Analiza podatkov raziskave TIMSS s paketom BIFIEsurvey v R-ju

Mihaela Štiglic
Pedagoški inštitut Ljubljana, Toulouse School of Economics

Paket BIFIEsurvey v R-ju predstavlja alternativo programu IDBAnalyzer in omogoča analizo podatkov mednarodnih raziskav v izobraževanju, kot so TIMSS, PISA in PIRLS. Poster predstavlja, kako v programu narediti izračun osnovnih statistik in iz podatkov narisati zemljevid. Prav tako je predstavljena uporaba linearne regresije in hierarhičnih linearnih modelov.

Uvod

Analiza podatkovne baze

1. Uvoz podatkovne baze iz SPSS

Najprej izberemo delovno mapo, na primer `setwd("C:/Users/MS/T11")`, in nato s pomočjo knjižnice `foreign`, ki omogoča branje podatkov iz SPSS-a, uvozimo podatke v R. Dobimo razpredelnico `df.timss`. Priporočeno je, da pri `use.value.label` uporabimo `false`, saj so tako spremenljivke obravnavane kot numerične.

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

2. BIFIE objekt

Za nadaljnje delo razpredelnico `df.timss` pretvorimo v BIFIE objekt `bf.data`. Obstaja več načinov, kako to narediti. Glede na to, da se standardne napake statistik pri TIMSS izračunavajo s pomočjo *jackknife* metode, to upoštevamo pri pretvorbi. Pred tem definiramo vektor `pv_vars` z imeni vseh spremenljivk, ki zavzamejo pet vrednosti (ang. *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. Delo z BIFIE objektom

Osnovne deskriptivne statistike

`BIFIE.univar` izračuna povprečje in standardni odklon za izbrane spremenljivke. Pri tem upošteva, da je vzorec utežen. Ukaz `group` omogoči izračun statistik po skupinah (regijah, šolah, razredih ...). Primer prikazuje izračun povprečnega dosežka in standardnega odklona pri matematiki po regijah. Rezultate pokličemo z ukazoma `res1$stat_M` in `res1$stat_SD`.

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

`BIFIE.by` izračuna statistike za poljubne funkcije, ki jih definiramo sami. Če smo definirali funkcijo `moja.funkcija`, potem `BIFIE.by` izračuna statistike za to funkcijo.

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

Rezultate pokličemo z ukazom `res2$stat`. Če je `moja.funkcija` povprečje, bosta `BIFIE.univar` in `BIFIE.by` izračunala enak rezultat.

Zemljevid

Za risanje zemljevida moramo najprej prenesti *shapefile* datoteko z geografskimi podatki Slovenije, dostopno na *www.gadm.org*. Podatke uvozimo v R s pomočjo knjižnice `rgdal`. Ker želimo podatke o regijah, izberemo SVN_adm1. Za podatke o občinah bi izbrali SVN_adm2.

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

Nato naredimo tabelo `res.df`, ki vsebuje povprečne matematične dosežke po regijah in pripadajoče oznake regij iz izvirnih geografskih podatkov `geo@data`. Pomurska regija ima na primer v naših podatkih oznako 1, v `geo@data` pa 8, zato definiramo vektor novih oznak regij `IDnovi`, ki regijam pripiše oznake iz izvirnih podatkov.

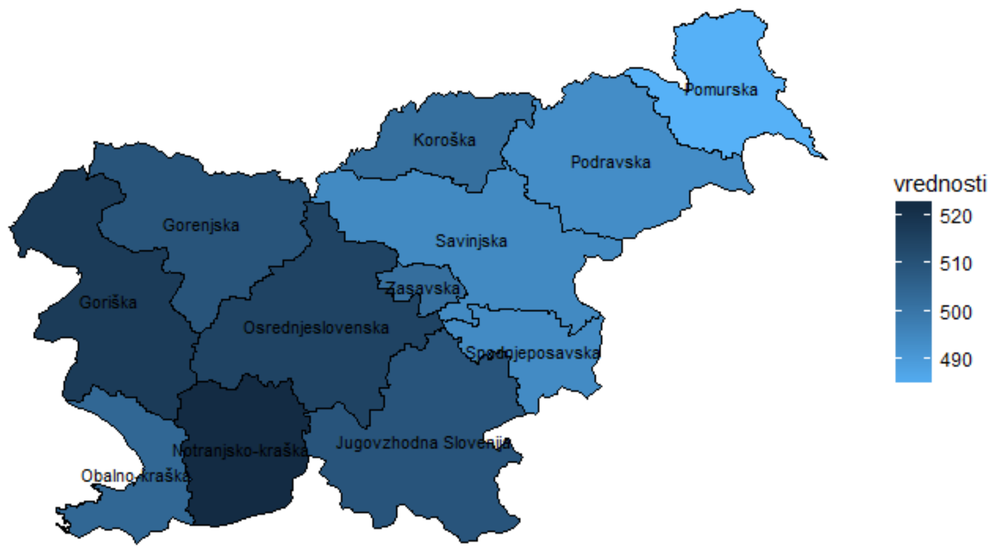
```
IDnovi <- c(8, 7, 3, 9, 11, 10, 2, 6, 0, 4, 1, 5)
res.df <- data.frame(values = res2$stat$est, region = IDnovi)
```

Naredimo tudi tabelo `geo.df`, ki vsebuje koordinate središč regij in njihova imena. To tabelo bomo uporabili za izpis imen regij na zemljevidu.

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

S pomočjo knjižnice `ggplot2`, tabel in geografskih koordinat `k` v podatkih `geo` narišemo zemljevid. Ukaza `expand_limits` in `coord_map` poskrbita za ustrezno razmerje med dolžino in širino, `geom_text` pa na zemljevid zapiše imena regij.

```
library(ggplot2)
k <- map_data(geo)
zemljevid <- 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)
```



Slika 1: Povprečni dosežki osmošolcev pri matematiki po regijah - TIMSS 2011

Linearna regresija

`BIFIE.linreg` izračuna linearno regresijo za izbrani model, ki ga definiramo v `formula`. Primer prikazuje regresijo dosežka na domačo podporo otroku pri izobraževanju glede na spol učenca. Spremenljivka `BSBGHER` je zvezna, v primeru diskretne spremenljivke bi uporabili `as.factor(BSDGHER)`.

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

Hierarhični linearni model

Na dosežek učenca pri TIMSS vplivajo tako šolske kot individualne spremenljivke. Ker so podatki ugnezdeni (učenec $\rightarrow$ razred $\rightarrow$ šola), lahko s pomočjo hierarhičnih linearnih modelov ugotovimo, katere šolske spremenljivke vplivajo na dosežke. Predpostavimo model, kjer na dosežek Y_{ij} otroka i v šoli j vpliva njegov socialno-ekonomski status X_{ij}, hkrati pa se zaradi šolske spremenljivke W_j, ki naj predstavlja število računalnikov na šoli, ta vpliv od šole do šole razlikuje.

$$\begin{aligned}
&\text{Model 1:} \quad Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}, \quad r_{ij} \sim \mathcal{N}(0, \sigma^2) \\
&\text{Model 2:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}, \quad \beta_{1j} = \gamma_{10} + u_{1j}, \quad u_{0j} \sim \mathcal{N}(0, \tau^2_{00}),\ u_{1j} \sim \mathcal{N}(0, \tau^2_{11}) \\
&\Rightarrow Y_{ij} = \underbrace{\gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{ij}}_{\textit{fixed}} + \underbrace{u_{0j} + u_{1j}X_{ij} + r_{ij}}_{\textit{random}}
\end{aligned}$$

Iz modela razberemo fiksni in slučajni del. To upoštevamo pri zapisu v `R`.

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

Če je koeficient pri W statistično značilen, potem lahko rečemo, da število računalnikov na šoli vpliva na dosežek pri TIMSS. Ker je varianca dosežka vsota variance učenčevih in šolskih spremenljivk, to je τ^2 in σ^2, nam o vplivu šole na dosežek dosti pove že delež variance dosežkov med šolami $\frac{\tau^2}{\tau^2+\sigma^2}$.

Literatura

[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.