ТЕОРИЈА УЗОРАКА час 9

30. април ’14.

КОЛИЧНИЧКО И РЕГРЕСИОНО ОЦЕЊИВАЊЕ

(*Ratio and Regression Estimation*)

- Ради се о двема техникама оцењивања у теорији узорака, код којих се користе допунске информације за "поправљање" прецизности оцена. При томе уводи се помоћно обележје x, које је у блиској вези (корелацији) са "главним" обележјем y, и за које се претпоставља да су његове вредности познате на свим јединицама у узорку. Мора бити познат и тотал X обележја x.
- Циљ примене баш ових техника је, дакле, смањење дисперзија оцена коришћењем корелације између y и x.

Количник обележја популације R једнак је $R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}$

$\hat{X} = \sum_{i \in s} X_i = N\hat{\bar{X}}$

- Оцена тотала:
  Ако су $\hat{Y}$ и $\hat{X}$ непристрасне оцене тотала Y и X главног и помоћног обележја, редом, количничка оцена тотала обележја популације је $\hat{Y}_R = \frac{\hat{Y}}{\hat{X}} X \left(= \frac{\hat{\bar{Y}}}{\hat{\bar{X}}} X \right)$

  Ова оцена није непристрасна, али се може показати да је пристрасност оцене обрнуто пропорционална обиму узорка, па је она занемарљиво мала за (довољно) велике узорке.
  Апроксимација средње квадратне грешке оцене $\hat{Y}_R$ је $MSE[\hat{Y}_R] \approx V[\hat{Y}] + R^2 V[\hat{X}] - 2R\mathrm{cov}(\hat{X}, \hat{Y})$.

- Код СУ без понављања:

  Оцена количника је $\hat{R} = \frac{\sum_{i \in s} Y_i}{\sum_{i \in s} X_i}$;

  оцена тотала је $\hat{Y} = \hat{R}X$; оцена средине је $\hat{\bar{Y}} = \hat{R}\bar{X}$;

  оцена средње квадратне грешке је

  $$\frac{N^2(N-n)}{Nn(n-1)} \sum_{i \in s} (Y_i - \hat{R}X_i)^2 .$$

  Оцене непознатих параметара нису непристрасне; из израза за оцену средње квадратне грешке види се да је она мања што је веза између главног и помоћног обележја "јача", у смислу да вредности Y_i мање одступају од регресионе праве, која пролази кроз координатни почетак, тако да је Y_i "грубо" пропорционално X_i. Тада је, уједно, и пристрасност приближно једнака нули.

Услов при коме је количничка оцена ефикаснија од обичне (директне) оцене, и, стога, пожељнија у датој ситуацији:

$$\frac{1}{2}R\frac{S_x}{S_y} = \frac{1}{2}\frac{CV_x}{CV_y} < \rho$$

где је $CV_x = \frac{S_x}{\bar{X}}$, односно $CV_y = \frac{S_y}{\bar{Y}}$, коефицијент варијације помоћног, односно главног обележја, а ρ је коефицијент корелације ова два обележја $\rho = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{(N-1)S_xS_y} = \frac{S_{xy}}{S_xS_y}$

- Примери код којих би коришћење количничке оцене могло имати бенефите:
  - x – обим дрвета; y – запремина дрвета
  - x – број страница у једном броју часописа; y – број страница на којима се налази бар једна реклама
  - x – број запослених у фирми; y – приход фирме
  - ...

- Количничко оцењивање, дакле, може дати врло квалитетне резултате у случају када се веза између главног и помоћног обележја најбоље може описати (регресионом) правом која пролази кроз координатни почетак. Регресионо оцењивање је слична техника пригодна у ситуацијама када су подаци уједначено "разбацани" око праве која не пролази кроз координатни почетак.
- Заправо, ради се о моделу линеарне зависности између главног и помоћног обележја: $y = a + bx$.
- Оцена (методом најмањих квадрата) коефицијента правца регресионе праве (*slope*) је $\hat{b} = \frac{s_{xy}}{s_x^2}$, а одсечка ординате (*intercept*) $\hat{a} = \hat{\bar{Y}} - \hat{b}\hat{\bar{X}}$.

$$\hat{X} = \sum_{i \in s} X_i = N\hat{\bar{X}}$$

Претпостављајући да се бира узорак без понављања, вредности главног обележја оних јединица које нису одабране у узорак могу се (на основу модела линеарне регресије) оценити са $\hat{Y}_i = \hat{a} + \hat{b}X_i = \hat{\bar{Y}} - \hat{b}\hat{\bar{X}} + \hat{b}X_i \ , i \in \bar{s} = S \backslash s$

- Оцена тотала:

  Регресиона оцена тотала обележја популације је $\hat{Y}_{LR} = N\hat{\bar{Y}} + N\hat{b}(\bar{X} - \hat{\bar{X}})$.

  Ова оцена није непристрасна.

- Код СУ без понављања:

  Оцена средње квадратне грешке је

$$\frac{N^2(N-n)}{Nn(n-2)} \sum_{i \in s} (Y_i - \hat{a} - \hat{b}X_i)^2 .$$

- Поређење квалитета оцена − обичне, количничке и регресионе код СУ без понављања:

  Уведу се ознаке за дисперзије, односно средње квадратне грешке оцена − V_{srs}, MSE_R, MSE_{LR} .

  Важи: $V_{srs} > MSE_{LR}$; $MSE_R > MSE_{LR}$.
- Код стратификованог узорка:

  Количничка оцена може се формирати на два различита начина:
  - посебна оцена (*separate ratio estimator*) $\hat{Y}_{Rs} = \sum_{h=1}^{L} \frac{\hat{Y}_h}{\hat{X}_h} X_h$
  - комбинована оцена (*combined ratio estimator*) $\hat{Y}_{Rc} = \frac{\sum_{h=1}^{L} \hat{Y}_h}{\sum_{h=1}^{L} \hat{X}_h} X$

  где су $\hat{Y}_h$ и $\hat{X}_h$, $h = 1, 2, \ldots, L$, непристрасне оцене тотала главног и помоћног обележја за h -ти стратум.

Регресиона оцена може се формирати на два различита начина:

- посебна оцена $\hat{Y}_{LRs} = \sum_{h=1}^{L} \left(N_h \hat{\bar{Y}}_h + N_h \hat{b}_h \left(\bar{X}_h - \hat{\bar{X}}_h \right) \right)$
- комбинована оцена $\hat{Y}_{LRc} = N \hat{\bar{Y}}_{st} + N \hat{b} (\bar{X} - \hat{\bar{X}}_{st})$

где су $\hat{\bar{Y}}_h$ и $\hat{\bar{X}}_h$, $h = 1, 2, \ldots, L$, непристрасне оцене средине главног и помоћног обележја за h-ти стратум.

Посебна регресиона оцена примењује се у ситуацијама када се може сматрати да се прави регресиони коефицијенти b_h разликују по стратумима.

При рачунању математичких очекивања (тј. одређивању пристрасности) и средње квадратних грешака оцена средине и тотала обележја требало би разликовати два случаја:

- када су $\hat{b}_h$, односно $\hat{b}$, унапред изабране константе
- када се $\hat{b}_h$, односно $\hat{b}$, израчунавају на основу узорка

- Пакет `sampling` – корисне функције:
  - `ratioest()`
  - `regest()`
  - `ratioest_strata()`
  - `regest_strata()`
- Пример - API популација
  - оцењивање количника

У API популацији налази се променљива `api.stu` која чува број ученика тестираних ради одређивања индекса академског успеха, по школама. Циљ: одредити проценат ученика који су решавали поменуте тестове широм целе државе Калифорнија. Функција `svyratio()` даје оцену количника популацијских тотала.

```
> dstrat <- svydesign(id=~1, strata=~stype, fpc=~fpc, data=apistrat)
> svyratio(~api.stu, ~enroll, dstrat)
Ratio estimator: svyratio.survey.design2(~api.stu, ~enroll, dstrat)
Ratios=
           enroll
api.stu 0.8369569
SEs=
             enroll
api.stu 0.007757103
```

- количници за оцењивање унутар потпопулација
- количничко оцењивање тотала обележја

Дакле, резултати са претходног слајда показују да је, на основу стратификованог узорка обима 200 школа, оцењено да је $83.7\% \pm 0.78\%$ броја уписаних ученика било тестирано у овом истраживању. С обзиром да је познат и тачан број уписаних ученика у школама широм Калифорније (укупно: 3811472), циљ је: оценити укупан број ученика који су решавали тестове. Очигледна оцена била би: $0.837 \cdot 3811472 \approx 3190202$, са стандардном грешком $0.0078 \cdot 3811472 \approx 29730$.

HT оцена укупног броја ученика који су решавали тестове је 3086009 са стандардном грешком (приближно) 99478.

```
> library(sampling)
> HTstrata(apistrat$api.stu, pik=1/apistrat$pw, strata=apistrat$stype)
[1] 3086009
> #kodovi za izracunavanje YG ocene disperzije HT ocene totala obelezja
'api.stu' dati su u dodatnom .r fajlu "cas9.r"
> #...
> (se <- sqrt(disp_strat1+disp_strat2+disp_strat3))
[1] 99477.38
```

Метод `predict()` може се користити за даља израчунавања базирана на излазу функције `svyratio()`.

```
> r <- svyratio(~api.stu, ~enroll, dstrat)
> predict(r, total=3811472)
$total
          enroll
api.stu 3190038

$se
           enroll
api.stu 29565.98
```

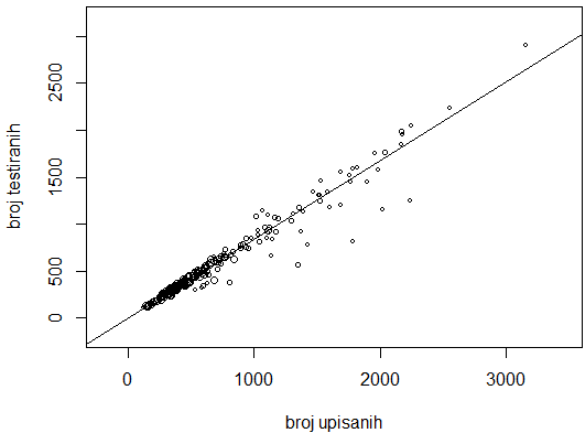
Коришћење количника у овој ситуацији мотивисано је идејом да је број ученика који су решавали тест “грубо” пропорционалан броју ученика у школи, тј. моделом

$$api.stu = \alpha \times enroll + \varepsilon$$

где ε има математичко очекивање приближно једнако нули, а α је поменути количник.

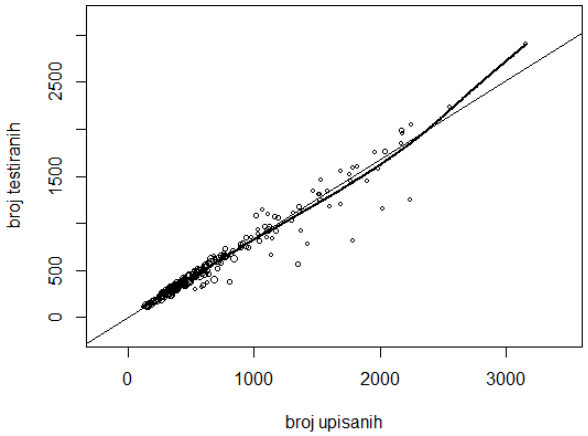
Графички приказ података показује оправданост избора овог модела као доброг за расподелу података.

```
> svyplot(api.stu~enroll, design=dstrat, style="bubble", xlab="broj upisanih",
          ylab="broj testiranih")
> abline(a=0, b=coef(svyratio(~api.stu, ~enroll, dstrat)))
```



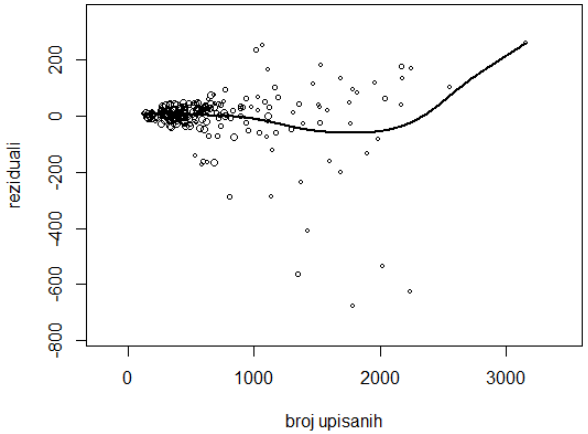
На горњем графику уцртана је права линија која одговара моделу, а следећим кодом додаје се “глатка” крива оцењена функцијом `svysmooth()`.

```
> lines(svysmooth(api.stu~enroll, dstrat), lwd=2)
Loading required package: KernSmooth
KernSmooth 2.23 loaded
Copyright M. P. Wand 1997-2009
```



Када је реч о резидуалима, они се израчунавају по формули $resid = api.stu - 0.837 \text{ x } enroll$.

```
> resid <- apistrat$api.stu-coef(svyratio(~api.stu, ~enroll, dstrat))*apistrat$enroll
> svyplot(resid~enroll, design=dstrat, style="bubble", xlab="broj upisanih",
        ylab="reziduali")
> lines(svysmooth(resid~enroll, dstrat), lwd=2)
```



Са графика делује да варијабилност резидуала расте са повећањем величине школе. Постоји наговештај да је количник систематски мањи код школа које броје између 1000 и 2000 ученика тако да модел не "фитује" у потпуности перфектно.

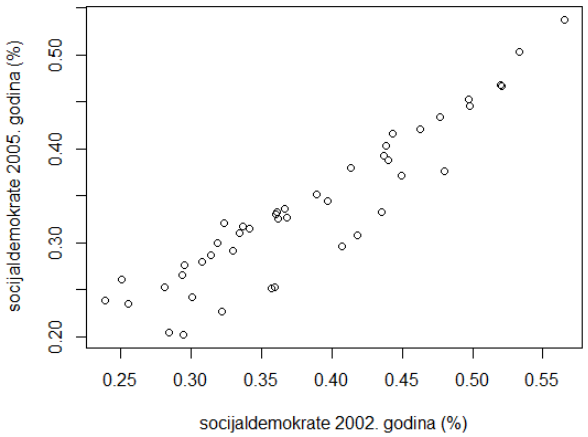
Код великих узорака посебна количничка оцена биће прецизнија него оцена која се изводи на основу стратификованог узорка узетог као целина, јер користи више информација о популацији (претпоставља се познавање тотала помоћног обележја X_h за сваки стратум) и јер већи број оцењених параметара омогућава одабраном моделу да боље одговара подацима.

```
> sep <- svyratio(~api.stu, ~enroll, dstrat, separate=T)
> com <- svyratio(~api.stu, ~enroll, dstrat)
> # poznati su totali pomocnog obelezja po stratumima
> strat.totals <- list(E=1877350, H=1013824, M=920298)
> predict(sep, total=strat.totals)
$total
          enroll
api.stu 3190022

$se
          enroll
api.stu 29756.44

> predict(com, total=3811472)
$total
          enroll
api.stu 3190038

$se
          enroll
api.stu 29565.98

> svyby(~api.stu, ~stype, dstrat, denom=~enroll, svyratio)
  stype api.stu/enroll se.api.stu/enroll
E     E      0.8518163        0.00703236
H     H      0.8105702        0.02047726
M     M      0.8356958        0.01818744
```

- # Пример за регресионо оцењивање

  Ради се са базом података `election` смештеном у пакету `samplingbook`. Она садржи податке о укупном броју грађана са правом гласа у 16 немачких федералних држава и резултатима избора за немачки Бундестаг 2002. и 2005. године. Функција `mbes()` из истог пакета коришћена је за оцењивање (базирано на жељеном моделу) средње вредности обележја од интереса.

```
> data(election)
> N <- nrow(election)
> set.seed(300314)
> sample <- election[sort(sample(1:N, size=45)),]
> plot(SPD_05 ~ SPD_02, data=sample, xlab="socijaldemokrate 2002. godina (%)",
      ylab="socijaldemokrate 2005. godina (%)")
```



```
> X.mean <- mean(election$SPD_02)
> mbes(SPD_05 ~ SPD_02, data=sample, aux=X.mean, N=N, method="regr")

mbes object: Model Based Estimation of Population Mean
Population size N = 299, sample size n = 45

Values for auxiliary variable:
X.mean.1 = 0.3861, x.mean.1 = 0.3816
------------------------------------------------------------------
Linear Regression Estimate

Mean estimate:  0.3388
Standard error:  0.0042

95% confidence interval [0.3305,0.3471]

------------------------------------------------------------------
Linear Regression Model:
Call:
lm(formula = formula, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.061920 -0.004367  0.011318  0.016667  0.044409

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01082    0.02177  -0.497    0.622
SPD_02       0.90539    0.05575  16.239   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03085 on 43 degrees of freedom
Multiple R-squared:  0.8598,   Adjusted R-squared:  0.8565
F-statistic: 263.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

Тачна вредност процента гласова који су освојили социјалдемократе 2005. године је 34.27%.

Што се тиче функција из пакета `survey` – функција `svyglm()` "фитује" линеарне и генералне линеарне моделе подацима који су мештени у *survey design object*-у.

```
> sample <- cbind(sample, fpc=rep(N, 45))
> srsdesign <- svydesign(id=~1, data=sample, fpc=~fpc)
> (socdemo05 <- svyglm(SPD_05 ~ SPD_02, srsdesign))
Independent Sampling design
svydesign(id = ~1, data = sample, fpc = ~fpc)

Call:  svyglm(formula = SPD_05 ~ SPD_02, srsdesign)

Coefficients:
(Intercept)       SPD_02
   -0.01082      0.90539

Degrees of Freedom: 44 Total (i.e. Null);  43 Residual
Null Deviance:     0.2918
Residual Deviance: 0.04091     AIC: -181.4
```

Пример се може проширити, у смислу да се размотри утицај још неких променљивих на модел. Наведен је и код у коме је искоришћен и резултат "зелених" на изборима 2002.

```
> X.mean2 <- c(mean(election$SPD_02), mean(election$GREEN_02))
> #isprobati: mbes(SPD_05 ~ SPD_02 + GREEN_02, data=sample, aux=X.mean2, N=N,
                  method="regr")
```

На врло једноставан начин може се вршити и предвиђање.

```
> (socdemo05_i_zel <- svyglm(SPD_05 ~ SPD_02 + GREEN_02, srsdesign))
Independent Sampling design
svydesign(id = ~1, data = sample, fpc = ~fpc)

Call:  svyglm(formula = SPD_05 ~ SPD_02 + GREEN_02, srsdesign)

Coefficients:
(Intercept)       SPD_02     GREEN_02
   -0.04502      0.91171      0.39245

Degrees of Freedom: 44 Total (i.e. Null);  42 Residual
Null Deviance:      0.2918
Residual Deviance: 0.02899     AIC: -194.9
> data2002 <- data.frame(SPD_02=X.mean2[1], GREEN_02=X.mean2[2])
> predict(socdemo05_i_zel, type="response", newdata=data2002)
  response     SE
1  0.34032 0.0035
```

СИСТЕМАТСКИ УЗОРАК
(*Systematic Sampling*)

- Постоји колекција планова узорковања под заједничким именом систематски / периодични узорак. Они имају неколико предности приликом примене у пракси.
- У најкраћем, код систематског узорка, уместо избора n јединица из популације на случајан начин, одлучивање о јединицама које ће се наћи у узорку врши на основу избора само једног случајног броја.

- *Linear Systematic Sampling*
  Идеја се састоји у следећем:
  Претпостави се да се популација састоји од N јединица и да су оне означене бројевима од 1 до N, по неком реду, а да треба одабрати узорак обима n. Јединице се, затим, у узорак бирају периодично − у једнаким размацима. Наиме, ако је интервал (период, корак) избора k, од првих k јединица бира се (најчешће на случајан начин) једна јединица као почетна, нека је то нпр. јединица r, а потом свака k-та јединица.
  Тако се формира узорак од јединица означених са $r, r+k, \ldots, r+(n-1)k$.
  Корак избора зависи, очигледно, од обима популације N и од обима узорка n, тако да се обично узима да је $k = \frac{N}{n} = \frac{1}{f}$.

Јасно је да се систематски узорак може замислити и као стратификован узорак, код кога је популација подељена на n стратума, од којих се сваки састоји од по k јединица. При томе, из сваког стратума се у узорак бира по једна јединица.

Стога се може очекивати да систематски узорак пружа исту прецизност као и стратификовани случајни узорак са једном јединицом у сваком стратуму.

Ипак, систематски узорак може, у неким ситуацијама, бити и лошији од стратификованог, јер се код њега стратуми формирају без разматрања интерне хомогености јединица.

Извесне тешкоће могу се јавити када N није умножак од n, тј. $N \neq k \cdot n$, али се у конкретној ситуацији могу направити мања одступања од задатог обима узорка.

Предности у односу на СУ без понављања:

- правило одабира јединица у узорак је сасвим једноставно, не захтева генерисање (псеудо)случајних бројева велики број пута нити потпуну нумерацију јединица у популацији
- број могућих узорака је знатно мањи и до њих се, по правилу, једноставније долази
- интуитивно је прихватљив – "равномерно" је распоређен на целој популацији, не допушта случајна груписања или пропуштање заступљености неких делова популације
- у извесним ситуацијама, које се ипак често сусрећу у пракси, има мању стандардну грешку

Мане:

- прикривена периодичност или трендови у популацији могу резултирати оценама са великом пристрасношћу
- присутан је недостатак непристрасних оцена дисперзија оцена праметара; заправо, дисперзија се може апроксимирати ако су на располагању информације о структури популације

- Оцена тотала:

  Непристрасна оцена тотала Y обележја популације код систематског узорка, са случајно одабраном почетном јединицом r је $\hat{Y}_{lss} = \frac{N}{n}\sum_{j=1}^{n} Y_{r+(j-1)k}$

  а њена дисперзија је $V(\hat{Y}_{lss}) = \frac{1}{k}\sum_{r=1}^{k}[\hat{Y}_r - Y]^2$, где је $\hat{Y}_r$ вредност $\hat{Y}_{lss}$, која одговара r-том узорку (тј. узорку који би као почетну имао јединицу r).

  Могућност одређивања дисперзије, код раније приказаних планова узорковања заснивала се на случајном избору јединица популације, што овде није случај.

Ако се претпостави да је списак јединица популације на основу кога се бира узорак уређен на случајан начин (а могућа је и ситуација да такав списак и не постоји), или да је обележје, које се посматра, независно од ознака јединица у популацији, систематски узорак може се поистоветити са СУ без понављања и могу се применити истоветне формуле за оцењивање дисперзије.

У пракси се најчешће тако и ради, понекад без довољно основа, што може довести до стицања потпуно погрешне слике о обележју од интереса (нпр. када популација има периодичан тренд).

“Systematic sampling performs badly when the list is ordered in cycles of values of the survey variables and when the sampling interval coincides with a multiple of the length of the cycle.”