Solution to Exercise 1: Creating a working dataset

Key Learning Points
   a. You should clean the final dataset so as to remove 'undesirable records' and correct obvious gross errors. Records removed from the dataset should be documented as well as the reason.
   b. The 'cleaned' working dataset will then be used for data analysis.

Task:

Create a combined dataset C_EX01_COMBINE.REC from all 107 files with a program C_EX01_1_COMBINE.PGM.

Solution

This is the dataset by country and year that should result from your program:

Year of registration	Moldova	Mongolia	Uganda	Zimbabwe	Total
		Country			
1999	0	0	17308	0	17308
2000	0	0	18655	0	18655
2001	0	0	18087	1213	19300
2002	0	149	0	29307	29456
2003	17725	22406	0	3958	44089
Total	17725	22555	54050	34478	128808

A possible solution is the following `C_EX01_1_COMBINE.PGM`:

```
* Part C, Exercise 1, first program

* Produce combined dataset for
* Moldova, Mongolia, Uganda, Zimbabwe
* and remove empty records

* Data courtesy:
* Moldowa: Dumitru Laticevschi, OR Paris 2003
* Mongolia:  Nymadawa Naranbat, OR Paris 2004
* Uganda:    Achilles Katamba, OR Paris 2003
* Zimbabwe:    Biggie Mabaera, OR Paris 2004

* Written by:   Hans L Rieder
* First version: 17 Jan 2010
* Last revision: 03 Jun 2013

cls
close
logclose


*********************************************
* Combine original final Moldova datasets
* Create mol_1.rec

cls
logclose
```

```
close

read   "mol_01.rec"
append /file="mol_02.rec"
append /file="mol_03.rec"
append /file="mol_04.rec"
append /file="mol_05.rec"
append /file="mol_06.rec"
append /file="mol_07.rec"
append /file="mol_08.rec"
append /file="mol_09.rec"
append /file="mol_10.rec"
append /file="mol_11.rec"
append /file="mol_12.rec"
append /file="mol_13.rec"
append /file="mol_14.rec"
append /file="mol_15.rec"
append /file="mol_16.rec"
append /file="mol_17.rec"
append /file="mol_18.rec"
append /file="mol_19.rec"
append /file="mol_20.rec"
append /file="mol_21.rec"
append /file="mol_22.rec"
append /file="mol_23.rec"
append /file="mol_24.rec"
append /file="mol_25.rec"
savedata "mol_0.rec" /replace

cls
logclose
close
read   "mol_0.rec"
define country #
country=1
label country "Study country"
* Exlclude laboratory BND with 13 records
* collected during 1 week only
select labcode<>"BND"
* remove 1 empty record
select serno<>.
var drop unique serno
savedata "mol_1.rec" /replace

close
*******************************************
* Combine original final Mongolia datasets
* Create mon_1.rec

cls
logclose
close

read   "mon_01.rec"
append /file="mon_02.rec"
append /file="mon_03.rec"
append /file="mon_04.rec"
append /file="mon_05.rec"
append /file="mon_06.rec"
append /file="mon_07.rec"
append /file="mon_08.rec"
* Note: 1 record in MON_09.REC had a corrupted
* date which prevented appending.  This record
* was manually changed in EpiData from "203" to "2003"
append /file="mon_09.rec"
append /file="mon_10.rec"
append /file="mon_11.rec"
append /file="mon_12.rec"
```

```
append /file="mon_13.rec"
append /file="mon_14.rec"
append /file="mon_15.rec"
append /file="mon_16.rec"
append /file="mon_17.rec"
append /file="mon_18.rec"
append /file="mon_19.rec"
append /file="mon_20.rec"
append /file="mon_21.rec"
append /file="mon_22.rec"
append /file="mon_23.rec"
append /file="mon_24.rec"
append /file="mon_25.rec"
append /file="mon_26.rec"
append /file="mon_27.rec"
append /file="mon_28.rec"
append /file="mon_29.rec"
append /file="mon_30.rec"
append /file="mon_31.rec"
savedata "mon_0.rec" /replace
close

read "mon_0.rec"
define country #
country=2
label country "Study country"
* The following removes 10 empty records
select serno<>.
savedata "mon_1.rec" /replace

close
*********************************************
* Combine original final Uganda datasets
* Create uga_1.rec

cls
close
logclose

read    "uga_01.rec"
append /file="uga_02.rec"
append /file="uga_03.rec"
append /file="uga_04.rec"
append /file="uga_05.rec"
append /file="uga_06.rec"
append /file="uga_07.rec"
append /file="uga_08.rec"
append /file="uga_09.rec"
append /file="uga_10.rec"
append /file="uga_11.rec"
append /file="uga_12.rec"
append /file="uga_13.rec"
append /file="uga_14.rec"
append /file="uga_15.rec"
append /file="uga_16.rec"
append /file="uga_17.rec"
append /file="uga_18.rec"
append /file="uga_19.rec"
append /file="uga_20.rec"
append /file="uga_21.rec"
append /file="uga_22.rec"
append /file="uga_23.rec"
append /file="uga_24.rec"
append /file="uga_25.rec"
append /file="uga_26.rec"
append /file="uga_27.rec"
append /file="uga_28.rec"
append /file="uga_29.rec"
```

```
append /file="uga_30.rec"
savedata "uga_0.rec" /replace

cls
logclose
close
read  "uga_0.rec"
define country #
let country=3
label country "Study country"
define labcode _____
let labcode=labno
var drop labno serno
savedata "uga_1.rec" /replace

close
*********************************************
* Combine original final Zimbabwe datasets
* Create zim_1.rec

cls
logclose
close

read  "zim_01.rec"
append /file="zim_02.rec"
append /file="zim_03.rec"
append /file="zim_04.rec"
append /file="zim_05.rec"
append /file="zim_06.rec"
append /file="zim_07.rec"
append /file="zim_08.rec"
append /file="zim_09.rec"
append /file="zim_10.rec"
append /file="zim_11.rec"
append /file="zim_12.rec"
append /file="zim_13.rec"
append /file="zim_14.rec"
append /file="zim_15.rec"
append /file="zim_16.rec"
append /file="zim_17.rec"
append /file="zim_18.rec"
append /file="zim_19.rec"
append /file="zim_20.rec"
append /file="zim_21.rec"
append /file="zim_22.rec"
append /file="zim_23.rec"
savedata "zim_0.rec" /replace
close

read "zim_0.rec"
define country #
country=4
label country "Study country"
* Note: if you freq on laboratory then
* you have a lab without a code.  When you sort
* on laboratory, then you see it on the top with
* 4 dots.  Curiously, an ID was created nevertheless
* it is laboratory "MW_L"
* Thus, from the following recoding, we get
* an appropriate laboratory and can retain the record
if ID="MW_L-2002-554" then laboratory="MW_L"
* Laboratory coded as "G867" is actually "ML_L"
* Thus, from the following recoding, we get
* an appropriate laboratory and can retain the record
if laboratory="G867" then laboratory="ML_L"
savedata "zim_1.rec" /replace
```

```
close
************************************************
* Combine 4 country sets

cls
close
logclose

cls
read "mon_1.rec"
drop serno id result pattern
savedata "montemp.rec" /replace
close

read "mol_1.rec"
define laboratory ____
laboratory=labcode
define regdate <dd/mm/yyyy>
regdate=labdate
drop labcode labdate
savedata "moltemp.rec" /replace
close

read "uga_1.rec"
define laboratory ____
laboratory=labcode
define regdate <dd/mm/yyyy>
regdate=labdate
drop labcode labdate
savedata "ugatemp.rec" /replace
close

cls
read "zim_1.rec"
drop serno id result pattern
savedata "zimtemp.rec" /replace
close

read "moltemp.rec"
append /file="montemp.rec"
append /file="ugatemp.rec"
append /file="zimtemp.rec"
labelvalue country /1="Moldova"
labelvalue country /2="Mongolia"
labelvalue country /3="Uganda"
labelvalue country /4="Zimbabwe"
savedata "c_ex01_combine.rec" /replace
close

read "c_ex01_combine.rec"
freq country

*************************
* Clean up and erase temporary session files

set echo=off
close
define yesno # global
yesno=?Delete all temporary files: 1=yes 0=no?
imif yesno=1 then
  cls
  type "Be patient, you will be alerted to completion" /h2
  erase "mon_0.chk  "
  erase "mon_0.rec  "
  erase "mon_1.chk  "
  erase "mon_1.rec  "
  erase "mol_0.chk  "
  erase "mol_0.rec  "
```

```
   erase "mol_1.chk  "
   erase "mol_1.rec  "
   erase "uga_0.chk  "
   erase "uga_0.rec  "
   erase "uga_1.chk  "
   erase "uga_1.rec  "
   erase "zim_0.chk  "
   erase "zim_0.rec  "
   erase "zim_1.chk  "
   erase "zim_1.rec  "
   erase "moltemp.chk"
   erase "moltemp.rec"
   erase "montemp.chk"
   erase "montemp.rec"
   erase "ugatemp.chk"
   erase "ugatemp.rec"
   erase "zimtemp.chk"
   erase "zimtemp.rec"
   cls
   type "All temporary files erased" /h2
 else
   type "All temporary files retained" /h2
endif
set echo=on
```

Task:

o *Create a "cleaned" final working dataset C_EX01.REC with a program C_EX01_2_RESTRUCTURE.PGM which excludes non-sensically coded result sequences, and with all fields codes numerically (including COUNTRY and LABORATORY).*

Solution

A possible solution is the following C_EX01_2_RESTRUCTURE.PGM:

```
* Part C, Exercise 1, second program

* Produce cleaned dataset for
* Moldova, Mongolia, Uganda, Zimbabwe
* Removing results with nonsensical sequence

* Data courtesy:
* Moldowa: Dumitru Laticevschi, OR Paris 2003
* Mongolia:  Nymadawa Naranbat, OR Paris 2004
* Uganda:    Achilles Katamba, OR Paris 2003
* Zimbabwe:   Biggie Mabaera, OR Paris 2004

* Written by:   Hans L Rieder
* First version: 17 Jan 2010
* Last revision: 03 Jun 2013

cls
close
logclose

read "c_ex01_combine.rec"

                  define res1b _
          if res1=0 then res1b="N"
if res1>0 and res1<9 then res1b="P"
          if res1=9 then res1b="9"
```

```
                 define res2b _
           if res2=0 then res2b="N"
if res2>0 and res2<9 then res2b="P"
           if res2=9 then res2b="9"

                 define res3b _
           if res3=0 then res3b="N"
if res3>0 and res3<9 then res3b="P"
           if res3=9 then res3b="9"


define sequence _____
label sequence "Sequence of serial results"
let sequence=res1b+"-"+res2b+"-"+res3b


* The following removes records with an impossible
* sequence of results
cls
select sequence<>"9-9-9"
select sequence<>"9-9-N"
select sequence<>"9-9-P"
select sequence<>"9-N-9"
select sequence<>"9-N-N"
select sequence<>"9-P-P"
select sequence<>"N-9-N"
select sequence<>"N-9-P"
select sequence<>"P-9-P"
select sequence<>"9-P-9"
select sequence<>"9-P-N"
select sequence<>"P-9-N"
select sequence<>"9-N-P"


cls
define result1 #.#
label result1 "Result of 1st examination"
let result1=res1
if res1=8.0 then result1=4.0

define result2 #.#
label result2 "Result of 2nd examination"
let result2=res2
if res2=8.0 then result2=4.0

define result3 #.#
label result3 "Result of 3rd examination"
let result3=res3
if res3=8.0 then result3=4.0

cls
define reason0 ##
if reason="D" then reason0=00
if reason="F" then reason0=10
if reason="9" then reason0=99
if reason="1" then reason0=01
if reason="2" then reason0=02
if reason="3" then reason0=03
if reason="4" then reason0=04
if reason="5" then reason0=05
if reason="6" then reason0=06
if reason="7" then reason0=07
if reason="8" then reason0=08


cls
define sex0 #
if sex="F" then sex0=1
if sex="M" then sex0=2
if sex="9" then sex0=9
```

```
cls
define lab0 ###
* Moldova laboratories
if laboratory="ANR" then lab0=101
if laboratory="BLM" then lab0=102
if laboratory="BRL" then lab0=103
if laboratory="BSR" then lab0=104
if laboratory="CCE" then lab0=105
if laboratory="CDR" then lab0=106
if laboratory="CHR" then lab0=107
if laboratory="CLR" then lab0=108
if laboratory="CMN" then lab0=109
if laboratory="CMR" then lab0=110
if laboratory="CNR" then lab0=111
if laboratory="CRR" then lab0=112
if laboratory="CTR" then lab0=113
if laboratory="DNR" then lab0=114
if laboratory="EDR" then lab0=115
if laboratory="FLR" then lab0=116
if laboratory="FRR" then lab0=117
if laboratory="HNR" then lab0=118
if laboratory="LVR" then lab0=119
if laboratory="PRB" then lab0=120
if laboratory="RZR" then lab0=121
if laboratory="SRR" then lab0=122
if laboratory="STR" then lab0=123
if laboratory="VLR" then lab0=124

cls
* Mongolia laboratories
if laboratory="AR_B" then lab0=201
if laboratory="BG_B" then lab0=202
if laboratory="BN_B" then lab0=203
if laboratory="BU_B" then lab0=204
if laboratory="BZ_B" then lab0=205
if laboratory="CH_B" then lab0=206
if laboratory="DA_B" then lab0=207
if laboratory="DD_B" then lab0=208
if laboratory="DG_B" then lab0=209
if laboratory="DU_B" then lab0=210
if laboratory="GA_B" then lab0=211
if laboratory="GS_B" then lab0=212
if laboratory="KE_B" then lab0=213
if laboratory="KH_B" then lab0=214
if laboratory="KO_B" then lab0=215
if laboratory="KU_B" then lab0=216
if laboratory="NA_B" then lab0=217
if laboratory="OR_B" then lab0=218
if laboratory="PR_B" then lab0=219
if laboratory="RE_B" then lab0=220
if laboratory="SB_B" then lab0=221
if laboratory="SK_B" then lab0=222
if laboratory="SU_B" then lab0=223
if laboratory="TU_B" then lab0=224
if laboratory="UM_B" then lab0=225
if laboratory="US_B" then lab0=226
if laboratory="UV_B" then lab0=227
if laboratory="ZA_B" then lab0=228
if laboratory="SE_B" then lab0=229
if laboratory="BK_B" then lab0=230
if laboratory="B-UB" then lab0=231

cls
* Uganda laboratories
if trim(laboratory)="1"  then lab0=301
if trim(laboratory)="2"  then lab0=302
if trim(laboratory)="3"  then lab0=303
if trim(laboratory)="4"  then lab0=304
```

```
if trim(laboratory)="5"  then lab0=305
if trim(laboratory)="6"  then lab0=306
if trim(laboratory)="7"  then lab0=307
if trim(laboratory)="8"  then lab0=308
if trim(laboratory)="9"  then lab0=309
if trim(laboratory)="10" then lab0=310
if trim(laboratory)="11" then lab0=311
if trim(laboratory)="12" then lab0=312
if trim(laboratory)="13" then lab0=313
if trim(laboratory)="14" then lab0=314
if trim(laboratory)="15" then lab0=315
if trim(laboratory)="16" then lab0=316
if trim(laboratory)="17" then lab0=317
if trim(laboratory)="18" then lab0=318
if trim(laboratory)="19" then lab0=319
if trim(laboratory)="20" then lab0=320
if trim(laboratory)="21" then lab0=321
if trim(laboratory)="22" then lab0=322
if trim(laboratory)="23" then lab0=323
if trim(laboratory)="24" then lab0=324
if trim(laboratory)="25" then lab0=325
if trim(laboratory)="26" then lab0=326
if trim(laboratory)="27" then lab0=327
if trim(laboratory)="28" then lab0=328
if trim(laboratory)="29" then lab0=329
if trim(laboratory)="30" then lab0=330

cls
* Zimbabwe laboratories
if laboratory="BY_A" then lab0=401
if laboratory="MC_A" then lab0=402
if laboratory="MC_B" then lab0=403
if laboratory="MC_C" then lab0=404
if laboratory="MC_G" then lab0=405
if laboratory="MC_I" then lab0=406
if laboratory="MC_J" then lab0=407
if laboratory="MD_G" then lab0=408
if laboratory="ME_A" then lab0=409
if laboratory="ME_C" then lab0=410
if laboratory="ME_L" then lab0=411
if laboratory="ME_O" then lab0=412
if laboratory="ML_E" then lab0=413
if laboratory="ML_G" then lab0=414
if laboratory="ML_I" then lab0=415
if laboratory="ML_L" then lab0=416
if laboratory="MN_G" then lab0=417
if laboratory="MV_A" then lab0=418
if laboratory="MV_C" then lab0=419
if laboratory="MV_E" then lab0=420
if laboratory="MW_B" then lab0=421
if laboratory="MW_E" then lab0=422
if laboratory="MW_L" then lab0=423

drop sequence
drop res1 res2 res3
drop reason
drop sex
drop laboratory

rename reason0 to reason
rename sex0 to sex
rename lab0 to laboratory

savedata "temp0.rec" /replace

*********************************
cls
close
```

```
read "temp0.rec"

define regyear0 ####
regyear0=year(regdate)

define regyear ####
regyear=regyear0

* correct errors in year of recording
if regyear0=1990 and laboratory=301 then regyear=1999
if regyear0=1990 and laboratory=306 then regyear=1999
if regyear0=1990 and laboratory=319 then regyear=1999
if regyear0=1990 and laboratory=320 then regyear=2000
if regyear0=1990 and laboratory=410 then regyear=2002

if regyear0=2000 and laboratory=408 then regyear=2002
if regyear0=2000 and laboratory=416 then regyear=2002
if regyear0=2000 and laboratory=419 then regyear=2002

if regyear0=2004 and laboratory=211 then regyear=2003
if regyear0=2004 and laboratory=223 then regyear=2003
if regyear0=2004 and laboratory=401 then regyear=2002
if regyear0=2004 and laboratory=408 then regyear=2002
if regyear0=2004 and laboratory=412 then regyear=2003
if regyear0=2004 and laboratory=413 then regyear=2002

if regyear0=2005 and laboratory=207 then regyear=2003
if regyear0=2033 and laboratory=207 then regyear=2003

label regyear "Year of registration"
labelvalue sex /1="Female" /2="Male" /9="Missing"
label sex "Sex of examinee"
labelvalue reason /0="Diagnosis"
labelvalue reason /1="Follow-up at 1 month"
labelvalue reason /2="Follow-up at 2 months"
labelvalue reason /3="Follow-up at 3 months"
labelvalue reason /4="Follow-up at 4 months"
labelvalue reason /5="Follow-up at 5 months"
labelvalue reason /6="Follow-up at 6 months"
labelvalue reason /7="Follow-up at 7 months"
labelvalue reason /8="Follow-up at 8 months or later"
labelvalue reason /10="Follow-up, month not stated"
labelvalue reason /99="Reason not stated"
label reason "Reason for examination"

labelvalue result1 /0.0="Negative"
labelvalue result1 /4.0="Positive, not quantified"
labelvalue result1 /5.0="Scanty, not quantified"
labelvalue result1 /0.1="Scanty, 1 AFB per 100 fields"
labelvalue result1 /0.2="Scanty, 2 AFB per 100 fields"
labelvalue result1 /0.3="Scanty, 3 AFB per 100 fields"
labelvalue result1 /0.4="Scanty, 4 AFB per 100 fields"
labelvalue result1 /0.5="Scanty, 5 AFB per 100 fields"
labelvalue result1 /0.6="Scanty, 6 AFB per 100 fields"
labelvalue result1 /0.7="Scanty, 7 AFB per 100 fields"
labelvalue result1 /0.8="Scanty, 8 AFB per 100 fields"
labelvalue result1 /0.9="Scanty, 9 AFB per 100 fields"
labelvalue result1 /1.0="1+ positive"
labelvalue result1 /2.0="2+ positive"
labelvalue result1 /3.0="3+ positive"
labelvalue result1 /9.0="No result recorded"
label result1 "Result of 1st examination"

labelvalue result2 /0.0="Negative"
labelvalue result2 /4.0="Positive, not quantified"
labelvalue result2 /5.0="Scanty, not quantified"
labelvalue result2 /0.1="Scanty, 1 AFB per 100 fields"
```

```
labelvalue result2 /0.2="Scanty, 2 AFB per 100 fields"
labelvalue result2 /0.3="Scanty, 3 AFB per 100 fields"
labelvalue result2 /0.4="Scanty, 4 AFB per 100 fields"
labelvalue result2 /0.5="Scanty, 5 AFB per 100 fields"
labelvalue result2 /0.6="Scanty, 6 AFB per 100 fields"
labelvalue result2 /0.7="Scanty, 7 AFB per 100 fields"
labelvalue result2 /0.8="Scanty, 8 AFB per 100 fields"
labelvalue result2 /0.9="Scanty, 9 AFB per 100 fields"
labelvalue result2 /1.0="1+ positive"
labelvalue result2 /2.0="2+ positive"
labelvalue result2 /3.0="3+ positive"
labelvalue result2 /9.0="No result recorded"
label result2 "Result of 2nd examination"

labelvalue result3 /0.0="Negative"
labelvalue result3 /4.0="Positive, not quantified"
labelvalue result3 /5.0="Scanty, not quantified"
labelvalue result3 /0.1="Scanty, 1 AFB per 100 fields"
labelvalue result3 /0.2="Scanty, 2 AFB per 100 fields"
labelvalue result3 /0.3="Scanty, 3 AFB per 100 fields"
labelvalue result3 /0.4="Scanty, 4 AFB per 100 fields"
labelvalue result3 /0.5="Scanty, 5 AFB per 100 fields"
labelvalue result3 /0.6="Scanty, 6 AFB per 100 fields"
labelvalue result3 /0.7="Scanty, 7 AFB per 100 fields"
labelvalue result3 /0.8="Scanty, 8 AFB per 100 fields"
labelvalue result3 /0.9="Scanty, 9 AFB per 100 fields"
labelvalue result3 /1.0="1+ positive"
labelvalue result3 /2.0="2+ positive"
labelvalue result3 /3.0="3+ positive"
labelvalue result3 /9.0="No result recorded"
label result3 "Result of 3rd examination"
label regdate "Date of registration"
label laboratory "Laboratory code"
keep country laboratory regdate regyear age sex reason result1 result2 result3
savedata "c_ex01.rec" /replace

********************************
* Test labels, sorting, and count

cls
close
read "c_ex01.rec"

tables country result1 /SLA /vl
tables country regyear

********************************
* Clean up and erase temporary session files

set echo=off
close
define yesno # global
yesno=?Delete all temporary files: 1=yes 0=no?
imif yesno=1 then
  cls
  erase "temp0.rec"
  erase "temp0.chk"
  cls
  type "All temporary files erased" /h2
 else
  type "All temporary files retained" /h2
endif
set echo=on
```