intsvy: An R Package for Analysing International Large-Scale Assessment Data

Daniel H. Caro
University of Oxford

Przemyslaw Biecek
University of Warsaw

Abstract

This paper introduces **intsvy**, an R package for working with international assessment data from PISA (Programme for International Student Assessment), TIMSS (The Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study), and PIAAC (Programme for the International Assessment of Adult Competencies). The package includes functions for importing data, performing data analysis, and visualising results. The paper describes the underlying methodology and provides real data examples. Tools for importing data allow useRs to select variables from student, home, school, and teacher survey instruments as well as for specific countries. Data analysis functions take into account the complex sample design (with replicate weights) and rotated test forms (with plausible achievement values) in the calculation of point estimates and standard errors of means, standard deviations, regression coefficients, correlation coefficients, and frequency tables. Visualization tools present data aggregates in standardised graphical form.

Keywords: international assessments, complex survey analysis, replicate weights, plausible values.

1. Introduction

International large-scale assessments (LSA) studies measure student performance through standardised achievement tests and administer questionnaires to collect data on students, their families, and schools that shed light on the mechanisms responsible for student performance in a number of countries. The results used by researchers and policymakers around the world have had significant impact on educational policy and on the educational debate. The Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS) stand out for their impact, comparative trend data, and number of participating countries. The Programme for the International Assessment of Adult Competencies (PIAAC) is another international LSA study with focus on adults' skills that has received great attention from policymakers since its release in 2013. The data from PISA, TIMSS, PIRLS, and PIAAC are publicly available, but its use is somewhat limited by available analytical tools for handling the complex design of LSA studies.

The design of international LSA studies involves complex sampling and testing procedures that have consequences on the analysis stage. Sampling is conducted in two stages, schools are selected in the first stage and students in the second stage. Testing uses a rotated design consisting of different test versions comparable through a common core of items. Datasets contain sampling variables (e.g., replicate weights) and plausible values of achievement scores in order to account for the complex sampling and test design, respectively. Traditional statistical procedures cannot handle these design complexities. Further, the organisation of public datasets from TIMSS and PIRLS in a large number of files by country and survey instrument is not straightforward for users and requires commercial software alternatives (e.g., IDB Analyzer in combination with SPSS) in order to merge and select data. Package **intsvy** facilitates the analysis of PISA, TIMSS, PIRLS and PIAAC by providing tools for importing data and conducting analysis whilst soundly considering the sample and test design in the calculation of statistics and associated standard errors. **intsvy** is an acronym for international surveys.

2. Complex design of international LSA

Obtaining point estimates of any statistic of interest θ (e.g., mean, correlation, percentage, regression coefficient) is not particularly complicated with international assessment data. Standard procedures weighted by the total sampling weight can be used to calculate θ for the observed data. For student performance, the average of plausible values estimates yields the estimate of group-level student performance:

$$\theta = \frac{1}{M}\sum_{i=1}^{M}\theta_i \tag{1}$$

where M is the number of imputations, typically 5 in international assessments.

What is particularly challenging is the calculation of the standard error of θ, that is, the uncertainty associated with its estimation. This is because the complex test and sampling design introduce two sources of error in the estimation of θ: imputation error and sampling error, respectively. And these errors cannot be calculated with standard routines of statistical software. The calculation of correct standard errors is important for making valid comparisons of performance between countries or boys and girls, for example. It is for this reason that specialized software like **intsvy** package is required.

2.1. Rotated test design

The total item pool of international assessments consists of hundreds of items and demand hours of testing time in order to produce valid and reliable measures of student achievement constructs. Clearly, it is not feasible to administer a test including the entire item pool for logistic, fatigue, and testing time issues in general. International assessments therefore employ a rotated design form in order to achieve a balance between validity and reasonable testing time. Test items are arranged into clusters that in turn are distributed between booklets administered to students. Clusters are distributed such that it is possible to link test booklets through clusters in common. Cluster linkage between booklets ensures the comparability of results between students and reporting on the same scale. Rotated test forms introduce technical complexities in the estimation of student performance, since students respond only to a subset of items, the ones in the booklet, but inferences on student performance are made as if the students had responded to the entire assessment through plausible value techniques.

The plausible values approach combines item response theory and latent regression techniques to produce unbiased estimates of student performance at the population level (von Davier, Gonzalez, and Mislevy 2009). Importantly, plausible values are not used to infer performance at the individual level, since students responded only to a subset of the items and measurement errors at the individual level tend to be large. Plausible values are random draws from the estimated posterior distribution of student performance given student responses to the subset of test items and background information collected in questionnaires. The average of plausible values estimates was calculated in equation 1. The variance reflects uncertainty in the estimation associated with making multiple imputations of plausible values based on the posterior distribution of student performance. The formula of the imputation variance, $Var_{imp}[\theta]$, is as follows (Little and Rubin 1987):

$$Var_{imp}[\theta] = \frac{1}{M-1}\sum_{i=1}^{M}(\theta_i - \theta)^2 \tag{2}$$

2.2. Complex sample design

Student samples in international LSA are selected in two stages: schools are sampled in the first stage and students within the school in the second stage. For example, 15-year-olds are sampled randomly within schools in PISA and intact classes within schools are sampled randomly in TIMSS and PIRLS. The sampling error takes into account the uncertainty related with the sample selection, as different samples of schools and students from the population not necessarily yield the same estimates. The sampling error formula under two-stage sampling cannot assume that observations are independent as in random sampling because students within schools tend to share similar characteristics, for example, family socio-economic status (SES) and the instructional setting. Compared to random sampling, the dependency of observations within schools in two-stage sampling tends to reduce the amount of information and increase the uncertainty of estimates, that is, the standard error. For example, a two-stage sample of 100 students per school in 10 schools will likely yield less information than a random sample of 1000 students. In one extreme scenario, if all students within schools are identical the two-stage sample will represent 10 students and not 1000. In the other extreme, if all students within schools are uncorrelated the two-stage sample size will be 1000. In real data the dependency of observations lies between these two scenarios (i.e., a sample size of 10 and 1000 students).

Replicate weights are used in international LSA to calculate sampling errors. Each replicate weight represents a sample of schools and the variability between estimates of the replicate weights samples the uncertainty due to school sample selection or the sampling error. Like multilevel models, replicate weights estimation introduces randomness in the selection of schools. Multilevel models does it by introducing random effects and replicate weights estimation by creating different samples in the data whilst maintaining the traditional ordinary least squares (OLS) model. From this perspective, replicate weights can be regarded as a case of adapting the data to the model and multilevel models as one of adapting the model to the data. Further, sample variation with replicate weights of international LSA is not entirely random but takes into account stratification (e.g., one school is selected at random within each stratum for each replicate weight). As a result, multilevel models and replicate weights estimation do not yield exactly the same results. To the extent that multilevel models do not take into account stratification information, they tend and produce standard errors that are larger than for regression analysis using replicate weights. There are different replication techniques for two-stage sampling. TIMSS and PIRLS employ the Jackknife technique and PISA employs Balanced Repeated Replication (BRR) with Fay's modification. The principles underpinning these techniques and worked examples are presented in technical reports of international assessments (e.g. OECD 2012). Here we will just present the formulas.

The sampling variance for PIRLS and TIMSS is:

$$Var_{sml}[\theta] = \sum_{j=1}^{R} (\theta_j - \theta)^2 \tag{3}$$

The sampling variance in PISA is:

$$Var_{sml}[\theta] = \frac{1}{G(1-k)} \sum_{j=1}^{R} (\theta_j - \theta)^2 \tag{4}$$

where R is the number of replicate weights, in PIRLS and TIMSS 75 Jackknife replicate weights and in PISA 80 BRR replicate weights.

For PIAAC study the story is slighty more complicated. Different replication method and different numbers of replications were used in different countries. Thus the general formula for the sampling variance in PIAAC is:

$$Var_{sml}[\theta] = c \sum_{j=1}^{R} (\theta_j - \theta)^2 \tag{5}$$

where $c = \frac{G-1}{G}$ (so called random groups (delete-one) approach) for Australia, Austria, Canada, Denmark and Germany while $c = 1$ (so called paired jackknife) for other countries. See `intsvy::piaacReplicationScheme` table or PIAAC Technical Report (2014) for more details.

For student performance data, the sampling variance is the average across the 5 plausible values:

$$Var_{sml}[\theta] = \frac{1}{5} (Var_1[\theta] + Var_2[\theta] + Var_3[\theta] + Var_4[\theta] + Var_5[\theta]) \tag{6}$$

TIMSS and PIRLS, however, use an unbiased shortcut for calculating the sampling variance. Instead of the average, the sampling variance is equal to the sampling variance for the first plausible value, $Var_1[\theta]$.

2.3. Standard error formula

The total standard error for single observed variables in international assessment data is equal to the sampling error. For the plausible values of student performance the standard error additionally takes into account imputation error. The formula combines the sampling error and the imputation error as follows:

$$Var_{tot}[\theta] = Var_{sml}[\theta] + \left(1 + \frac{1}{M}\right) \times Var_{imp}[\theta] \tag{7}$$

The standard error is the square root:

$$SE[\theta] = \sqrt{Var_{tot}[\theta]} \tag{8}$$

3. Applied examples

Package **intsvy** uses the formulas above to calculate correct standard errors for different statistics, including means, standard deviations, percentages, correlations, and regression coefficients with data from observed variables or plausible values of student performance. It is installed and loaded into R with:

```
> install.packages("intsvy")
> library(intsvy)
```

3.1. Select and merge data

intsvy provides tools for selecting and importing data into R. Data can be imported in two steps. First, functions `var.label` facilitate data selection by reporting variable names, variable labels and names of participating countries in available datasets. Secondly, functions `select.merge` produce a single data frame for selected variables and countries. Sampling variables (i.e., replicate weights and total weights) and plausible variables are selected automatically and a country identifier variable with the long version of the country name: `IDCNTRYL` is created.

TIMSS and PIRLS

Data importing tools are particularly useful for TIMSS and PIRLS, because original datasets available from the IEA Data Repository (`http://rms.iea-dpc.org/`) are organised in a number of files by country, school grade, and survey instrument (e.g., student questionnaire, home questionnaire, teacher questionnaire). With functions `timssg8.select.merge`, `timssg4.select.merge`, and `pirls.select.merge` the useR can produce a single data frame containing selected data (i.e., variables and countries) without needing to understand the original data structure in multiple files.

For example, the following command outputs variable names and variable labels for TIMSS 2011 data of 8th Grade students by survey instrument as well as the name and abbreviations of participating countries.

```
> timssg8.var.label(folder = "C:/TIMSS/TIMSS 2011/Grade 8/Data")
```

The `folder` argument indicates where the data are located. The output is automatically stored on a text file saved in the working directory (i.e., `getwd()`). The file location and name can be modified with the `output` and `name` arguments.

Subsequently, selected data of specific variables and countries can be imported into a data frame with `timssg8.select.merge`. For example, selected variables from the student and school questionnaire in TIMSS 2011 Grade 8 can be imported for Australia, Bahrain, Armenia, and Chile:

```
> timss8g <- timssg8.select.merge(folder="C:/TIMSS/TIMSS 2011/Grade 8/Data",
countries=c("AUS", "BHR", "ARM", "CHL"), student =c("BSDGEDUP", "ITSEX",
"BSDAGE", "BSBGSLM", "BSDGSLM"),school=c("BCBGDAS", "BCDG03"))
```

Similarly, selected PIRLS 2011 data from the student, home, and school questionnaires can be imported into a data frame:

```
> pirls <- pirls.select.merge(folder= "C:/PIRLS/PIRLS 2011/Data",
countries= c("AUS", "AUT", "AZE", "BFR"), student= c("ITSEX", "ASDAGE",
"ASBGSMR"), home= c("ASDHEDUP", "ASDHOCCP", "ASDHELA", "ASBHELA"),
school= c("ACDGDAS", "ACDGCMP", "ACDG03"))
```

PISA and PIAAC

The structure of `pisa.select.merge` is different, because original datasets available from the OECD website (`http://www.oecd.org/pisa/pisaproducts/`) are organised in large files for the student, school, and parent questionnaire containing data for all participating countries. As before, `pisa.var.label` reports names of variables and countries in the data for data selection. And `pisa,select.merge` creates a data frame with the selected data. For example, selected data from the student and school questionnaire can be imported for Hong Kong, the United States, Sweden, Poland, and Peru:

```
> pisa <- pisa.select.merge(folder = "C:/PISA/PISA 2012/Data",
school.file="INT_SCQ12_DEC03.sav", student.file="INT_STU12_DEC03.sav",
student= c("ST01Q01", "ST04Q01", "ESCS", "PARED"), school =
c("CLSIZE", "TCSHORT"), countries = c("HKG", "USA", "SWE", "POL", "PER"))
```

An alternative way to access data from PIAAC or PISA studies is by using R packages with converted data. Since these datasets have significant size, up to few hundreds MB, they are not available on CRAN. But they can be downloaded from pbiecek account on github.

Packages with consecutive releases of PISA data are named **PISA2000lite**, **PISA2003lite**, **PISA2006lite**, **PISA2009lite**, **PISA2012lite**) while the package with PIAAC data is named **PIAAC**. For example, the following code installs the package with PISA 2012 data:

```
> library(devtools)
> install_github("pbiecek/PISA2012lite")
```

Dictionaries with variable names are available in `student2012dict`, `school2012dict` and `parent2012dict` vectors. With aid of the `grep` function it is possible to find a desired variable. Here is an example for finding the variable with the number of books at home.

```
> library(PISA2012lite)
> grep(student2012dict, pattern = "books", value = TRUE)
                                    ST26Q10
                  "Possessions - textbooks"
                                    ST26Q11
"Possessions - <technical reference books>"
                                    ST28Q01
                   "How many books at home"
```

Variable names, such as `ST28Q01` can be used to extract information of specific variables from data frames `student2012`, `school2012` and `parent2012`. For example:

```
> table(student2012["ST28Q01"])

        0-10 books         11-25 books        26-100 books       101-200 books
             95042               97335              135184               68350
     201-500 books  More than 500 books
             49267               28587
```

For PIAAC, the following code installs the package with the data:

```
> library(devtools)
> install_github("pbiecek/PIAAC")
```

A single data frame with PIAAC data is available in the `piaac` data frame while a dictionary for variable names is stored in the `piaacdict` vector.

```
> library(PIAAC)

> grep(piaacdict, pattern="Number of books", value = TRUE)
                                 J_Q08
"Background - Number of books at home"

> table(piaac["J_Q08"])

   10 books or less      11 to 25 books     26 to 100 books    101 to 200 books
              21590               23069               47999               25938
   201 to 500 books More than 500 books
              20125               10760
```

3.2. Average achievement scores with plausible values

Functions `pisa.mean.pv`, `piaac.mean.pv`, `timss.mean.pv`, and `pirls.mean.pv` calculate average estimates and associated standard errors for achievement variables with plausible values. Three main arguments are supplied by the useR: `pvlabel`, `by`, and `data`. Argument `pvlabel` indicates the part of the label in common for the plausible values variables (e.g., `"READ"`, "MATH"). Argument `by` defines the level of grouping for the analysis (e.g., "IDCNTRYL") and may contain more than one level (e.g., `c("IDCNTRYL", "SEX")`). And argument `data` defines the dataset to be used in the analysis.

PISA and PIAAC

For example, in PISA 2012, the average math performance by education system and associated standard errors can be calculated as follows (see OECD 2014, p. 305):

```
> pisa.mean.pv(pvlabel = "MATH", by = "IDCNTRYL", data = pisa)

                  IDCNTRYL Freq   Mean s.e.    SD  s.e
1         China, Hong Kong 4670 561.24 3.22 96.31 1.92
2                     Peru 6035 368.10 3.69 84.36 2.20
3                   Poland 4607 517.50 3.62 90.37 1.89
4                   Sweden 4736 478.26 2.26 91.75 1.28
5 United States of America 4978 481.37 3.60 89.86 1.30
```

The argument `pvlabel="MATH"` refers to the name suffix in common of the variables containing the plausible values variables: `PV1MATH`, `PV2MATH`, `PV3MATH`, `PV4MATH`, and `PV5MATH`. For science and reading, this argument should be changed to `pvlabel="READ"` and `pvlabel="SCIE"`, for example.

More levels of grouping can be included in the analysis. For example the following code produces results by education system (`IDCNTRYL`) and the student's sex (`ST04Q01`), whilst exporting results (`export=TRUE`) into a comma-separated value (csv) file (see OECD 2014, p. 305):

```
> pisa.mean.pv(pvlabel = "MATH", by = c("IDCNTRYL", "ST04Q01"), data = pisa,
export=TRUE, name="PISA mean by sex", folder="C:/PISA/PISA 2012/Results")

                   IDCNTRYL ST04Q01 Freq   Mean s.e.     SD  s.e
1          China, Hong Kong  Female 2161 552.96 3.94  90.51 2.23
2          China, Hong Kong    Male 2509 568.38 4.55 100.49 2.18
3                      Peru  Female 3118 358.92 4.75  83.44 2.61
4                      Peru    Male 2917 377.82 3.65  84.24 2.51
5                    Poland  Female 2388 515.53 3.76  86.38 1.59
6                    Poland    Male 2219 519.56 4.25  94.32 2.65
7                    Sweden  Female 2378 479.63 2.41  87.60 1.60
8                    Sweden    Male 2358 476.92 2.97  95.63 1.88
9  United States of America  Female 2453 479.00 3.91  87.08 1.71
10 United States of America    Male 2525 483.65 3.81  92.40 1.61
```

The resulting csv file is named "PISA mean by sex.csv" and is located in "C:/PISA/PISA 2012/Results". It can be imported directed into a spreadsheet for further analysis or for formatting for publication.

For PIAAC, numeracy average performance can be calculated with `piaac.mean.pv` function as follows:

```
> head(piaac.mean.pv(pvlabel="NUM", by="CNTRYID", data=piaac,
   export=FALSE))

         CNTRYID  Freq   Mean s.e.    SD  s.e
1        Austria  5130 275.04 0.88 48.84 0.64
2        Belgium  5463 280.39 0.83 49.27 0.67
3         Canada 26683 265.24 0.70 55.60 0.54
4 Czech Republic  6102 275.73 0.93 43.59 0.78
5        Denmark  7328 278.28 0.73 51.13 0.59
6        Estonia  7632 273.12 0.53 45.45 0.48
```

Also, results by country and age group can be produced with:

```
> head(piaac.mean.pv(pvlabel="NUM", by=c("CNTRYID", "AGEG10LFS"), data=piaac,
   export=FALSE))

  CNTRYID  AGEG10LFS Freq   Mean s.e.    SD  s.e
1 Austria 24 or less  898 279.27 1.63 46.15 1.82
2 Austria      25-34  958 282.06 1.73 49.98 1.63
3 Austria      35-44 1117 281.35 2.01 50.26 1.40
4 Austria      45-54 1188 274.48 1.67 46.49 1.24
5 Austria    55 plus  969 257.48 1.74 46.83 1.47
6 Belgium 24 or less  994 282.82 1.74 45.07 1.63
```

TIMSS and PIRLS

Similar analysis can be conducted with TIMSS and PIRLS data.

In TIMSS 2011, Grade 8, math average performance by education system can be calculated as follows (see Foy, Arora, and Stanco 2013, p. 15):

```
> timss.mean.pv(pvlabel="BSMMAT", by= "IDCNTRYL", data=timss8g)

   IDCNTRYL  Freq   Mean s.e.    SD  s.e
1   Armenia 23384 466.59 2.73 90.68 1.73
2 Australia 30224 504.80 5.09 85.42 3.36
3   Bahrain 18560 409.22 1.96 99.57 1.72
4     Chile 23340 416.27 2.59 79.65 1.85
```

And results by education system and student's sex as follows (see Foy *et al.* 2013, p. 18):

```
> timss.mean.pv(pvlabel="BSMMAT", by= c("IDCNTRYL", "ITSEX"), data=timss8g)
```

```
   IDCNTRYL ITSEX  Freq   Mean s.e.     SD  s.e
1   Armenia  GIRL 11576 471.52 3.07  87.13 1.81
2   Armenia   BOY 11808 461.86 3.21  93.72 2.24
3 Australia  GIRL 14988 500.41 4.72  82.72 3.59
4 Australia   BOY 15236 509.16 7.26  87.80 4.82
5   Bahrain  GIRL  9152 430.78 2.51  87.23 1.93
6   Bahrain   BOY  9408 387.89 3.07 106.20 2.26
7     Chile  GIRL 12532 409.46 3.23  79.97 2.39
8     Chile   BOY 10808 423.94 3.05  78.59 2.03
```

In PIRLS 2011, reading performance results by country can be calculated as follows (see Foy and Drucker 2013, p. 15):

```
> pirls.mean.pv(pvlabel="ASRREA", by= "IDCNTRYL", data=pirls)
```

```
          IDCNTRYL Freq   Mean s.e.    SD  s.e
1        Australia 6126 527.37 2.21 80.22 1.31
2          Austria 4670 528.88 1.95 63.38 0.95
3       Azerbaijan 4881 462.30 3.33 67.83 1.68
4 Belgium (French) 3727 506.12 2.88 64.67 1.57
```

And results by country and student's sex as follows (see Foy and Drucker 2013, p. 18):

```
> pirls.mean.pv(pvlabel="ASRREA", by= c("IDCNTRYL", "ITSEX"), data=pirls)
```

```
          IDCNTRYL ITSEX Freq   Mean s.e.    SD  s.e
1        Australia  GIRL 3048 535.79 2.67 78.20 1.62
2        Australia   BOY 3078 519.20 2.73 81.30 1.75
3          Austria  GIRL 2274 532.76 2.18 62.00 1.21
4          Austria   BOY 2396 525.19 2.32 64.44 1.48
5       Azerbaijan  GIRL 2241 469.57 3.56 67.31 1.94
6       Azerbaijan   BOY 2640 455.82 3.47 67.63 1.85
7 Belgium (French)  GIRL 1815 508.85 3.11 63.11 2.01
8 Belgium (French)   BOY 1912 503.51 3.11 66.02 1.62
```

Unlike PISA, the argument `pvlabel` for TIMSS and PIRLS refers to the prefix of the variable names containing the plausible values. For example, variable names of math plausible values in TIMSS are `BSMMAT01`, `BSMMAT02`, `BSMMAT03`, `BSMMAT04`, and `BSMMAT01` and variable names of reading plausible values in PIRLS are `ASRREA01`, `ASRREA02`, `ASRREA03`, `ASRREA04`, and `ASRREA05`. Equally, results can be exported into a .csv file using the `export=TRUE` argument.

3.3. Average estimates without plausible values

It is also possible to calculate means and standard errors for variables without plausible values, that is, for all of the other variables in the datasets, using functions `pisa.mean`, `piaac.mean`, `timss.mean`, and `pirls.mean`.

PISA and PIAAC

For example, the following code calculates the average highest level of education of parents in years of schooling (PARED) by education system in PISA 2012 (see OECD 2013, p. 183):

```
> pisa.mean(variable="PARED", by="IDCNTRYL", data=pisa)

                  IDCNTRYL Freq  Mean Std.err.
1         China, Hong Kong 4477 11.41     0.14
2                     Peru 5960 11.46     0.14
3                   Poland 4481 12.68     0.06
4                   Sweden 4496 14.09     0.04
5 United States of America 4869 13.65     0.09
```

The following example calculates the average age of participants by country for the PIAAC data.

```
> head(piaac.mean(variable="AGE_R", by="CNTRYID", data=piaac, export=FALSE))

          CNTRYID Freq  Mean s.e.
1         Belgium 5463 41.78 0.03
2  Czech Republic 6102 40.54 0.04
3         Denmark 7328 41.03 0.04
4         Estonia 7632 40.05 0.03
5         Finland 5464 41.40 0.04
6          France 6993 40.76 0.03
```

TIMSS and PIRLS

For TIMSS 2011, the following code calculates the average of the index *Students Like Learning Mathematics* (BSBGSLM) by education system (see Foy *et al.* 2013, p. 27):

```
> timss.mean(variable="BSBGSLM", by='IDCNTRYL', data=timss8g)

   IDCNTRYL     n  Mean Std.err.
1   Armenia 22504 10.87     0.05
2 Australia 29556  9.32     0.06
3   Bahrain 18324  9.77     0.03
4     Chile 23088  9.76     0.04
```

For PIRLS 2011, the following calculates the average of the index *Early Literacy Activities before Beginning Primary School* by education system (see Foy and Drucker 2013, p. 28):

```
> pirls.mean(variable='ASBHELA', by= 'IDCNTRYL', data=pirls)

          IDCNTRYL    n  Mean Std.err.
```

```
1        Australia 3232 10.84     0.06
2          Austria 4393  9.98     0.03
3       Azerbaijan 4509  9.47     0.07
4 Belgium (French) 3383  9.69     0.04
```

3.4. Regression analysis

Functions `pisa.reg.pv`, `timss.reg.pv`, and `pirls.reg.pv` perform regression analysis. For example, differences in mean performance calculated previously for boys and girls can be tested for statistical significance using a regression approach.

PISA and PIAAC

For example, significance tests can be conducted in PISA 2012 as follows (see OECD 2014, p. 305):

```
> pisa.reg.pv(pvlabel="MATH", x="ST04Q01", by = "IDCNTRYL", data=pisa)

$`China, Hong Kong`
            Estimate Std. Error t value
(Intercept)   552.96       3.94  140.18
ST04Q01Male    15.42       5.69    2.71
R-squared       0.64       0.49    1.31

$Peru
            Estimate Std. Error t value
(Intercept)   358.92       4.75   75.53
ST04Q01Male    18.90       3.92    4.82
R-squared       1.26       0.54    2.33

$Poland
            Estimate Std. Error t value
(Intercept)   515.53       3.76  137.28
ST04Q01Male     4.03       3.42    1.18
R-squared       0.05       0.09    0.59

$Sweden
            Estimate Std. Error t value
(Intercept)   479.63       2.41  199.08
ST04Q01Male    -2.71       2.98   -0.91
R-squared       0.02       0.05    0.41

$`United States of America`
            Estimate Std. Error t value
(Intercept)   479.00       3.91  122.52
ST04Q01Male     4.65       2.80    1.66
```

```
R-squared       0.07        0.09    0.81
```

Argument `x` defines the independent variable(s), in this case `ST04Q01`, but more variable can be included separated by commas (e.g., `x=c("ST04Q01", "ESCS")`). The output is a list with regression results by education system. Coefficient `ST04Q01Male` captures differences between boys and girls and its t-value indicates whether they are statistically significant. R-squared values range from 0 to 100.

The following provides an example of regression with literacy scores as dependent variable and student's sex and country as independent variable for PIAAC data.

```
> rmodelLG <- piaac.reg.pv(pvlabel="LIT", x="GENDER_R", by = "CNTRYID", data=piaac,
                        export=FALSE)
> rmodelLG[1:3]

$Austria
               Estimate Std. Error t value
(Intercept)      271.53       1.04  259.90
GENDER_RFemale    -4.14       1.32   -3.13
R-squared          0.22       0.14    1.58

$Belgium
               Estimate Std. Error t value
(Intercept)      278.09       0.97  287.08
GENDER_RFemale    -5.27       1.21   -4.36
R-squared          0.31       0.15    2.17

$Canada
               Estimate Std. Error t value
(Intercept)      274.49       0.86  317.75
GENDER_RFemale    -2.30       1.20   -1.92
R-squared          0.06       0.05    1.04
```

TIMSS and PIRLS

Similarly, tests of mean differences between boys and girls in TIMSS 2011, Grade 8 can be performed using a regression approach as follows (see Foy *et al.* 2013, p. 21):

```
> timss.reg.pv(pvlabel="BSMMAT", by="IDCNTRYL", x="ITSEX", data=timss8g)

$Armenia
            Estimate Std. Error t value
(Intercept)   471.52       3.07  153.75
ITSEXBOY       -9.66       3.10   -3.12
R-squared       0.29       0.18    1.61
```

```
$Australia
             Estimate Std. Error t value
(Intercept)    500.41       4.72  105.93
ITSEXBOY         8.75       6.90    1.27
R-squared        0.27       0.32    0.83

$Bahrain
             Estimate Std. Error t value
(Intercept)    430.78       2.51  171.50
ITSEXBOY       -42.89       3.99  -10.74
R-squared        4.64       0.85    5.44

$Chile
             Estimate Std. Error t value
(Intercept)    409.46       3.23  126.86
ITSEXBOY        14.48       3.63    3.99
R-squared        0.82       0.44    1.89
```

The same mean different test can be performed for PIRLS 2011 with a regression (see Foy and Drucker 2013, p. 21):

```
> pirls.reg.pv(pvlabel="ASRREA", by="IDCNTRYL", x="ITSEX", data=pirls)

$Australia
             Estimate Std. Error t value
(Intercept)    535.79       2.67  200.57
ITSEXBOY       -16.58       3.11   -5.33
R-squared        1.07       0.40    2.69

$Austria
             Estimate Std. Error t value
(Intercept)    532.76       2.18  244.47
ITSEXBOY        -7.58       2.31   -3.28
R-squared        0.36       0.24    1.50

$Azerbaijan
             Estimate Std. Error t value
(Intercept)    469.57       3.56  131.76
ITSEXBOY       -13.75       2.34   -5.87
R-squared        1.02       0.36    2.83

$`Belgium (French)`
             Estimate Std. Error t value
(Intercept)    508.85       3.11  163.70
ITSEXBOY        -5.34       2.34   -2.28
R-squared        0.18       0.14    1.26
```

Also, functions `pisa.reg`, `timss.reg`, and `pirls.reg` perform regression analysis for observed variables without plausible values.

3.5. Frequency tables

Functions `pisa.table`, `piaac.table`, `timss.table`, and `pirls.table` produce frequency tables including percentages and associated standard errors.

For example, the following code produces the frequency and percentage of students in each school grade level (i.e., `variable="ST01Q01"`) by education system in PISA 2012 (see OECD 2014, p. 274):

```
> pisa.table(variable="ST01Q01", by="IDCNTRYL", data=pisa)

                  IDCNTRYL ST01Q01 Freq Percentage Std.err.
1          China, Hong Kong       7   51       1.06     0.14
2          China, Hong Kong       8  300       6.47     0.41
3          China, Hong Kong       9 1205      25.94     0.72
4          China, Hong Kong      10 3088      65.01     0.91
5          China, Hong Kong      11   26       1.51     1.36
6                      Peru       7  150       2.69     0.44
7                      Peru       8  466       7.79     0.54
8                      Peru       9 1056      18.10     0.67
9                      Peru      10 2907      47.68     0.95
10                     Peru      11 1456      23.74     0.82
11                   Poland       7   20       0.53     0.13
12                   Poland       8  158       4.08     0.37
13                   Poland       9 4416      94.89     0.42
14                   Poland      10   13       0.50     0.22
15                   Sweden       7    1       0.03     0.03
16                   Sweden       8  159       3.69     0.35
17                   Sweden       9 4496      94.05     0.64
18                   Sweden      10   80       2.23     0.54
19 United States of America       8    6       0.26     0.14
20 United States of America       9  538      11.74     1.06
21 United States of America      10 3633      71.21     1.10
22 United States of America      11  794      16.58     0.83
23 United States of America      12    7       0.21     0.11
```

With PIAAC data, the percentages of age groups by country can be calculated as follows:

```
head(piaac.table(variable="AGEG10LFS", by="CNTRYID", data=piaac))
  CNTRYID  AGEG10LFS Freq Percentage Std.err.
1 Austria 24 or less  898      16.00     0.04
2 Austria      25-34  958      19.11     0.06
3 Austria      35-44 1117      22.18     0.07
4 Austria      45-54 1188      23.83     0.07
```

```
5 Austria    55 plus  969      18.89     0.04
6 Belgium 24 or less  994      15.33     0.03
```

With TIMSS data, it is possible to calculate the percentage of students according to how much they like learning mathematics reported by own students (see Foy *et al.* 2013, p. 29):

```
> timss.table(variable="BSDGSLM", by="IDCNTRYL", data=timss8g)

    IDCNTRYL                              BSDGSLM  Freq Percentage Std.err.
1    Armenia          LIKE LEARNING MATHEMATICS  9684      42.92     0.97
2    Armenia SOMEWHAT LIKE LEARNING MATHEMATICS  8724      39.48     0.76
3    Armenia   DO NOT LIKE LEARNING MATHEMATICS  4096      17.60     0.97
4  Australia          LIKE LEARNING MATHEMATICS  4272      15.67     0.94
5  Australia SOMEWHAT LIKE LEARNING MATHEMATICS 11940      39.81     0.87
6  Australia   DO NOT LIKE LEARNING MATHEMATICS 13344      44.53     1.41
7    Bahrain          LIKE LEARNING MATHEMATICS  4288      23.75     0.64
8    Bahrain SOMEWHAT LIKE LEARNING MATHEMATICS  7024      38.37     0.86
9    Bahrain   DO NOT LIKE LEARNING MATHEMATICS  7012      37.88     0.84
10     Chile          LIKE LEARNING MATHEMATICS  5156      22.06     0.86
11     Chile SOMEWHAT LIKE LEARNING MATHEMATICS  9164      40.21     0.89
12     Chile   DO NOT LIKE LEARNING MATHEMATICS  8768      37.73     0.97
```

And using school level data, we can calculate the percentage of students in schools classified by the socio-economic composition reported by principals (see Foy *et al.* 2013, p. 36):

```
> timss.table(variable="BCDG03", by="IDCNTRYL", data=timss8g)

    IDCNTRYL                      BCDG03  Freq Percentage Std.err.
1    Armenia               MORE AFFLUENT  8340      34.78     3.70
2    Armenia         NEITHER MORE AFFLUENT...  5316      24.25     3.59
3    Armenia          MORE DISADVANTAGED  8632      40.97     3.68
4  Australia               MORE AFFLUENT  8472      32.49     3.36
5  Australia     NEITHER MORE AFFLUENT... 10140      38.54     3.74
6  Australia          MORE DISADVANTAGED  7200      28.97     3.11
7    Bahrain               MORE AFFLUENT  7816      45.30     0.32
8    Bahrain         NEITHER MORE AFFLUENT...  4572      27.87     0.23
9    Bahrain          MORE DISADVANTAGED  4204      26.83     0.34
10     Chile               MORE AFFLUENT  3244      12.16     2.32
11     Chile         NEITHER MORE AFFLUENT...  5564      31.66     4.07
12     Chile          MORE DISADVANTAGED  8476      56.18     3.86
```

3.6. Performance benchmarks

Functions `pisa.ben.pv`, `timss.ben.pv`, and `pirls.ben.pv` calculate percentages of students in each proficiency level and associated standard errors. Proficiency levels are defined by PISA, TIMSS, and PIRLS studies and can be modified by the user.

For example, in PISA 2012 the percentage of students in each math proficiency level can be calculated as follows (see OECD 2014, p. 298):

```
> pisa.ben.pv(pvlabel="MATH", cutoff=c(357.77, 420.07, 482.38, 544.68,
606.99, 669.30), by="IDCNTRYL", data=pisa)

                   IDCNTRYL       Benchmarks Percentage Std. err.
1          China, Hong Kong        <=357.77        2.57      0.36
2          China, Hong Kong (357.77, 420.07]       5.94      0.61
3          China, Hong Kong (420.07, 482.38]      12.02      0.77
4          China, Hong Kong (482.38, 544.68]      19.69      0.97
5          China, Hong Kong (544.68, 606.99]      26.07      1.09
6          China, Hong Kong  (606.99, 669.3]      21.45      0.96
7          China, Hong Kong           >669.3      12.26      0.95
8                      Peru         <=357.77      46.97      1.79
9                      Peru (357.77, 420.07]      27.61      0.88
10                     Peru (420.07, 482.38]      16.13      1.00
11                     Peru (482.38, 544.68]       6.66      0.68
12                     Peru (544.68, 606.99]       2.06      0.38
13                     Peru (606.99, 669.30]       0.55      0.20
14                     Peru           >669.3       0.03      0.03
15                   Poland         <=357.77       3.28      0.38
16                   Poland (357.77, 420.07]      11.10      0.77
17                   Poland (420.07, 482.38]      22.08      0.93
18                   Poland (482.38, 544.68]      25.46      0.94
19                   Poland (544.68, 606.99]      21.34      1.12
20                   Poland (606.99,  669.3]      11.74      0.78
21                   Poland           >669.3       5.00      0.80
22                   Sweden         <=357.77       9.55      0.68
23                   Sweden (357.77, 420.07]      17.53      0.76
24                   Sweden (420.07, 482.38]      24.69      0.92
25                   Sweden (482.38, 544.68]      23.93      0.78
26                   Sweden (544.68, 606.99]      16.30      0.69
27                   Sweden (606.99,  669.3]       6.46      0.49
28                   Sweden           >669.3       1.55      0.25
29 United States of America         <=357.77       7.96      0.73
30 United States of America (357.77, 420.07]      17.89      0.98
31 United States of America (420.07, 482.38]      26.25      0.84
32 United States of America (482.38, 544.68]      23.34      0.93
33 United States of America (544.68, 606.99]      15.79      0.91
34 United States of America (606.99.  669.3]       6.58      0.61
35 United States of America           >669.3       2.19      0.34
```

The argument `cutoff` specifies proficiency levels for math performance in PISA 2012. These values are the default, can be omitted for 2012 data and should be modified for data with different proficiency levels.

Likewise, percentage of students according to performance levels established by TIMSS and PIRLS can be calculated. For TIMSS 2011, for example (see Foy *et al.* 2013, p. 24):

```
> timss.ben.pv(pvlabel="BSMMAT", by="IDCNTRYL", cutoff =
c(400, 475, 550, 625), data=timss8g)

    IDCNTRYL       Benchmark Percentage Std. err.
1    Armenia At or above 400      76.38      1.16
2    Armenia At or above 475      49.02      1.37
3    Armenia At or above 550      17.65      0.88
4    Armenia At or above 625       3.23      0.40
5  Australia At or above 400      89.17      1.08
6  Australia At or above 475      62.94      2.40
7  Australia At or above 550      28.65      2.63
8  Australia At or above 625       8.68      1.68
9    Bahrain At or above 400      53.49      0.79
10   Bahrain At or above 475      26.19      0.65
11   Bahrain At or above 550       7.97      0.68
12   Bahrain At or above 625       1.26      0.25
13     Chile At or above 400      56.86      1.57
14     Chile At or above 475      22.95      1.11
15     Chile At or above 550       5.35      0.62
16     Chile At or above 625       0.56      0.16
```

And for PIRLS 2011 (see Foy and Drucker 2013, p. 24):

```
> pirls.ben.pv(pvlabel="ASRREA", by="IDCNTRYL", data=pirls)

           IDCNTRYL       Benchmark Percentage Std. err.
1         Australia At or above 400      92.93      0.67
2         Australia At or above 475      75.62      1.03
3         Australia At or above 550      41.91      1.14
4         Australia At or above 625       9.93      0.65
5           Austria At or above 400      97.10      0.35
6           Austria At or above 475      80.38      0.94
7           Austria At or above 550      39.05      1.50
8           Austria At or above 625       5.22      0.54
9        Azerbaijan At or above 400      81.86      1.60
10       Azerbaijan At or above 475      45.16      2.10
11       Azerbaijan At or above 550       8.94      0.93
12       Azerbaijan At or above 625       0.44      0.28
13 Belgium (French) At or above 400      93.79      1.08
14 Belgium (French) At or above 475      70.39      1.67
15 Belgium (French) At or above 550      25.50      1.39
16 Belgium (French) At or above 625       2.25      0.49
```

As before, the argument `cutoff` can be omitted since these are the benchmark levels established by PIRLS 2011. For different benchmarks, the cut-off values should be modified by the useR. Also, more grouping levels for the analysis can be added with `by`.

3.7. Data visualisation

The functions presented above allow to precisely estimate averages, frequencies or regression coefficients together with their standard errors. Since large tables filled with numbers could be difficult to understand at first sigh, **intsvy** provides functions for data visualization that facilitate interpretation of results.

Function	Class of returned object	Generic plot function
`pisa.table()`, `piaac.table()`, `pirls.table()`, `timms.table()`	`intsvy.table`	`plot.intsvy.table()`
`pisa.mean.pv()`, `piaac.mean.pv()`, `pirls.mean.pv()`, `timms.mean.pv()`, `pisa.mean()`, `piaac.mean()`, `pirls.mean()`, `timms.mean()`	`intsvy.mean`	`plot.intsvy.mean()`
`pisa.reg.pv()`, `piaac.reg.pv()`, `pirls.reg.pv()`, `timms.reg.pv()`, `pisa.reg()`, `piaac.reg()`, `pirls.reg()`, `timms.reg()`	`intsvy.reg`	`plot.intsvy.reg()`

Table 1: Analytical functions implemented in **intsvy** package are presented in first column. The second column presents classes of returned objects. For each class, a generic version of `plot()` function, full name of these functions is presented in the third column.

The architecture of developed solution is presented in Table 1. Below examples for each class of analytical functions are presented.

The output of functions `piaac.table`, `timms.table`, `pirls.table` and `pisa.table` is an object of the class `intsvy.table`. The overloaded `plot` function produces a **ggplot2** based barplot that summarizes frequency tables. Optional arguments for the `plot.intsvy.table()` are `stacked` (should bars be stacked or not) and `se` (should standard error be plotted or not).

The following example calculates and plots two tables based on PIAAC dataset.

```
> library(PIAAC)
> ptable <- piaac.table(variable="AGEG10LFS", data=piaac)
> plot(ptable)
> ptableCA <- piaac.table(variable="AGEG10LFS", by=c("CNTRYID", "GENDER_R"),
              data=piaac)
> plot(na.omit(ptableCA), stacked=TRUE)
```

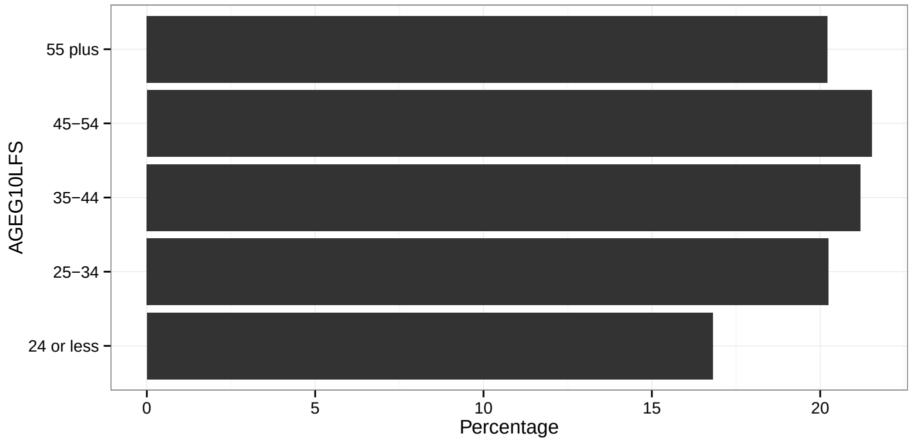

Figure 1: Graphical summary of a frequency table. This example presents structure of age groups in PIAAC dataset.

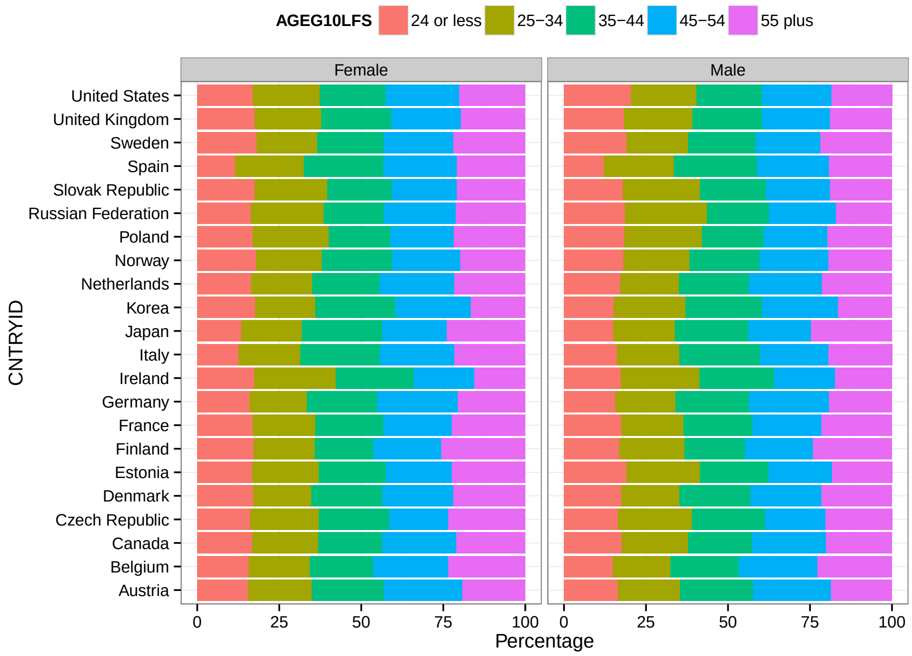

Figure 2: Graphical summary of a frequency table with grouping variable. This example presents structure of age groups by country and gender in PIAAC dataset.

Functions `*.mean.pv`, and `*.mean` (where * stands for pisa, piaac, timms and pirls) produce objects of the class `intsvy.mean`. The overloaded `plot` function produces a **ggplot2** based dotplot that resents calculated averages and their standard errors.

Optional arguments for the `plot.intsvy.mean()` are `sort` (should groups be sorted along the average or not) and `se` (should standard error be plotted or not).

The following example calculates and plots average numeracy performance in groups (by country / country and age group) based on the PIAAC dataset.

```
> pmeansNC <- piaac.mean.pv(pvlabel="NUM", by="CNTRYID", data=piaac,
             export=FALSE)
> plot(pmeansNC, sort=TRUE)

> pmeansNCA <- piaac.mean.pv(pvlabel="NUM", by=c("CNTRYID", "AGEG10LFS"),
              data=piaac, export=FALSE)
> plot(pmeansNCA, sort=TRUE)
```

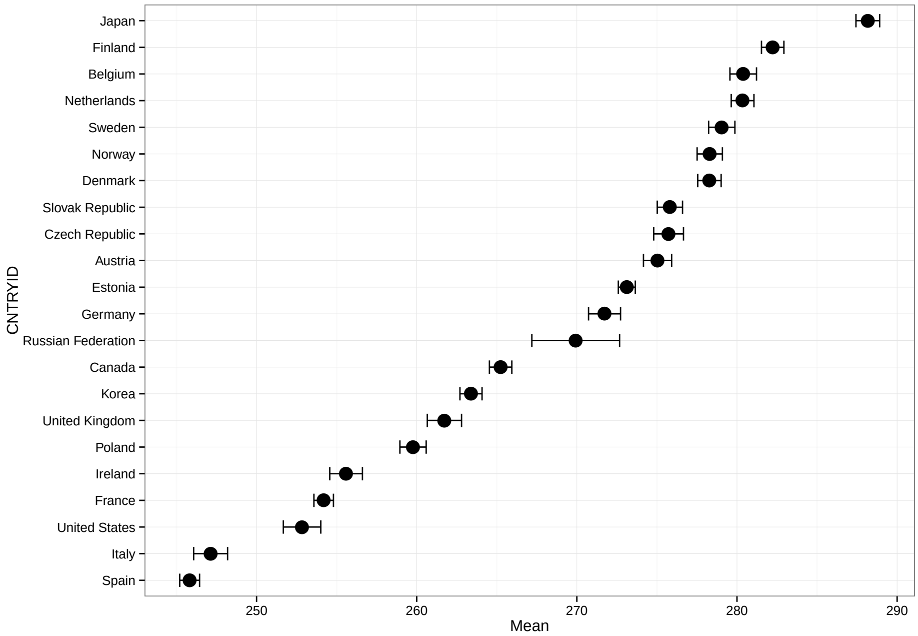

Figure 3: Graphical summary of averages and their standard errors. This example presents average numeracy scores and their standard errors for different countries based on the PIAAC dataset.

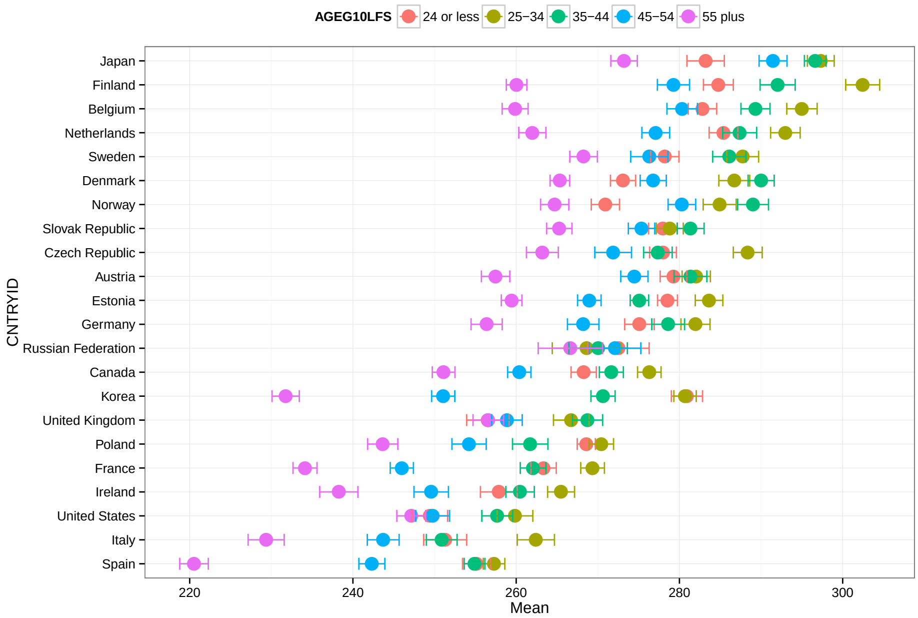

Figure 4: Graphical summary of averages in groups and their standard errors. This example presents average numeracy scores and their standard errors for different countries and age groups based on the PIAAC dataset.

Functions `*.reg.pv`, `*.reg` (where * stands for pisa, piaac, timms and pirls) produce objects of the class `intsvy.reg`. The overloaded `plot` function produces a **ggplot2** based dotplot that summarizes regression based model coefficients and their standard errors.

Optional arguments for the `plot.intsvy.reg()` are `sort` (should groups be sorted along the average or not) and `se` (should standard error be plotted or not).

The following example calculates and plots regression coefficients, intercepts and R^2 coefficients in groups defined by country based on the PIAAC dataset.

```
> rmodelLG <- piaac.reg.pv(pvlabel="LIT", x="GENDER_R", by = "CNTRYID",
data=piaac, export=FALSE)
> plot(rmodelLG, se=TRUE, sort=TRUE)
```

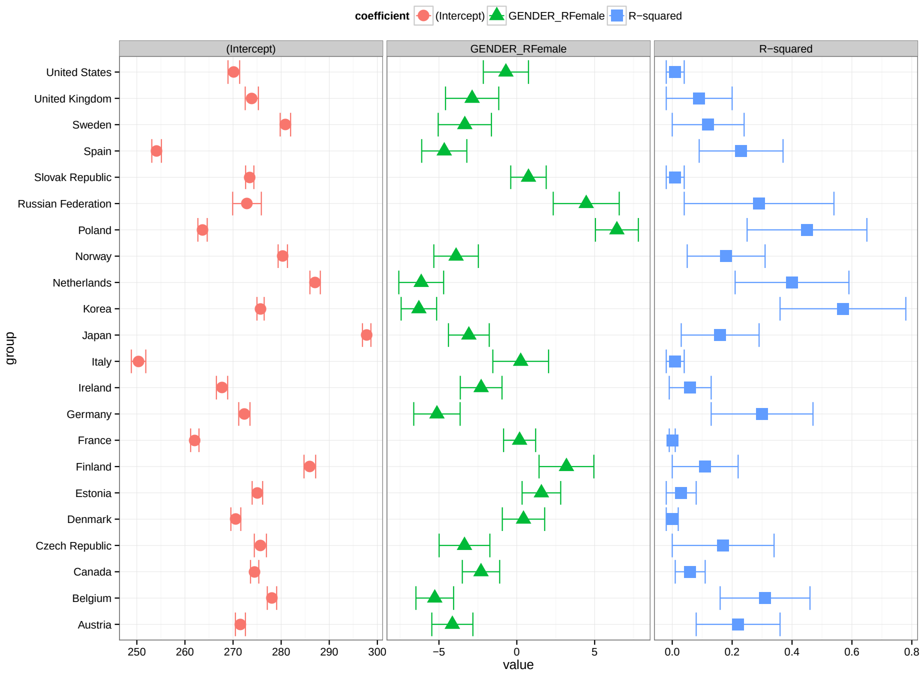

Figure 5: Graphical summary of regression models. This example presents outcomes for regression models with literacy scores as dependent variable and gender as independent variable. Panels present intercepts, gender coefficients and R-square coefficients based on the PIAAC dataset.

4. Summary

This article introduced **intsvy** and demonstrated its use with data from PISA, PIRLS, TIMSS, and PIAAC. **intsvy** provides another alternative within R to soundly handle data from international LSA and, to our knowledge, is the only available package for merging data from PIRLS and TIMSS. There are several limitations and plans for incorporating new features in future releases of this package. Currently **intsvy** can only deal with continuous data in regression analysis, handles missing data using listwise deletion, cannot analyse trend data from international LSA, cannot perform tests of statistical significance beyond those provided by regressions, cannot plot PIRLS and TIMSS data, to mention some limitations.

References

Foy P, Arora A, Stanco G (2013). *TIMSS 2011 User Guide for the International Database.* TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill, MA.

Foy P, Drucker K (2013). *PIRLS 2011 User Guide for the International Database.* TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill, MA.

Little R, Rubin D (1987). *Statistical Analysis with Missing Data.* John Wiley and Sons, New York.

OECD (2012). *PISA 2009 Technical Report.* OECD Publishing, Paris.

OECD (2013). *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (Volume II).* OECD Publishing, Paris. URL http://dx.doi.org/10.1787/9789264201132-en.

OECD (2014). *PISA 2012 Results: What Students Know and Can Do âĂŞ Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014).* OECD Publishing, Paris. URL http://dx.doi.org/10.1787/9789264201118-en.

von Davier M, Gonzalez E, Mislevy RJ (2009). "What are plausible values and why are they useful?" *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, **2**, 9–36.

Affiliation:

Daniel H. Caro
Oxford University Centre for Educational Assessment
Department of Education
University of Oxford
15 Norham Gardens

Oxford, OX2 6PY, United Kingdom
E-mail: daniel.caro@education.ox.ac.uk

Przemysław Biecek
Interdisciplinary Centre for Mathematical and Computational Modelling
University of Warsaw
Pawińskiego 5a, 02-106 Warsaw, Poland
E-mail: Przemyslaw.Biecek@gmail.com