Artur Pokropek
Institute of Educational Research (IBE), Warsaw
artur.pokropek@gmail.com

Maciej Jakubowski
Faculty of Economic Sciences, Warsaw University
mjakubowski@uw.edu.pl

PIAACTOOLS: Stata® programs for statistical computing using PIAAC data

Contents

INTRODUCTION

The OECD *Programme for the International Assessment of Adult Competencies* (PIAAC) is a study based on a complex survey design and advanced statistical methods ensuring the highest quality and reliability of final results. To analyse microdata from the PIAAC study, one needs to use special methods developed to obtain correct results. Typically, these methods cannot be used directly in available statistical packages. It is necessary to code statistical commands that facilitate analysis of this data. Three commands presented below, *piaacdes*, *piaacreg* and *piaactab* were developed in Stata® programming language to obtain correct estimates of basic statistics and facilitate regression analysis with PIAAC. These commands allow analyzing plausible values available in PIAAC datasets and account for complex derivation of standard errors using the jackknife method implemented in PIAAC.

The two commands are straightforward to use even for beginning users of Stata® and guarantee that users will obtain correct point estimates and standard errors. The results obtained with these three Stata® commands are virtually identical to those researchers can obtain using IDB analyzer provided by the PIAAC consortium for SPSS® users. The Stata® command *piaacdes* allows calculating basic statistics like mean, median, percentiles, standard deviation etc., *piaacreg* allows using several regression models, while *piaactab* computes frequencies of students at each of proficiency level. Output is saved as HTML tables that can be easily opened and edited in Excel® or any spreadsheet program or internet browser.

STATISTICAL BACKGROUND

In PIAAC complex sample designs were used. This implies two consequences. First, all point estimates must be computed using sampling weights. Second, it is necessary to use special procedures for standard error computations. While some analytical procedures are available for standard error computations in PIAAC, study methods based on replication approach were chosen as they showed to be unbiased and efficient and relatively easy to implement for wide range of applications. For PIAAC research so called *jackknife* replicate procedure was chosen (for details see Efron 1982; Levy and Lemeshow, 1999).

Computation of standard errors not involving cognitive components (not involving *plausible values* methodology) is based on computations that summarize variability of estimates in subsequent subset of samples called replicates:

$$SE_\theta = \sqrt{f \sum_{r=1}^{R} \left(\hat{\theta}_r - \hat{\theta}_0 \right)^2} \qquad (1)$$

Where:

R is the number of replicates;

$\hat{\theta}_r$ represents any statistic of interest (percent, mean, variance, regression coefficient, etc.) not involving *plausible values* for replicate r=(1,...,R);

$\hat{\theta}_0$ represents the statistic of interest (not involving *plausible values*) estimated using the whole sample and final sample weight.

Constant f depends of sampling procedures used in each country. In PIAAC survey two types of sampling were used: non-stratified and stratified for which the constant has different values $f = \frac{R-1}{R}$ or $f = 1$.

When statistics of interests involve cognitive components and *plausible values* methodology (see Wu 2005 for details) computations get more complex. In this case measurement error for cognitive data must be also taken into account. Standard error for a statistic involving plausible values methodology might be expressed as:

$$SE_{\theta_P} = \sqrt{\left(\text{Sampling error}\right)^2 + \left(\text{Measurment error}\right)^2} \tag{2}$$

Implementation of equation (2) for PIAAC data is presented in equation (3):

$$SE_{\theta_P} = \sqrt{\left[\sum_{p=1}^{P}\left(f\sum_{r=1}^{R}\left(\hat{\theta}_{r,p} - \bar{\theta}_{0,P}\right)^2\right)\frac{1}{P}\right] + \left[\left(1+\frac{1}{P}\right)\frac{\sum_{p=1}^{P}\left(\hat{\theta}_{0,p} - \bar{\theta}_{0,P}\right)^2}{P-1}\right]} \tag{3}$$

Where:

$\bar{\theta}_{0,P} = \frac{\sum_{p=1}^{P}\theta_{0,p}}{P}$;

P is the number of plausible values, *p*=(1,...,*P*);

$\hat{\theta}_{r,p}$ represents the statistic estimate for replicate r and the p[th] *plausible value*;

$\hat{\theta}_{0,p}$ represents the statistic estimate using the final sample weight for the p[th] *plausible value*;

$\bar{\theta}_{0,P}$ represents the unweighted average of the statistic for each *plausible value* using whole sample and the final weight

As in PIAAC data we find 10 *plausible values* and 80 replicates for each country, implementing this formula requires 810 repetitions of computations of statistic of interests (800 statistics for each of 80 replicates for each of 10 *plausible vales* and 10 statistics using whole sample and final weights).

For the logistic regression when option "or" is specified and odds ratios are reported, the standard errors are derived using the delta rule. See http://www.stata.com/support/faqs/statistics/delta-rule/.

USING COMMANDS

The package of three commands has six files (three ado files and three help files):

- piaacdes.ado and piaacdes.hlp

- piaacreg.ado and piaacreg.hlp

- piaactab.ado and piaactab.hlp

Before using the commands you need to save these files in Stata® folder with user-written commands. In Stata® type *„sysdir"* or *„personal"* to see where folder called PERSONAL is located and copy these four files to this folder. After re-starting Stata® *piaacdes*, *piaacreg* and *piaactab* will be already available for your use. Please note that you have to repeat this for any new installation of Stata®.

The final versions of PIAAC macro for Stata® will be published on Statistical Software Components (SSC) archive, which is often called the Boston College Archive and is provided by http://www.repec.org. The SSC has become the premier Stata® download site for user-written software on the web. To install PIAAC command from this archive user will need to type:

```
ssc install piaactools
```

Thereafter, the commands *piaacdes*, *piaacreg* or *piaactab* can be invoked directly from Stata®.

FUNCTIONS THAT CAN BE USED WITH PIAACTOOLS

Piaacdes command facilitates calculating basic statistics with PIAAC data. This includes calculating mean, percentage, percentiles, standard deviation or variance. *Piaacreg* helps users to use PIAAC data to estimate regression models including linear regression, logistic regression, poisson regression or ordered logistic regression. *Piaactab* computes frequencies of students at each of proficiency level

All commands compute correct results for standard variables as well as for plausible values, while calculations with plausible values will take much longer. In regression models, plausible values can be used as dependent (with option *pvdep()*) or independent (with option *pvindep1(), pvindep2(), pvindep3()*) variables, even simultaneously.

In all cases, standard errors are calculated using the *jackknife* resampling method which might be time consuming, especially with regression models for which numerical solutions have to be found, e.g. logistic regression. To see preliminary results from *piaacreg*, users can specify *fast* option which speeds up calculations but provides analytical standard errors that are usually incorrect. Final results should always be calculated without *fast* option specified.

Commands can use standard variables available in PIAAC to identify countries and replication method or users can provide names of their own variables. If nothing is specified in *countryid()* option then the commands will look for cntryid or CNTRYID variable in the dataset and will use it to calculate results over countries defined by this variable. Similarly, if *vemethodn()* option is not specified, all commands will

look for variable vemethodn or VEMETHODN that is available in original PIAAC dataset to recognize which jackknife replication method should be used for each country. Weights are also pre-specified - if *weight()* option for final sample weight and *rep()* option for replication weights are not specified commands automatically search for variables spfwt0-spfwt80 or SPFWT0-SPFWT80. Thus, if users are using the original PIAAC dataset they do not have to specify these options. However, if users plan to calculate results for different groups of countries, economies or regions then *countryid()* option can be used. Similarly, different variable denoting jackknife methods can be provided through *vemethodn()* option and weights through options *weight()* and *rep()*.

Commands allow calculations for subgroups and "if" or "in" restrictions in Stata®. While "if" and "in" can be specified as for any other Stata® command, calculations for subgroups are performed using *over()* option. Variable specified in *over()* option must be numeric with a sequence of numbers. For example, if female is 0/1 variable defining male and female respondents, using *over(female)* will produce results by gender.

For *piaacreg* command additional options are available. Option *cons* will add estimates for regression constant in output tables. Option *r2()* can be used to return different summary statistic from the default r-squared. Option *cmd()* allows using different estimation command from the default *regress* command. Option *cmdops()* will pass options to this command.

The commands return results as HTML tables that are saved to disk. These files can easily be opened and further edited in Excel, while they are ready-to-use tables similar to those presented in OECD publications. Users can specify the number of decimal places using *round()* option. Users have to provide the name of the results file using *save()* option (the output folder or file path can be also specified using this option).

For more advanced users, results are returned in matrices that can be further processed directly in Stata®. Type *return list* after running *piaacdes*, *piaacreg* or *piaactab* to see returned matrices with results.

In next section short notes for quick use of Stata® for beginner users are provided. Then full syntax three commands are presented (equivalent with Stata® help files for those commands) followed by extended list of examples of applications of three commands.

NOTES FOR QUICK USE OF STATA®

Open Stata® dataset (*.dta) file in Stata®

Change the directory where macro is located by following command:

adopath + "C:\Documents and Settings\user\ \My Documents\ \Stata®Tool"

Change the working directory to locate output quickly:

cd "C:\Documents and Settings\ user\ \My Documents\ \Stata®Tool "

Since Stata® is case sensitive run the following command to make all variable names in upper case letters:

rename *, upper

Run examples from the list below in the command window (or use do-file), e.g.,

```
piaacdes AGE_R , stats(mean sd) centile(50) save(test1) round(5)
```

FULL SYNTAX FOR PIAACDES (HELP FILE)

Syntax

piaacdes varlist [if] [in], save(string) [, options]

Options Description

Main options

save(filename, ...) save results to filename. You have to specify this option

Optional

countryid(varname) Allows providing the name of a variable containing a list of countries for which you want to obtain results. The list can be numeric or string with any possible values. Missing categories will be omitted from calculations and average over all countries will be calculated. When this option is not specified, CNTRYID or cntryid variable will be used as to identify countries and the OECD average will be calculated.

vemethodn(varname numeric)

Provides the name of the numeric variable specifying the jackknife variance estimation method for each country. The variable can only contain values of 1 or 2. When this option is not specified, VEMETHODN or vemethodn variable will be used to identify jackknife method.

weight(varlist max=1) rep(varlist)

Gives the main weight and a list of jackknife replication weights. You don't have to specify these options if your dataset contains original weights spfwt0-spfwt80 or SPFWT0-SPFWT80.

stats(string) Gives a list of statistics to be calculated. You can list any statistic calculated by -summarize-, e.g. stats(mean sd) will calculate mean and standard deviation. If stats() and centile() are not specified, means will be calculated.

centile(string) Gives percentiles to be calculated. For example, centile(5 50 75) will result in calculating the 5th, median and the 75th percentile.

pv(string)	Gives a list of plausible values prefixes, e.g. use pv(pvlit) to calculate statistics for plausible value in literacy.
over(varname)	Specifies a categorical variable or plausible value for which you want to obtain regression results by each category. The variable must be numerical with a sequence of integers denoting each category. For proficiency levels, specify the prefix of plausible values without ending numbers (e.g. pvlit, pvnum, pvpsl).
round(int)	Specifies how many decimal places you want to see in results tables. The default is 2.

Description

piaacdes calculates basic statistics with PIAAC data for the given variables and plausible values listed in the pv() option. The pv() option takes the prefix of plausible values variable without ending numbers (e.g. pvlit, pvnum, pvpsl). Similarly, the prefix of plausible values can be specified in over() option. In this case, the statistics will be calculated over proficiency levels. Piaacdes saves results as html file that can be, for example, further edited in a spreadsheet application. Piaacdes also returns results in Stata matrices. Type -return list- after executing the command.

FULL SYNTAX FOR PIAACREG (HELP FILE)

Syntax

piaacreg varlist [if] [in], save(string)) [options]

options	**Description**

Main options

save(filename, ...) save results to filename. You have to specify this option

Optional

countryid(varname) Provides the name of a variable containing a list of countries for which you want to obtain results. The list can be numeric or string with any possible values. Missing categories will be omitted from calculations. When this option is not specified, CNTRYID or cntryid variable will be used to identify countries.

vemethodn(varname numeric)
Provides the name of a numeric variable specifying jackknife variance estimation method for each country. The variable can only contain values of 1 or 2. When this option is not specified, VEMETHODN or vemethodn variable will be used to identify the jackknife method.

weight(varlist max=1) rep(varlist)
Gives the main weight and a list of jackknife replication weights. You don't have to specify these options if your dataset contains original weights spfwt0-spfwt80 or SPFWT0-SPFWT80.

pvdep(pv prefix) Specifies plausible values as dependent variables. Provide the plausible values prefix without ending numbers, e.g. pv(pvlit) asks for plausible values pvlit1-pvlit10 to be used as dependent variables.

pvindep1(pv prefix) pvindep1(pv prefix) pvindep3(pv prefix)
Specifies plausible values to be used as independent variables. Provide the plausbile values prefix without ending numbers, e.g. pvindep1(pvlit) asks for plausible values pvlit1-pvlit10 to be used as independent variables. If additional plausible values need to be used one could specify option pvindep2() or pvindep3().

over(varname) Specifies a categorical variable or plausible value for which you want to obtain regression results by each category. The variable must be numerical with a sequence of integers denoting each category. For

	proficiency levels, specify the prefix of plausible values without ending numbers (e.g. pvlit, pvnum, pvpsl).
round(int)	Specifies how many decimal places you want to see in results tables. The default is 2.
fast	Specifying this option speeds up calculations at the cost of not fully valid estimates of standard errors. Point estimates are correct, while standard errors are obtained analytically and usually differ from those obtained with jackknife method.
cons	Specify this option if you want to save estimates for the regression constant.
cmd() cmdops()	Specify these options if you want to run a regression model different from -regress-. You can pass options to the regression command using cmdops(). For example, specifying cmd("logit") and will estimate logistic regression and report odds ratios.
or	Specify this option together with cmd("logit") or cmd("logistic") to obtain odds ratios instead of coefficients. In this case, a standard Stata approach is taken and the standard errors are derived using the delta rule. (see www.stata.com/support/faqs/statistics/delta-rule/)
r2()	Specify r2(r2_a) to report adjusted R-square or any other scalar returned in e().

Description

piaacreg runs regression with PIAAC data. First variable listed after piaacreg command is the dependent variable unless you specify pvdep(). If your dependent variable is a vector of plausible values, you should specify pvdep() option providing the prefix of plausible values variable without ending numbers (e.g. pvlit, pvnum, pvpsl). The remaining variables listed after piaacreg are treated as independent variables. Options pvindep1(), pvindep2() and pvindep3() allow for the use of plausible variables as independent variables. Similarly, the prefix of plausible values can be specified in over() option. In this case, the regressions will be run over proficiency levels. Piaacreg saves results as html file that can be, for example, further edited in a spreadsheet application. Piaacreg also returns results in Stata matrices. Type -return list- after executing the command.

FULL SYNTAX FOR PIAACTAB (HELP FILE)

Syntax

piaactab varname [if] [in], save(string) [, options]

Options Description

Main options

save(filename, ...) save results to filename. You have to specify this option

Optional

twoway Calculates cell percentages as in a twoway table (over both the main and the over() variable).

fast Speeds up calculations by not reporting standard errors.

countryid(varname) Allows for providing the name of a variable containing a list of countries for which you want to obtain results. The list can be numeric or string with any possible values. Missing categories will be omitted from calculations and average over all countries will be calculated. When this option is not specified, CNTRYID or cntryid variable will be used to identify countries and OECD average will be calculated.

vemethodn(varname numeric)

Provides the name of the numeric variable specifying jackknife variance estimation method for each country. The variable can only contain values of 1 or 2. When this option is not specified, VEMETHODN or vemethodn variable will be used to identify jackknife method.

weight(varlist max=1) rep(varlist)

Gives the main weight and a list of jackknife replication weights. You don't have to specify these options if your dataset contains original weights spfwt0-spfwt80 or SPFWT0-SPFWT80.

over(varname)	Specifies a categorical variable or plausible value for which you want to obtain regression results by each category. The variable must be numerical with a sequence of integers denoting each category. For proficiency levels, specify the prefix of plausible values without ending numbers (e.g. pvlit, pvnum, pvpsl).
round(int)	Specifies how many decimal places you want to see in results tables. Default is 2.

Description

piaactab calculates cell percentages for a variable with PIAAC data. To calculate percentages for proficiency levels, specify the prefix of plausible values variable without ending numbers (e.g. pvlit, pvnum, pvpsl). Similarly, the prefix of plausible values can be specified in over() option. In this case, the percentages will be calculated over proficiency levels. Piaactab saves results as html file that can be, for example, further edited in a spreadsheet application. Piaactab also returns results in Stata matrices. Type –return list- after executing the command.

EXAMPLES OF SYNTAX FOR DIFFERENT FUNCTIONS

piaacdes

Computation of mean, standard deviation and median of age (AGE_R) across countries. Results will be saved in file test1(in working directory):

```
piaacdes AGE_R ,stats(mean sd) centile(50)save(test)
```

The same as above but results are presented with 5 decimal points precision:

```
piaacdes AGE_R ,stats(mean sd) centile(50)save(test) round(5)
```

Computation of mean and chosen centiles of PVPSL across countries:

```
piaacdes , pv(PVPSL) stats(mean) centile(5 10 25 50 75 90 95) save(test)
```

Computation of mean and chosen centiles of PVPSL across groups defined by user (in this case groups are identified by variable CNTRY):

```
piaacdes , pv(PVPSL) stats(mean) centile(5 10 25 50 75 90 95) save(test)
countryid(CNTRY)
```

The same computation as above but also over gender (GENDER_R):

```
piaacdes , pv(PVPSL) stats(mean) centile(5 10 25 50 75 90 95)
weight(SPFWT0) save(test) countryid(CNTRY) over(GENDER_R)
```

Computation of statistics for gender (GENDER_R) over ability levels. Ability thresholds for each cognitive domain are stored by program so there is no need to specify thresholds manually:

```
piaacdes GENDER_R, stats(mean sd)  save(test) over(PVPSL)
```

piaacreg

Regression without plausible values AGE_R = β_0 + β GENDER_R + e:

```
piaacreg AGE_R GENDER_R, save(test)
```

Regression with plausible values as a dependent variable: PVPLS=β_0 + βAGE_R + βGENDER_R + e, by countries:

```
piaacreg AGE_R GENDER_R, pvdep(PVLIT) save(test)
```

The same as above but including information about constant and R2 in output file:

```
piaacreg AGE_R GENDER_R, pvdep(PVLIT) save(test) r2(r2) cons
```

Regression with plausible values a dependent variable PVPLS=β_0 + βAGE_R + e , by countries and over gender (GENDER_R):

```
piaacreg AGE_R, pvdep(PVPLS) save(test) over(GENDER_R)
```

Regression where plausible values are specified both as dependent and independent variables PVPLS=β_0 + βAGE_R + βPVNUM+ e:

```
piaacreg AGE_R, pvdep(PVPLS) pvindep1(PVNUM) save(test)
```

Regression where two plausible values variables are specified as independent variables AGE_R=β_0 + βPVPLS + βPVNUM+ e:

```
piaacreg AGE_R, pvindep1(PVPLS) pvindep2(PVNUM) save(test)
```

Running other types of regressions than simple linear OLS requires using option `cmd("")` and `cmdops("")`. The first one specifies the regression command. You could use for instance: `logit`, `probit`, `poisson` etc. (essentially all regression models provided by Stata®).

Logistic regression with plausible values as independent variables logit(GENDER_R)=β_0 + βAGE_R + e

```
piaacreg GENDER_R AGE_R, cmd("logit") pvindep1(pvpsl) save(test) cons
```

The same as above but odds ratio are reported in a html file:

```
piaacreg GENDER_R AGE_R, cmd("logit") pvindep1(pvpsl) save(test)cons or
```

Option `cmdops("")` will pass options to any regression command in Stata®. For instance option `cmdops("noconstant")` might be used to estimate regression model without constant PVPLS= βAGE_R + βGENDER_R + e, by countries:

```
piaacreg AGE_R GENDER_R, pvdep(PVLIT) save(test) cmdops("noconstant")
```

piaactab

Computation of percentage of males and females by country with appropriate standard errors:

```
piaactab GENDER_R, save(test)
```

Computation of percentage of respondents at each of ability level for plausible values in literacy (PVLIT*) with appropriate standard errors using plausible values methodology (thresholds are stored by program):

```
piaactab pvlit, save(test)
```

The same as above but over GENDER_R

```
piaactab pvlit, over(GENDER_R) save(test)
```

LITERATURE

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia Pennsylvania: SIAM

Levy, P. S. and Lemeshow. S. (1999), *Sampling of Populations: Methods and Applications*, 3rd edition, Wiley, New York.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2), 114-128.