Zadanie z planowania eksperymentu

dane apistrat

Zadanie

zrobić analizę średniej wyników testów w roku 2000 (api00) z uwzględnieniem faktu, że próba jest wylosowana warstwowo, gdzie warstwy stanowią typy szkoły (elementary, middle, high). Porównać wyniki ze średnią zmiennej api00 całej populacji (apipop) oraz ze średnią liczoną bez uwzględniania faktu, iż próba jest warstwowa (ale z tego samego zbioru czyli apistrat). Na koniec przeprowadzić estymację ilorazową wykorzystując silny związek pomiędzy api00 a api99. Ocenić obciążenie tych trzech estymatorów (zwyklej średniej, średniej z uwzględnieniem warstw i średniej ilorazowej) oraz ich błędy standardowe estymacji. Związek pomiędzy api00 a api99 przedstawić graficznie.

Rozwiązanie

Wczytanie zbioru danych apistrat

```
library(tidyverse)
library(survey)
data(api)
head(apistrat)
```

```
##              cds stype            name                                    sname
## 1 19647336097927     E Open Magnet: Ce Open Magnet: Center for Individual (Char
## 2 19647336016018     E Belvedere Eleme                     Belvedere Elementary
## 3 19648816021505     E Altadena Elemen                      Altadena Elementary
## 4 19647336019285     E Soto Street Ele                   Soto Street Elementary
## 5 56739406115430     E Walnut Canyon E                 Walnut Canyon Elementary
## 6 56726036084917     E Atherwood Eleme                     Atherwood Elementary
##   snum               dname dnum       cname cnum flag pcttest api00 api99
## 1 2077 Los Angeles Unified  401 Los Angeles   18   NA      99   840   816
## 2 1622 Los Angeles Unified  401 Los Angeles   18   NA     100   516   476
## 3 2236    Pasadena Unified  541 Los Angeles   18   NA      99   531   544
## 4 1921 Los Angeles Unified  401 Los Angeles   18   NA     100   501   457
## 5 6140    Moorpark Unified  460     Ventura   55   NA     100   720   659
## 6 6077 Simi Valley Unified  689     Ventura   55   NA     100   805   780
##   target growth sch.wide comp.imp both awards meals ell yr.rnd mobility acs.k3
## 1     NA     24      Yes       No   No     No    33  25     No       11     20
## 2     16     40      Yes      Yes  Yes    Yes    98  77    Yes       26     19
## 3     13    -13       No       No   No     No    64  23     No       17     20
## 4     17     44      Yes      Yes  Yes    Yes    83  63     No       13     17
## 5      7     61      Yes      Yes  Yes    Yes    26  17     No       31     20
## 6      1     25      Yes      Yes  Yes    Yes     7   0     No       12     19
##   acs.46 acs.core pct.resp not.hsg hsg some.col col.grad grad.sch avg.ed full
## 1     29       NA        0       0   0        0        0        0   3.32  100
## 2     28       NA        0       0   0        0        0        0   1.67   57
```

```
## 3     30       NA        0       0   0        0        0        0   2.34   81
## 4     30       NA        0       0   0        0        0        0   1.86   64
## 5     30       NA        0       0   0        0        0        0   3.17   90
## 6     29       NA        0       0   0        0        0        0   3.64   95
##   emer enroll api.stu    pw  fpc
## 1    0    276     241 44.21 4421
## 2   40    841     631 44.21 4421
## 3   26    441     415 44.21 4421
## 4   24    298     288 44.21 4421
## 5    7    354     319 44.21 4421
## 6    0    330     315 44.21 4421
```

Warstwy ze względu na typy szkoły

```
any(is.na(apistrat$stype))
```

```
## [1] FALSE
```

```
nrow(apipop)
```

```
## [1] 6194
```

```
summary(apipop$stype)
```

```
##    E    H    M
## 4421  755 1018
```

```
summary(apistrat$stype)
```

```
##   E   H   M
## 100  50  50
```

W populacji jest 6194 przypadków, zaś w próbie 200. Powyżej jest przedstawiona także liczebność szkół ze względu na jej typ.

Następnie tworzę trzy warstwy ze względu na typ szkoły.

```
elementary <- filter(apistrat, stype=="E")
middle <- filter(apistrat, stype=="M")
high <- filter(apistrat, stype=="H")
```

Analiza średniej wyników testów w roku 2000

```
sredniapop <- mean(apipop$api00)
sredniapop
```

```
## [1] 664.7126
```

Średnia wyników w całej populacji wynosi 664,7126.

a) z uwzględnieniem faktu, że próba jest wylosowana warstwowo Łączę zbiory wierszami, następnie obliczam wagi obserwacji z próby.

```
proba <- rbind.data.frame(elementary, middle, high)
wagi <- table(apipop$stype)/table(apistrat$stype)
wagi
```

```
##
##     E     H     M
## 44.21 15.10 20.36
```

```
proba$wagi <- wagi[proba$stype]
```

Ustawiam schemat losowania próby.

```
dstrat <- svydesign(ids=~1, strata = ~stype, weights = ~wagi, data = proba)
```

Liczę średnią.

```
svymean(~api00,dstrat)
```

```
##         mean     SE
## api00 662.29 9.5361
```

Średnia wynosi 662,29, zaś błąd standardowy estymacji w przybliżeniu 9,54.

```
dstrat2 <- svydesign(ids=~1,data = proba)
svymean(~api00,dstrat2)
```

b) bez uwzględniania faktu, iż próba jest warstwowa

```
##         mean    SE
## api00 652.82 8.554
```

Średnia wynosi 652,82, zaś błąd standardowy estymacji w przybliżeniu 8,55.

c) estymacja ilorazowa Podane jest, że występuje silny związek pomiędzy api00, a api99.

Zmienną towarzyszącą, silnie skorelowaną z api00 jest api99. Najpierw określam B, które jest stosunkiem total api00 do total api99. Następnie liczę średnią i błąd standardowy estymacji.

```
B <- svyratio(~api00,~api99,design=dstrat)
B
```

```
## Ratio estimator: svyratio.survey.design2(~api00, ~api99, design = dstrat)
## Ratios=
##          api99
## api00 1.052261
## SEs=
##              api99
## api00 0.003691607
```

```
x <- svymean(~api99,dstrat)

B$ratio*x[1]
```

```
##          api99
## api00 662.2874
```

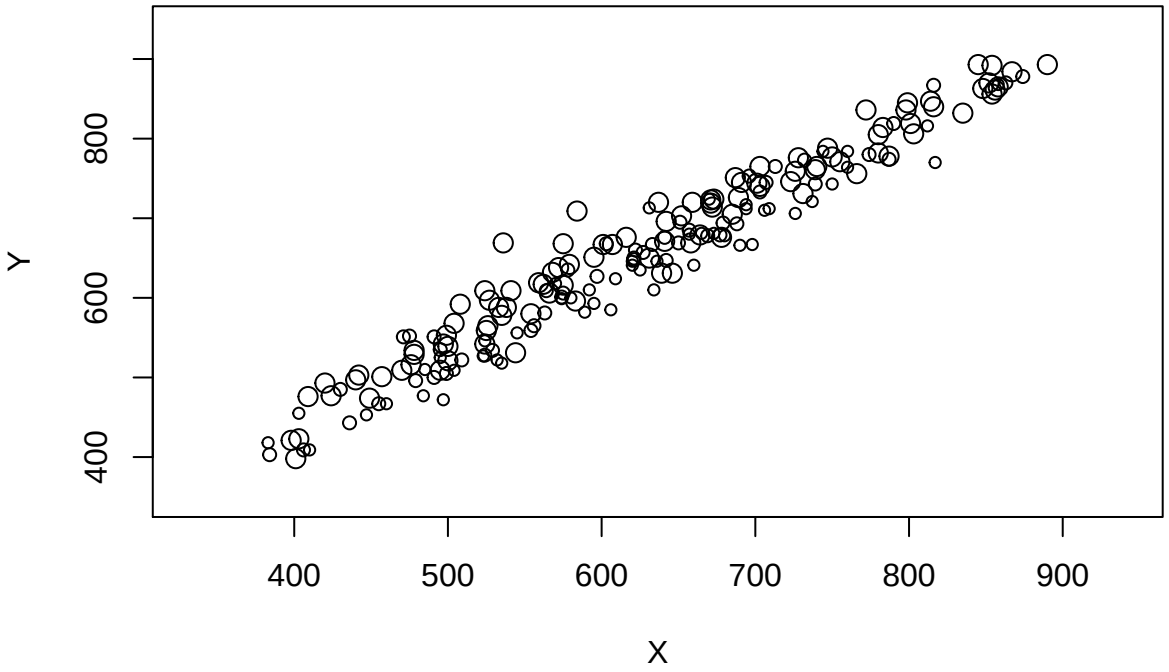
```
SE(B)*x[1]
```

```
## api00/api99
##    2.323479
```

Średnia wynosi 662,2874, zaś błąd standardowy estymacji w przybliżeniu 2,32.

Graficzny związek pomiędzy api00 a api99

```
svyplot(api00~api99,dstrat)
```



Wykres potwierdza założoną wcześniej wysoką korelację pomiędzy api00 i api99.

Podsumowanie

Wyniki:

- rzeczywisty (w całej populacji)
  - średnia 664,7126
- w próbie z podziałem na warstwy
  - średnia 662,29
  - SE 9,54
  - obciążenie estymatora 2,4226
- bez uwzględnienia podziału na warstwy
  - średnia 652,82
  - SE 8,55
  - obciążenie estymatora 11,8926
- estymator ilorazowy
  - średnia 662,2874
  - SE 2,32
  - obciążenie estymatora 2.4252

Patrząc na obciążenie estymatora stwierdzam, że estymator ilorazowy jest lepszy od pozostałych. Najbliżej prawdziwej wartości było oszacowanie za pomocą próby z podziałem na warstwy, ale estymator ilorazowy zapewniał prawie równie dobre przybliżenie wyniku. Ogólnie najlepiej wypada estymator ilorazowy.