAME AND DOSAGE	Lisinopril 20mg	Glipizide 10mg	
3	010075	030	P1E1A	MEDICATION: NAME AND DOSAGE	Atenolol 100mg	Norvasc 5mg	
4	010210	010	P1E1A	MEDICATION: NAME AND DOSAGE	Ursodiol 300mg	Prevacid 30mg	Isosorbide D
5	010210	020	P1E1A	MEDICATION: NAME AND DOSAGE	Zocor	Metoprolol	
6	010281	010	P1E1A	MEDICATION: NAME AND DOSAGE	Aciphex 20mg		
7	010378	010	P1E1A	MEDICATION: NAME AND DOSAGE	Taztia XT 180mg CAP	Estraderm Calendar Pack 8's 0.05mg	Micardis 30's
8	010394	010	P1E1A	MEDICATION: NAME AND DOSAGE	Uroxatral 10mg		
9	010395	010	P1E1A	MEDICATION: NAME AND DOSAGE	Allopurinol 300mg	Zoloft 100mg	Lisinopril 20m
10	010395	020	P1E1A	MEDICATION: NAME AND DOSAGE	Levoxyl 05mg	Lexapro 20mg	Torsemide 2

Step 5: Merge in Data step

```
data pds05wide ;
merge
 fivewidep1e1a (drop=_label_ drop=_name_)
 fivewidep1drugname (drop=_label_ drop=_name_)
 fivewidep1brandname (drop=_label_ drop=_name_)
 fivewideP1E2 (drop=_label_ drop=_name_)
 fivewideP1E3 (drop=_label_ drop=_name_)
 fivewideP1E6 (drop=_label_ drop=_name_)
 fivewideP1E1b (drop=_label_ drop=_name_)
 fivewideP1E1c (drop=_label_ drop=_name_)
 fivewideP1E5a (drop=_label_ drop=_name_)
 fivewideP1E5b (drop=_label_ drop=_name_)
 fivewidep1E5c (drop=_label_ drop=_name_)
 ...
 fivewidebr4_tclass_id_3 (drop=_name_)
 fivewidebr4_tclass_nm_3 (drop=_name_) ;
  by hhid pn ;
run;
```

- Now, we'll take all of the individual files created for each variable and merge those together into one large wide file.
- Note the 'dropping' of **_label** and **_name** to remove extraneous information from the created file

Step 6: Proc SQL

```
proc sql;
        create table HRSATHSMA as
            select

a.kc005, a.kc006, a.kc008, a.kc010, a.kc011, a.kc012, a.kc015, a.kc018, a.kc030, a.kc031, a.kc032, a.kc033,
b.*,
c.lc005, c.lc006, c.lc008, c.lc010, c.lc011, c.lc012, c.lc015, c.lc018, c.lc030, c.lc031, c.lc032, c.lc033, c.lc034,
d.mc005, d.mc006, d.mc008, d.mc010, d.mc011, d.mc012, d.mc015, d.mc018, d.mc030, d.mc031,
e.DA_06, e.BC_06, e.NA_06, e.DA_08, e.BC_08, e.NA_08,
f.ki807, f.ki808, f.ki809, f.ki812, f.ki815, f.ki816, f.ki851, f.ki852, f.ki853,
g.li807, g.li808, g.li809, g.li812, g.li815, g.li816, g.li851, g.li852, g.li853,
h.mi807, h.mi808, h.mi809, h.mi812, h.mi815, h.mi816, h.mi851, h.mi852, h.mi853,
i.kg014, i.kg016, i.kg021, i.kg023, i.kg025, i.kg030,
i.kg041, i.kg042, i.kg043, i.kg044, i.kg045, i.kg046, i.kg047, i.kg050, i.kg051, i.kg052, i.kg053, i.kg059, i.kg209,
j.lg041, j.lg042, j.lg043, j.lg044, j.lg045, j.lg046, j.lg047, j.lg050, j.lg051, j.lg052, j.lg053, j.lg059, j.lg209,
k.mg041, k.mg044, k.mg047, k.mg050,  k.mg209,
l.kn362,
m.ln362,
n.mn362,
o.*,
p.*,
q.*,
r.*,
s.*,
t.*,
u.h7itot, u.h8itot, u.h9itot,
v.regionb, region04, region06, region08, urbrur04, urbrur06, urbrur08, D04to06, D06to08

        from
          HRSdata.h06c_r a
        left join HRSdata.trk2010 b                          on (b.hhid=a.hhid and b.PN=a.PN)
        left join HRSdata.h08c_r c                           on (c.hhid=a.hhid and c.PN=a.PN)
        left join HRSdata.h10c_r d                           on (d.hhid=a.hhid and d.PN=a.PN)
        left join HRSdata.bigcog2008f_docrptaug26 e          on (e.hhid=a.hhid and e.PN=a.PN)
        left join HRSdata.h06i_r f                           on (f.hhid=a.hhid and f.PN=a.PN)
        left join HRSdata.h08i_r g                           on (g.hhid=a.hhid and g.PN=a.PN)
        left join HRSdata.h10i_r h                           on (h.hhid=a.hhid and h.PN=a.PN)
        left join HRSdata.h06g_r i                           on (i.hhid=a.hhid and i.PN=a.PN)
        left join HRSdata.h08g_r j                           on (j.hhid=a.hhid and j.PN=a.PN)
        left join HRSdata.h10g_r k                           on (k.hhid=a.hhid and k.PN=a.PN)
        left join HRSdata.h06n_r l                           on (l.hhid=a.hhid and l.PN=a.PN)
        left join HRSdata.h08n_r m                           on (m.hhid=a.hhid and m.PN=a.PN)
        left join HRSdata.h10n_r n                           on (n.hhid=a.hhid and n.PN=a.PN)
        left join HRSdata.PDS05a_R o                         on (o.hhid=a.hhid and o.PN=a.PN)
        left join work.pds05wide p                           on (p.hhid=a.hhid and p.PN=a.PN)
        left join HRSdata.PDS05S_R q                         on (q.hhid=a.hhid and q.PN=a.PN)
        left join HRSdata.PDS07a_R r                         on (r.hhid=a.hhid and r.PN=a.PN)
        left join work.pds07wide s                           on (s.hhid=a.hhid and s.PN=a.PN)
        left join HRSdata.PDS07S_R t                         on (t.hhid=a.hhid and t.PN=a.PN)
        left join work.Randdata u                            on (u.hhid=a.hhid and u.PN=a.PN)
        left join HRSdata.hrsxregion v                       on (v.hhid=a.hhid and v.PN=a.PN);
quit;
```

In this final step we combine other Unique ID information to our created ‘wide’ file.

Final Result

The final result is our ‘wide’ file:

	A	B	C	D	E	F	G	H	I	J	K	L
1	HHID	PN	p1drugname1	p1drugname2	p1drugname3	p1drugname4	p1drugname5	p1drugname6	p1drugname7	p1drugname8	p1drugname9	p1drugname10
2	1	010	ATORVASTATIN	HYDROCHLOROTHIAZIDE-LOSARTAN	EZETIMIBE	ESTRADIOL	ALENDRONATE					
3	2	020	ATORVASTATIN	FUROSEMIDE	VALSARTAN	MOMETASONE NASAL	FLUTICASONE-SALMETEROL	CARVEDILOL	EZETIMIBE	LISINOPRIL		
4	3	010	IRBESARTAN	DILTIAZEM	SIMVASTATIN	METFORMIN						
5	4	020	AMLODIPINE	METOPROLOL	FUROSEMIDE							
6	5	010	CAPTOPRIL	FLUVASTATIN	PROPRANOLOL	LATANOPROST OPHTHALMIC						

	A	B	C	D	
1	HHID	PN	POSITION	P1DRUGNAME	
2	1	10	1	ATORVASTATIN	
3	1	10	2	HYDROCHLOROTHIAZIDE-LOSARTAN	
4	1	10	3	EZETIMIBE	
5	1	10	4	ESTRADIOL	
6	1	10	5	ALENDRONATE	2
7	2	20	1	ATORVASTATIN	
8	2	20	2	FUROSEMIDE	
9	2	20	3	VALSARTAN	
10	2	20	4	MOMETASONE NASAL	
11	2	20	5	FLUTICASONE-SALMETEROL	3
12	2	20	6	CARVEDILOL	
13	2	20	7	EZETIMIBE	
14	2	20	8	LISINOPRIL	
15	3	10	1	IRBESARTAN	
16	3	10	2	DILTIAZEM	
17	3	10	3	SIMVASTATIN	4
18	3	10	4	METFORMIN	
19	4	20	1	AMLODIPINE	
20	4	20	2	METOPROLOL	5
21	4	20	3	FUROSEMIDE	
22	5	10	1	CAPTOPRIL	
23	5	10	2	FLUVASTATIN	
24	5	10	3	PROPRANOLOL	6
25	5	10	4	LATANOPROST OPHTHALMIC	