PanelWhiz Plugins: Automatic Vector-Oriented Data Cleaning for Large Scale Panel Datasets

Markus Hahn

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI Essen)

5th German Stata Users Group meeting
(April 2, 2007)
"Employed Again Last Year" (Source: GSOEP)

<table>
<thead>
<tr>
<th>Year 2004</th>
<th>Year 2005</th>
</tr>
</thead>
</table>
### "Number Employees in Firm" (Source: GSOEP)

<table>
<thead>
<tr>
<th>Year 2004</th>
<th>Year 2005</th>
</tr>
</thead>
</table>
What a plugin does

- Ensures time consistency of variables
- Data cleaning / Conversion (DM⇒EURO, ...)
- Creating new variables (real values, ...)

How a plugin works

- Executable ado-files for each item
- Plugin only describes what actions and when these actions are to be performed
- Plugin system executes the necessary actions
- Auto-documentation of what is done
Example Plugin 1/2

program define bp_lglohn

*! Plugin : [lglohn] Brutto Lohn- und Gehaltssumme
*! Author : Markus Hahn
*! Email : hahn@rwi-essen.de
*! Web : http://www.panelwhiz.eu
*! Copyright : Copyright 2006. All rights reserved. Unauthorized copying prohibited.
*! Version : 1.0
*! Years : a b c d e f g h i j k l
*! Possible : a b c d e f g h i j k l
*! Changes : a b c d e f g h i j k l
*! *-----------------------------------------------------------------------

pwplugin begin, item(lglohn) year(2004)
   .action.dm2euro

      .plugin.newitem
      .action.dm2euro
      .action.real, thisyear

pwplugin end

end
Example Plugin 2/2

program define bp_aawest

*! --------------------------------------------------------------------------------------
*! Plugin : [aawest] AA: West-/Ostdeutschland
*! Author : Markus Hahn
*! Email : hahn@rwi-essen.de
*! Web : http://www.panelwhiz.eu
*! Copyright : Copyright 2006. All rights reserved. Unauthorized copying prohibited.
*! Version : 1.0
*! Years : a b c d e f g h i j k l
*! Possible : _ _ _ d e f g h i j k l
*! Changes : _ _ _ _ _ g h i j k l
*! --------------------------------------------------------------------------------------

pwplugin begin, item(aawest) year(2004)
    .action.recode (2=0), g h i j k l
pwplugin end

drive
Automatically generated plugin-do-file (Auto-doc)

PanelWhiz

Homepage: http://www.panelwhiz.eu
Help: http://forum.panelwhiz.eu
Contact: Dr. John P. Haisken-DeNew

+++ ---------------------------------------------------------------------
++ [aawest] AA: West-/Ostdeutschland
++ ---------------------------------------------------------------------
// ---[ PREPARATION OF ITEMS ]------------------------------------------
// label variable aawest "West-/Ostdeutschland"
// ---[ BEFORE ]--------------------------------------------------------
// ---[ A - 1993 ]-----------------------------------------------------
// ---[ B - 1994 ]-----------------------------------------------------
// ---[ C - 1995 ]-----------------------------------------------------
// ---[ D - 1996 ]-----------------------------------------------------
// ---[ E - 1997 ]-----------------------------------------------------
// ---[ F - 1998 ]-----------------------------------------------------
// ---[ G - 1999 ]                                                     
// recode aawest (2=0) if year == 1999
// ---[ H - 2000 ]                                                     
// recode aawest (2=0) if year == 2000
// ---[ I - 2001 ]                                                     
// recode aawest (2=0) if year == 2001
// ---[ J - 2002 ]                                                     
// recode aawest (2=0) if year == 2002
// ---[ K - 2003 ]                                                     
// recode aawest (2=0) if year == 2003
// ---[ L - 2004 ]                                                     
// recode aawest (2=0) if year == 2004
// ---[ AFTER ]---------------------------------------------------------
Why we do not use normal ado-files

- Relative high complexity of code (loops, ...)
  ⇒ Not that easy to understand
- The code should be self-explanatory (transparency)
- Simpler code could is easier to process
  ⇒ for generation purposes ("plugin wizard")
  ⇒ for auto-documentation purposes beyond log-files
Implementation of the plugin system

- Stata and Mata code
- Class system (*.class)
- Dialog boxes (*.dlg)
- On-the-fly generation of smcl-files (Output for Plugin Wizard)
Recoding a variable

```
.action.recode (2=0), a b c d e
```
- Similar syntax to stata’s recode
- Item variable gets recoded: 2 becomes 0
- Action is done for the waves a-e

Changing a variable

```
.action.replace wage/12, a b c d e
```
- Similar syntax to stata’s replace
- Item variable gets divided by 12
- Action is done for the waves a-e
Handling money variables

.action.dm2euro
- Item variable will be converted to EURO (years < 2002)

.action.real, thisyear
- Item variable will be converted to real values
- The deflator (price index) of the current year is used
Label definition

- .action.label var(Variable Label)
  - Item variable gets new variable label

- .action.label val(1 "Yes" 0 "No")
  - Item variable gets new value label
Creating new variables

plugin.newitem

- A new variable will be created
- Variable name will be automatically assigned
- Following actions(commands) will refer to new variable
Plugin Wizard

- Tool for creating plugins by "clicking"
  ⇒ No manual typing necessary
- User interface: generated smcl-files + dialog boxes
- Supports the complete syntax of plugin code
- Once created plugins can be re-read
Data Browse Page: Start Plugin Wizard
Plugin Wizard
Create, view, erase, open, ...
Information about plugin

PanelWhiz IABB - Plugin Wizard
PanelWhiz IABB 1.0 Aug 2006 <john@panelwhiz.eu>

bp_aabula.ado - (p)
Based on: NEW Plugin
[Create] [View] [Erase] [New] [Open (Original)] [Open (Personal)]

Plugin Information
[Default] [Edit]
- Plugin: [aabula] Bundesland
- Author: Markus Hahn
- Email: hahn@wi-essen.de
- Web: http://www.panelwhiz.eu
- Copyright: Copyright 2006. All rights reserved. Unauthorized copying prohibited.
- Version: 1.0
- Years: a b c d e f g h i j k l
- Possible: a b c d e f g h i j k l
- Changes: = = = = = = = = = =

Plugin Initialization
[Default] [Edit]
- Update year: 2004
- Depends on: <nothing>

Plugin Action List
[Clear List]
[+][-][up][down]
Action List
Add an action to the list
Select Plugin Action: `dm2euro`

Converting DM to EURO
This action is for converting an amount of money from Deutsch Mark (DM) to EURO (€). The conversion is done automatically for the appropriate waves.

Plugin Action Manager
Select an action

Select Plugin Action: dm2euro

Converting DM to EURO
This action is for converting an amount of money from Deutsch Mark (DM) to EURO (€). The conversion is done automatically for the appropriate waves.
Select an action
Description of action

Converting DM to EURO
This action is for converting an amount of money from Deutsch Mark (DM) to EURO (€). The conversion is done automatically for the appropriate waves.
Action "recode"
Action "replace"
Two actions were added
Create plugin
View plugin
Created plugin code
When a plugin is finally created...

- Can be used directly with PanelWhiz retrievals
- Can be shared with the PanelWhiz community
  - Embedded knowledge, which is respected by contract ⇒ PanelWhiz Plugins must be cited
  - Modularized piece of code ⇒ Shareable
Plans for future releases

- Add more predefined actions
- More flexibility through "free coding"
- Extend the auto-documentation feature