Data Cleaning involving duplicate IDs and duplicate records

Lecture 10
July 26, 2005

Kate Cowles
374 SH
kcowles@stat.uiowa.edu

Duplicates

- may need to check for either duplicate ID codes or duplicate observations
- duplicate observations should just be eliminated
- duplicate IDs with different data values need to be resolved

Patients file: duplicate records added

<table>
<thead>
<tr>
<th>ID</th>
<th>Date</th>
<th>Gender</th>
<th>Age</th>
<th>Diagnosis Code</th>
<th>Country</th>
<th>State</th>
<th>Zip Code</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>002F</td>
<td>13/1998</td>
<td>F</td>
<td>78</td>
<td>84120</td>
<td></td>
<td></td>
<td>x0</td>
<td></td>
</tr>
<tr>
<td>003F</td>
<td>11/11/1998</td>
<td>F</td>
<td>78</td>
<td>84120</td>
<td></td>
<td></td>
<td>x0</td>
<td></td>
</tr>
<tr>
<td>003T</td>
<td>11/11/1998</td>
<td>T</td>
<td>78</td>
<td>84120</td>
<td></td>
<td></td>
<td>x0</td>
<td></td>
</tr>
<tr>
<td>004F</td>
<td>31/1998</td>
<td>F</td>
<td>31</td>
<td>68190100</td>
<td></td>
<td></td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>005M</td>
<td>12/1999</td>
<td>M</td>
<td>74</td>
<td>58112</td>
<td></td>
<td></td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>006F</td>
<td>01/01/1999</td>
<td>F</td>
<td>100</td>
<td>80120</td>
<td></td>
<td></td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>006M</td>
<td>12/1998</td>
<td>M</td>
<td>80</td>
<td>68120</td>
<td></td>
<td></td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>006F</td>
<td>07/07/1999</td>
<td>F</td>
<td>84</td>
<td>82148</td>
<td></td>
<td></td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>007M</td>
<td>11/1998</td>
<td>M</td>
<td>80</td>
<td>90190100</td>
<td></td>
<td></td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>008F</td>
<td>08/09/1998</td>
<td>F</td>
<td>70</td>
<td>8820100</td>
<td></td>
<td></td>
<td>8820100</td>
<td></td>
</tr>
<tr>
<td>009M</td>
<td>25/1999</td>
<td>M</td>
<td>41</td>
<td>88240180</td>
<td></td>
<td></td>
<td>88240180</td>
<td></td>
</tr>
<tr>
<td>010F</td>
<td>19/1999</td>
<td>F</td>
<td>10</td>
<td>40120</td>
<td></td>
<td></td>
<td>40120</td>
<td></td>
</tr>
<tr>
<td>011F</td>
<td>13/1998</td>
<td>F</td>
<td>20</td>
<td>68300</td>
<td></td>
<td></td>
<td>68300</td>
<td></td>
</tr>
<tr>
<td>012M</td>
<td>9/1998</td>
<td>M</td>
<td>74</td>
<td>60122</td>
<td></td>
<td></td>
<td>60122</td>
<td></td>
</tr>
<tr>
<td>013M</td>
<td>23/1998</td>
<td>M</td>
<td>64</td>
<td>74108</td>
<td></td>
<td></td>
<td>74108</td>
<td></td>
</tr>
<tr>
<td>014F</td>
<td>02/1999</td>
<td>F</td>
<td>90</td>
<td>22130</td>
<td></td>
<td></td>
<td>22130</td>
<td></td>
</tr>
<tr>
<td>002F</td>
<td>13/1998</td>
<td>F</td>
<td>78</td>
<td>84120</td>
<td></td>
<td></td>
<td>x0</td>
<td></td>
</tr>
<tr>
<td>003M</td>
<td>12/1999</td>
<td>M</td>
<td>74</td>
<td>58112</td>
<td></td>
<td></td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>003F</td>
<td>12/1999</td>
<td>F</td>
<td>74</td>
<td>82148</td>
<td></td>
<td></td>
<td>82148</td>
<td></td>
</tr>
<tr>
<td>015F</td>
<td>01/01/1999</td>
<td>F</td>
<td>51</td>
<td>88240180</td>
<td></td>
<td></td>
<td>88240180</td>
<td></td>
</tr>
<tr>
<td>016F</td>
<td>05/1999</td>
<td>F</td>
<td>20</td>
<td>82148</td>
<td></td>
<td></td>
<td>82148</td>
<td></td>
</tr>
<tr>
<td>017M</td>
<td>07/1999</td>
<td>M</td>
<td>70</td>
<td>58118</td>
<td></td>
<td></td>
<td>58118</td>
<td></td>
</tr>
<tr>
<td>018F</td>
<td>12/1999</td>
<td>F</td>
<td>50</td>
<td>1000000</td>
<td></td>
<td></td>
<td>1000000</td>
<td></td>
</tr>
<tr>
<td>019F</td>
<td>06/1999</td>
<td>F</td>
<td>80</td>
<td>82148</td>
<td></td>
<td></td>
<td>82148</td>
<td></td>
</tr>
<tr>
<td>020M</td>
<td>12/1999</td>
<td>M</td>
<td>60</td>
<td>80120</td>
<td></td>
<td></td>
<td>80120</td>
<td></td>
</tr>
<tr>
<td>021F</td>
<td>09/1999</td>
<td>F</td>
<td>51</td>
<td>88240180</td>
<td></td>
<td></td>
<td>88240180</td>
<td></td>
</tr>
<tr>
<td>022F</td>
<td>10/1999</td>
<td>F</td>
<td>82</td>
<td>48114</td>
<td></td>
<td></td>
<td>48114</td>
<td></td>
</tr>
<tr>
<td>023F</td>
<td>31/1998</td>
<td>F</td>
<td>78</td>
<td>22134</td>
<td></td>
<td></td>
<td>22134</td>
<td></td>
</tr>
<tr>
<td>024F</td>
<td>09/1998</td>
<td>F</td>
<td>80</td>
<td>1200120</td>
<td></td>
<td></td>
<td>1200120</td>
<td></td>
</tr>
<tr>
<td>025F</td>
<td>01/1999</td>
<td>F</td>
<td>68</td>
<td>74102</td>
<td></td>
<td></td>
<td>74102</td>
<td></td>
</tr>
<tr>
<td>026F</td>
<td>31/1998</td>
<td>F</td>
<td>70</td>
<td>166106</td>
<td></td>
<td></td>
<td>166106</td>
<td></td>
</tr>
<tr>
<td>027F</td>
<td>28/1998</td>
<td>F</td>
<td>90</td>
<td>66150</td>
<td></td>
<td></td>
<td>66150</td>
<td></td>
</tr>
</tbody>
</table>
Eliminating duplicates using PROC SORT

- NODUPKEY option on PROC SORT eliminates multiple observations where the BY variables have same value
- OUT= option is used to create new data set, leaving original data set unchanged
- when NODUPKEY removes multiple observations, the only indication is in the NOTE in the SAS log
- looking for duplicate IDs with NODUPKEY in PROC SORT is useful only if the SAS log shows that NO duplicates were remove
  -- otherwise, you need to see which IDs had duplicate data and the nature of the data
- if NODUPKEY is used with more than one BY variable, only those observations with identical values on all the BY variables will be deleted

Data Set SINGLE - Duplicated ID's Removed from PATIENTS

<table>
<thead>
<tr>
<th>PATNO</th>
<th>GENDER</th>
<th>VISIT</th>
<th>HR</th>
<th>SBP</th>
<th>DBP</th>
<th>DX</th>
<th>AE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>11/11/98</td>
<td>90</td>
<td>190</td>
<td>100</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>001</td>
<td>M</td>
<td>11/11/98</td>
<td>88</td>
<td>140</td>
<td>80</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>002</td>
<td>F</td>
<td>11/13/98</td>
<td>84</td>
<td>120</td>
<td>78</td>
<td>X</td>
<td>0</td>
</tr>
<tr>
<td>003</td>
<td>X</td>
<td>10/21/98</td>
<td>68</td>
<td>190</td>
<td>100</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>004</td>
<td>F</td>
<td>01/01/99</td>
<td>101</td>
<td>200</td>
<td>120</td>
<td>5</td>
<td>A</td>
</tr>
<tr>
<td>006</td>
<td></td>
<td>06/15/99</td>
<td>72</td>
<td>102</td>
<td>68</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>007</td>
<td></td>
<td></td>
<td>88</td>
<td>148</td>
<td>102</td>
<td></td>
<td></td>
</tr>
<tr>
<td>008</td>
<td>F</td>
<td>08/08/98</td>
<td>210</td>
<td></td>
<td></td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>009</td>
<td>M</td>
<td>09/25/99</td>
<td>86</td>
<td>240</td>
<td>180</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>010</td>
<td>F</td>
<td>10/19/99</td>
<td></td>
<td>40</td>
<td>120</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>011</td>
<td>M</td>
<td>.</td>
<td>68</td>
<td>300</td>
<td>20</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>012</td>
<td>M</td>
<td>10/12/98</td>
<td>60</td>
<td>122</td>
<td>74</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>013</td>
<td>2</td>
<td>08/23/99</td>
<td>74</td>
<td>108</td>
<td>64</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>014</td>
<td>M</td>
<td>02/02/99</td>
<td>22</td>
<td>130</td>
<td>90</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>015</td>
<td>F</td>
<td>.</td>
<td>82</td>
<td>148</td>
<td>88</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>017</td>
<td>F</td>
<td>04/05/99</td>
<td>208</td>
<td></td>
<td>84</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>019</td>
<td>M</td>
<td>06/07/99</td>
<td>58</td>
<td>118</td>
<td>70</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>020</td>
<td>F</td>
<td>10</td>
<td>20</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>022</td>
<td>M</td>
<td>10/10/99</td>
<td>48</td>
<td>114</td>
<td>82</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>023</td>
<td>F</td>
<td>12/31/98</td>
<td>22</td>
<td>34</td>
<td>78</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>024</td>
<td>F</td>
<td>11/09/98</td>
<td>76</td>
<td>120</td>
<td>80</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>025</td>
<td>M</td>
<td>01/01/99</td>
<td>74</td>
<td>102</td>
<td>68</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>027</td>
<td>F</td>
<td>.</td>
<td>166</td>
<td>106</td>
<td>7</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>028</td>
<td>F</td>
<td>03/28/98</td>
<td>66</td>
<td>150</td>
<td>90</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>029</td>
<td>M</td>
<td>05/15/98</td>
<td>.</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>M</td>
<td>.</td>
<td>60</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>321</td>
<td>F</td>
<td>.</td>
<td>900</td>
<td>400</td>
<td>200</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>XX5</td>
<td>M</td>
<td>05/07/98</td>
<td>68</td>
<td>120</td>
<td>80</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

The NODUP option of PROC SORT

- NODUP also deletes duplicates, but only for two observations where ALL the variables have identical values
- you have to sort by all of the variables to make this work correctly
- note use of _ALL_ as the by variable in the example for this purpose

Program 5-2 Demonstrating the NODUP option of PROC SORT

PROC SORT DATA=CLEAN.PATIENTS OUT=SINGLE NODUP;
   BY _ALL_;
RUN;

PROC PRINT DATA=SINGLE;
   TITLE 'SINGLE from NODUP';
   ID PATNO;
RUN;
What happens if you use NODUP and sort on only one variable

Program 5-3 Demonstrating a "feature" of the NODUP option

DATA MULTIPLE;
  INPUT PATNO $ X Y;
DATALINES;
  001 1 2
  006 1 2
  009 1 2
  001 3 4
  001 2
  009 1 2
  001 2
  ;
PROC SORT DATA=MULTIPLE OUT=SINGLE NODUP;
  BY PATNO;
RUN;
PROC PRINT DATA=SINGLE;
  TITLE "Listing of Data Set SINGLE";
RUN;

Listing of Data Set SINGLE

Obs  PATNO  X  Y
1   001    1 2
2   006    1 2
3   009    1 2
4   001    3 4
5   001    2
6   009    2 0

Data step method for identifying and listing duplicate IDs

Program 5-4 Identifying duplicate ID's

PROC SORT DATA=CLEAN.PATIENTS OUT=TMP;
  BY PATNO;
RUN;
DATA DUP;
  SET TMP;
  IF FIRST.PATNO AND LAST.PATNO THEN DELETE;
RUN;
PROC PRINT DATA=DUP;
  TITLE "Listing of Duplicates from Data Set CLEAN.PATIENTS";
  ID PATNO;
RUN;

Listing of Duplicates from Data Set CLEAN.PATIENTS

A new dataset for illustration purposes

Program 5-5 Creating the SAS data set, PATIENTS2 (a data set containing multiple visits for each patient)

DATA CLEAN.PATIENT2;
  INFILE "/space/kcowles/172/clean/PATIENT2.TXT" PAD;
  INPUT @1 PATNO $3.
   @4 VISIT MMDDYY10.
   @14 HR 3.
   @17 SBP 3.
   @20 DBP 3.;
  FORMAT VISIT MMDDYY10.;
  RUN;
PROC PRINT DATA=CLEAN.PATIENT2;
  TITLE 'PATIENT2 DATA SET' ;
run ;
Using PROC FREQ to detect duplicate IDs

- can use PROC FREQ to count the number of observations for each value of patient ID variable
- use patient ID variable and OUT= option of TABLES statement to create a SAS data set containing the value of the PATNO variable and the frequency count
  - PROC FREQ uses the variable name COUNT to hold the frequency information
- after you have this information, you can use it to select the original duplicate observations from your data set

Detecting duplicates having same visit date for same patient

/******************************************************************************
 Program 5-6 Identifying patient ID's with duplicate visit dates
*******************************************************************************/
PROC SORT DATA=CLEAN.PATIENT2 OUT=TMP;
  BY PATNO VISIT;
RUN;

DATA DUP;
  SET TMP;
  BY PATNO VISIT;
  IF FIRST.VISIT AND LAST.VISIT THEN DELETE;
RUN;

PROC PRINT DATA=DUP;
  TITLE "Listing of Duplicates from Data Set CLEAN.PATIENT2";
  ID PATNO;
RUN;

Listing of Duplicates from Data Set CLEAN.PATIENT2

08:41 Monday, July 7, 2003

PATNO VISIT HR SBP DBP
005 04/14/1998 72 118 74
005 04/14/1998 74 120 80

/******************************************************************************
 Program 5-7 Using PROC FREQ and an output data set to identify duplicate ID
*******************************************************************************/
PROC FREQ DATA=CLEAN.PATIENTS NOPRINT;
  TABLES PATNO / OUT=DUP_NO(KEEP=PATNO COUNT WHERE=(COUNT GT 1));
RUN;

PROC SORT DATA=CLEAN.PATIENTS OUT=TMP;
  BY PATNO;
RUN;

PROC SORT DATA=DUP_NO;
  BY PATNO;
RUN;

DATA DUP;
  MERGE TMP DUP_NO(IN=YES_DUP DROP=COUNT);
  BY PATNO;
  IF YES_DUP;
RUN;

PROC PRINT DATA=DUP;
  TITLE "Listing of Data Set DUP";
RUN;
Identifying subjects with ”n” observations each using data step

- sometimes we know how many observations there should be for each subject and we need to verify that the correct number are there in data set
- example program lists all patient IDs that do not have exactly two observations each

More efficient program to accomplish similar task

/***************************************************************************/
Program 5-8 Producing a list of duplicate patient numbers, using PROC FREQ
***************************************************************************/
PROC FREQ DATA=CLEAN.PATIENTS NOPRINT;
   TABLES PATNO / OUT=DUP_NO(KEEP=PATNO COUNT)
      WHERE=(COUNT GT 1);
RUN;

DATA _NULL_;
   TITLE "Patients with Duplicate Observations";
   FILE PRINT;
   SET DUP_NO;
   PUT "Patient number " PATNO "has " COUNT "observation(s).";
RUN;

Patients with Duplicate Observations 8
08:41 Monday, July 7, 2003
Patient number 002 has 2 observation(s).
Patient number 003 has 2 observation(s).
Patient number 006 has 2 observation(s).

/***************************************************************************/
Program 5-11 Using a Data Step to list all ID’s for subjects who do not have exactly two observations
***************************************************************************/
PROC SORT DATA=CLEAN.PATIENT2(KEEP=PATNO) OUT=TMP;
   BY PATNO;
RUN;

DATA _NULL_;
   TITLE "Patient ID’s for Patients with Other than Two Observations";
   FILE PRINT;
   SET TMP;
   BY PATNO;
   IF FIRST.PATNO THEN N = 1;
   ELSE N + 1;
   IF LAST.PATNO AND N NE 2 THEN PUT "Patient number " PATNO "has " N "observation(s).";
RUN;

Patient ID’s for Patients with Other than Two Observations 8
08:41 Monday, July 7, 2003
Patient number 002 has 3 observation(s).
Patient number 003 has 1 observation(s).
Patient number 006 has 1 observation(s).
Accomplishing same task using PROC FREQ

• usually easier to have a SAS procedure do the work than to code up a fancy DATA step

***************************************************************************
Program 5-12 Using PROC FREQ to list all ID’s for subjects who do not have exactly two observations
***************************************************************************

PROC FREQ DATA=CLEAN.PATIENT2 NOPRINT;
   TABLES PATNO / OUT=DUP_NO(KEEP=PATNO COUNT
            WHERE=(COUNT NE 2));
RUN;

DATA _NULL_;
   TITLE "Patient ID’s for Patients with Other than Two Observations";
   FILE PRINT;
   SET DUP_NO;
   PUT "Patient number " PATNO " has " COUNT "observation(s).";
RUN;

Patient ID’s for Patients with Other than Two Observations  12
08:41 Monday, July 7, 2003

Patient number 002 has 3 observation(s).
Patient number 003 has 1 observation(s).
Patient number 006 has 1 observation(s).