Data Analysis of TIMSS dataset using `BIFIEsurvey` package in `R`

Mihaela Štiglic
Educational Research Institute Ljubljana, Toulouse School of Economics

Introduction

The `BIFIEsurvey` package in `R` contains tools for survey statistics for data sets with replication designs. It is particularly useful for analyzing data in educational assessments such as TIMSS, PISA or PIRLS. This poster presents the computation of descriptive statistics, linear regression and two-level hierarchical regression. It also demonstrates how to present data on a map.

Analysis of Data Set

1. Importing Data from SPSS

We set our working directory, `setwd("C:/Users/MS/T11")`, then import data to `R` by using library `foreign`. We use the TIMSS 2011 database for eigth-graders in Slovenia. We obtain a data frame called `df.timss`. It is recommended to set the `use.value.label` to `F`, to assure that variables are imported as numerical.

```
library(foreign)
df.timss <- read.spss('mydata.sav',  to.data.frame = T, use.value.label = F)
```

2. BIFIE Object

For our analysis we convert a data frame `df.timss` to a BIFIE object `bf.data`. There are several ways of doing this. We need to take into account that standard errors of statistics at TIMSS are computed using the *jackknife* method. We also need to specify the `pv_vars` vector, which contains the names of all variables that have five plausible values.

```
library(BIFIEsurvey)
pv_vars <- c("BSMMAT", "BSSSCI")
bf.data <- BIFIE.data.jack(data = df.timss, pv_vars = pv_vars, jktype = "JK_TIMSS")
```

3. Working with BIFIE Object

Descriptive Statistics

`BIFIE.univar` computes means and standard deviations for a given variable. It already takes into account sample weights. Statistics can be computed for specific groups such as regions, schools or classes. The example shows how to compute the mean and standard deviation for mathematics achievement across regions. The results are obtained by calling `res1$stat_M` and `res1$stat_SD`.

```
res1 <- BIFIE.univar(bf.data, vars = "BSMMAT", group = "REGION")
```

`BIFIE.by` computes statistics for user-defined functions. If we define a function `my.function`, `BIFIE.by` will compute the statistics for this function. The results are obtained by calling `res2$stat`. If `my.function` computes weighted mean, then we get the same result by using either `BIFIE.univar` or `BIFIE.by`.

```
res2 <- BIFIE.by(bf.data, vars = "BSMMAT", userfct = my.function, group = "REGION")
```

Drawing Maps

In order to draw a map in `R`, we first have to download a *shapefile* containing geographical data for Slovenia, which is available at *www.gadm.org*. Since we want to analyze performance in different regions, we select `SVN_adm1`. If we wanted to do the same for the counties, we would choose `SVN_adm2`.

```
library(rgdal)
geo <- readOGR("C:/Users/MS/T11/SVN_adm", "SVN_adm1", verbose=T, stringsAsFactors=F)
```

We build data frame `res.df` where the first column contains the mean mathematics achievements in different regions and the second one the associated region labels from `geo@data`. For instance, the Pomurska Region is labeled as number 1 in our database, but as number 8 in `geo@data`. We therefore need to define a vector `IDnew`, which associates regions with their labels in `geo@data`.

```
IDnew <- c(8, 7, 3, 9, 11, 10, 2, 6, 0, 4, 1, 5)
res.df <- data.frame(values = res2$stat$est, region = IDnew)
```

We also build data frame `geo.df`, which consists of geographical coordinates of the central points of the regions and the regions' names. This data frame will be used in order to display the names of the regions on the map.

```
geo.df <- data.frame(long = coordinates(geo)[, 1], lat = coordinates(geo)[, 2], region
  = geo@data$NAME_1)
```

Using library `ggplot2`, the geographical coordinates `k` from `geo` and the data frames `res.df` and `geo.df`, we plot a map. Each region is colored according to its mean mathematics achievement. The `expand_limits` and `coord_map` arguments are used to ensure the right proportion between the length and height of the map, while `geom_text` adds the names of the regions on the map.

```
library(ggplot2)
k <- map_data(geo)
map <- ggplot(res.df, aes(map_id=region)) + expand_limits(x=k$long, y=k$lat) +
geom_map(aes(fill=values), map=k, colour="black") + coord_map() +
geom_text(data=geo.df, aes(label=region, x=long, y=lat, group=region), size=3)
```

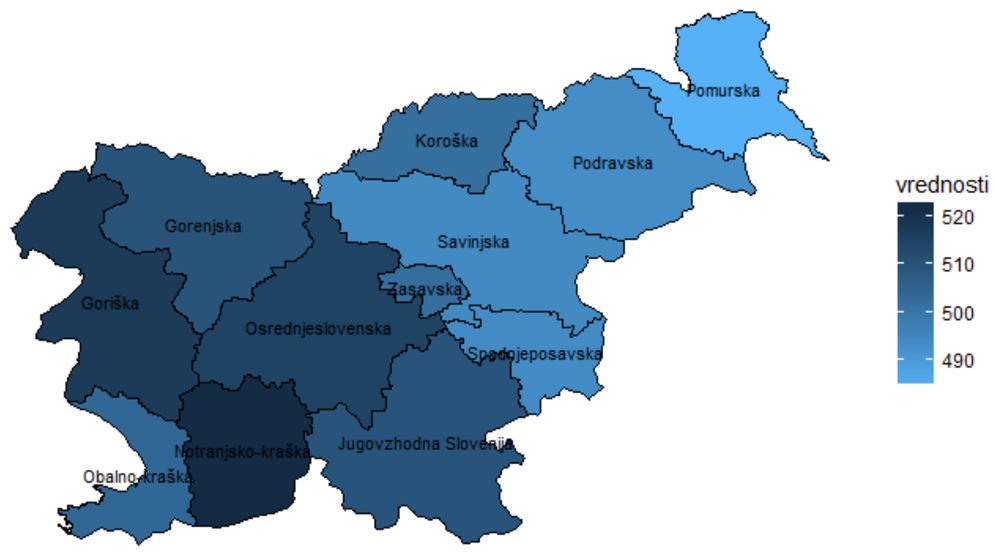

Figure 1: Average mathematics achievement of eighth-graders in different regions

Linear Regression

`BIFIE.linreg` computes linear regression for a model defined in `formula`. The example shows the regression of mathematics achievement on home resources for learning for boys and girls respectively. The `BSBGHER` variable is continuous. In case of a discrete variable, we would use `as.factor(BSDGHER)`.

```
mod.home <-  BIFIE.linreg(bf.data, formula = BSMMAT ~ BSBGHER, group = "ITSEX")
```

Hierarchical Linear Model

The variables observed at TIMSS could be related either to student's individual characteristics or to schools. Since our data set is nested (student $\rightarrow$ class $\rightarrow$ school), we can use a hierarchical linear model to identify which school variables influence student's performance. Let us assume a model where a socio-economic status X_{ij} of a student i in a school j has an impact on the student's achievement Y_{ij}. Due to a school variable W_j, which represents the number of computers at the school j, the impact of X_{ij} on Y_{ij} may differ from one school to another.

$$\text{Model 1:} \quad Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}, \quad r_{ij} \sim \mathcal{N}(0, \sigma^2)$$

$$\text{Model 2:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}, \quad \beta_{1j} = \gamma_{10} + u_{1j}, \quad u_{0j} \sim \mathcal{N}(0, \tau_{00}^2),\ u_{1j} \sim \mathcal{N}(0, \tau_{11}^2)$$

$$\Rightarrow Y_{ij} = \underbrace{\gamma_{00} + \gamma_{01} W_j + \gamma_{10} X_{ij}}_{\textit{fixed}} + \underbrace{u_{0j} + u_{1j} X_{ij} + r_{ij}}_{\textit{random}}$$

From the model we can identify a fixed and a random part. We take this into account when applying `BIFIE.twolevelreg`.

```
mod <- BIFIE.twolevelreg(BIFIEobj = bf.data, dep = "BSMMAT", formula.fixed = ~ W + X,
  formula.random = ~ X, idcluster="IDSCHOOL", wgtlevel2 = "TOTWGT" )
```

If the coefficient at W is statistically significant, we can say that the number of computers at school has an impact on how students perform. Moreover, the variance of Y_{ij} can be decomposed to the variance between and within groups, τ^2 and σ^2 respectively. Thus we consider the proportion of the variance accounted by group level by computing the intraclass coefficient $\rho_I = \frac{\tau^2}{\tau^2 + \sigma^2}$.

References

[1] BIFIE (2016). BIFIEsurvey: Tools for survey statistics in educational assessment. R package version 1.9.4-0.

[2] Raudenbush, S. W., Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, SAGE Publications.