R을 이용한 사회조사 자료의 분석 및 보고서 작성 방법

Jinseog Kim
Dongguk University
jinseog.kim@gmail.com

2017-10-28

여기서 다룰 문제

- 상황
  - 소규모 데이터
  - 단순(?)연산
  - 독립적 반복 연산
  - 정형화된 보고서의 작성
- 제약조건: 시스템화된 도구는 없다. R을 이용하자.
- 예제: 주기적이고 정형화된 설문조사에서 결과표의 생성
- 발표의 목적: R을 활용한 사회조사 분석에서의 경험을 공유

지자체 사회조사

- 통계법 제18조 및 시행령 제24조에 의해 승인된 일반통계
- 조사목적:
  - 지역사회 구성원의 사회적 관심, 삶의 질등을 파악
  - 지역균형개발, 복지시책추진 등 행정시책의 기초자료로 활용
- 조사주기: 매년
- 연도별 조사영역

조사 영역	2014년	2015년	2016년
가구·가족	✓		✓
교육	✓		✓
문화·여가	✓		✓
안전	✓		✓
환경	✓		✓
정보·통신	✓		
고용	✓		
노동		✓	
보건		✓	
사회통합		✓	
소득 소비		✓	
주거와 교통		✓	

사회조사 표본자료 현황

시군	2014		2015		2016	
	가구	가구원	가구	가구원	가구	가구원
경북도	19,980	37,341	22,714	41,758	22,724	41,342
포항시	1,935	3,880	1,939	3,882	1,939	3,889
경주시	1,494	2,879	1,500	2,810	1,497	2,806
김천시	1,060	2,100	1,060	2,088	1,060	2,024
안동시	1,260	2,319	1,258	2,272	1,260	2,272
구미시	1,670	3,203	1,667	3,239	1,674	3,347
영주시	1,000	1,943	1,001	1,876	1,000	1,888
			...			
군위군	580	1,043	801	1,406	800	1,375
의성군	800	1,395	801	1,374	800	1,316
청송군	600	1,044	800	1,358	800	1,358
영양군	297	511	800	1,329	801	1,337
			...			
봉화군	440	783	800	1,454	799	1,298
울진군	519	876	800	1,358	799	1,332
울릉군	292	504	392	656	400	619

문항의 종류

- 공통문항: 기초단체 공통 문항
- 특성문항: 기초단체별 특성화된 문항
- 연도별 영역별 공통문항의 수

영역	2014년		2015년		2016년	
	문항수	세부문항수	문항수	세부문항수	문항수	세부문항수
가구(주/원)관련	4	5	3	4	3	10
가구·가족	3	11			3	12
교육	3	13			5	46
문화·여가	5	26			7	30
안전	10	36			10	47
환경	3	17			2	12
고용	4	9				
정보·통신	7	15				
노동			2	7		
보건			1	3		
사회통합			13	63		
소득 소비			4	7		
주거와 교통			5	14		
합계	39	132	28	98	30	157

시군별 특성 문항

시군	2014년	2015년	2016년
포항시	8	9	9
경주시	6	6	6
김천시	6	6	6
안동시	5	5	5
	...		
문경시	30	30	30
경산시	10	7	10
군위군	6	6	6
의성군	8	8	8
청송군	5	5	8
영양군	7	7	7
	...		
울진군	8	5	15
울릉군	11	11	11
시군 특성 항목 전체	224	214	229

추계범위

- 공통항목: 도 전체 및 시군별, 분류변수별
  - 분류변수: 성별, 연령, 소득수준, …
- 특성항목: 시군별, 분류변수별

결과 테이블의 수 및 계산속도

연도	공통 세부문항수	지역수	특성 세부문항수	합
2016	157	24	229	3,997
2015	98	24	214	2,566
2014	132	24	224	3,392
전체				9,955

- Serial computing time: 약 1~2H
- 디버깅도 해야하는데 ㅠㅠ

사회조사 결과표의 형태

- 연도별/문항별/전체/기초단체별/분류별 추계

5.7 해외여행 여부 및 목적별 횟수

구분		응답수	여부	응답수	관광목적	가사목적	업무목적
전체		41,294	13.98 (3.34)	5,054	88.91 (0.82)	8.43 (9.19)	7.65 (6.91)
시군	포항시	3,885	17.90 (7.90)	690	90.84 (1.37)	7.12 (17.11)	7.84 (13.27)
	경주시	2,800	16.69 (9.07)	450	89.50 (2.31)	9.58 (28.45)	6.79 (21.76)
	김천시	2,023	11.85 (8.84)	221	91.91 (2.07)	5.01 (36.24)	6.68 (26.01)
	안동시	2,272	10.86 (11.02)	231	89.36 (3.70)	9.15 (36.61)	6.59 (28.98)
	구미시	3,341	16.39 (10.37)	516	87.26 (2.44)	7.66 (28.02)	12.72 (14.64)
	영주시	1,887	12.28 (12.79)	216	91.04 (3.42)	6.16 (30.10)	4.37 (43.34)
	영천시	1,731	13.74 (9.65)	237	91.08 (2.74)	7.69 (28.79)	3.84 (36.34)
	상주시	2,032	13.17 (13.31)	240	76.57 (8.64)	21.66 (40.64)	5.06 (47.45)
	문경시	1,758	11.93 (11.84)	205	89.62 (2.41)	5.89 (36.07)	4.32 (30.09)
	경산시	2,404	10.98 (9.26)	261	86.10 (4.07)	10.42 (32.60)	4.34 (31.87)
	군위군	1,375	10.20 (15.47)	140	86.94 (4.15)	13.47 (46.75)	4.65 (50.36)
	의성군	1,314	10.56 (19.42)	125	90.06 (3.16)	5.84 (47.07)	5.94 (38.37)
	청송군	1,356	8.96 (11.90)	125	91.27 (3.32)	10.79 (30.90)	4.83 (39.33)
	영양군	1,337	12.60 (11.07)	169	88.03 (3.81)	10.12 (40.83)	6.26 (37.66)
	영덕군	1,382	11.42 (14.77)	160	95.01 (1.86)	4.01 (41.21)	3.26 (56.45)
	청도군	1,415	9.44 (13.67)	139	90.53 (3.83)	8.72 (34.01)	4.93 (43.99)
	고령군	1,428	9.06 (15.02)	129	90.22 (4.93)	5.74 (44.70)	4.22 (48.85)
	성주군	1,381	8.67 (15.20)	114	87.23 (5.82)	11.68 (41.35)	2.18 (60.37)
	칠곡군	1,566	11.81 (13.67)	184	87.57 (3.26)	5.14 (38.18)	15.40 (22.75)
	예천군	1,364	10.78 (15.40)	135	91.31 (3.02)	9.91 (31.19)	0.57 (100.20)
	봉화군	1,296	10.44 (13.84)	137	82.86 (5.42)	13.49 (31.83)	5.96 (53.57)
	울진군	1,328	10.36 (12.45)	141	94.59 (1.89)	12.17 (37.46)	3.60 (37.13)
	울릉군	619	14.15 (14.30)	89	85.07 (5.59)	14.61 (26.78)	17.76 (30.15)
성별	남	18,941	14.54 (3.61)	2,484	86.49 (1.07)	7.40 (11.28)	12.06 (7.63)
	여	22,345	13.49 (3.45)	2,570	91.20 (0.87)	9.40 (9.46)	3.47 (13.25)
연령	29세이하	3,598	13.86 (6.30)	473	86.65 (2.12)	7.66 (19.92)	6.11 (20.76)
	30-39세	3,464	16.89 (6.17)	539	85.88 (1.98)	9.19 (20.92)	13.71 (13.50)
	40-49세	5,471	15.98 (5.05)	809	83.52 (1.93)	9.44 (14.16)	14.56 (10.08)
	50-59세	7,661	19.59 (4.15)	1,422	91.62 (1.07)	7.60 (13.24)	6.67 (12.98)
	60세이상	21,095	9.93 (4.41)	1,811	92.08 (1.24)	8.51 (15.91)	2.10 (20.54)

추계 프로그램 설계시 고려사항

- 다양한 문항 유형
  - 문항의 응답유형 분류
  - 유형에 따른 함수 구성
- 추계단위별 많은 수의 조사문항에 대한 결과물 생성
  - 추계단위: 전체 및 23개 기초단체별
  - 동일한 연산의 반복, 오류 발견시 디버깅 ⇒ 계산 속도가 문제
  - ⇒ 병렬처리로 수행속도 개선
- 문서화된 결과표의 편집
  - 별도의 편집없이 자동 생성

문항의 응답 유형 분류

- 응답 특성에 따른 분류
  1. 다른 문항의 응답과의 관련 여부: 단순/조건부 문항
  2. 응답값의 속성: 범주형/연속형/정수형/이진형
  3. 중복응답 여부: 단일응답/중복응답
- 실제로는 아래 5가지 유형으로 구분됨
  1. 단순 단일응답
  2. 조건부 단일응답
  3. 단순 다중(복수)응답
  4. 조건부 다중(복수)응답
  5. 복수 이항 응답

응답 유형 예시

1. 귀 댁에는 현재 직장, 학업 등의 이유로 타 지역(해외포함)에 살고 있는 배우자나 미혼자녀가 있습니까? ( )

① 있다 (총 _____ 명) (☞ 1-1번으로)　　② 없다 (☞ 2번으로)

Figure 1: 단일 선택형 문항

19. 귀하는 평소 어떤 교통 수단을 이용하십니까? 주로 이용하는 교통수단 2개만 순서대로 표시해 주십시오.

1순위 ( ), 2순위 ( )

① 자전거　② 오토바이, 화물차 등　③ 시내버스　④ 철도　⑤ 도시철도(지하철)　⑥ 택시　⑦ 승용/ 승합차　⑧ 고속/ 시외버스　⑨ 걸어서　⑩ 통근, 통학용 버스　⑪ 기타( )

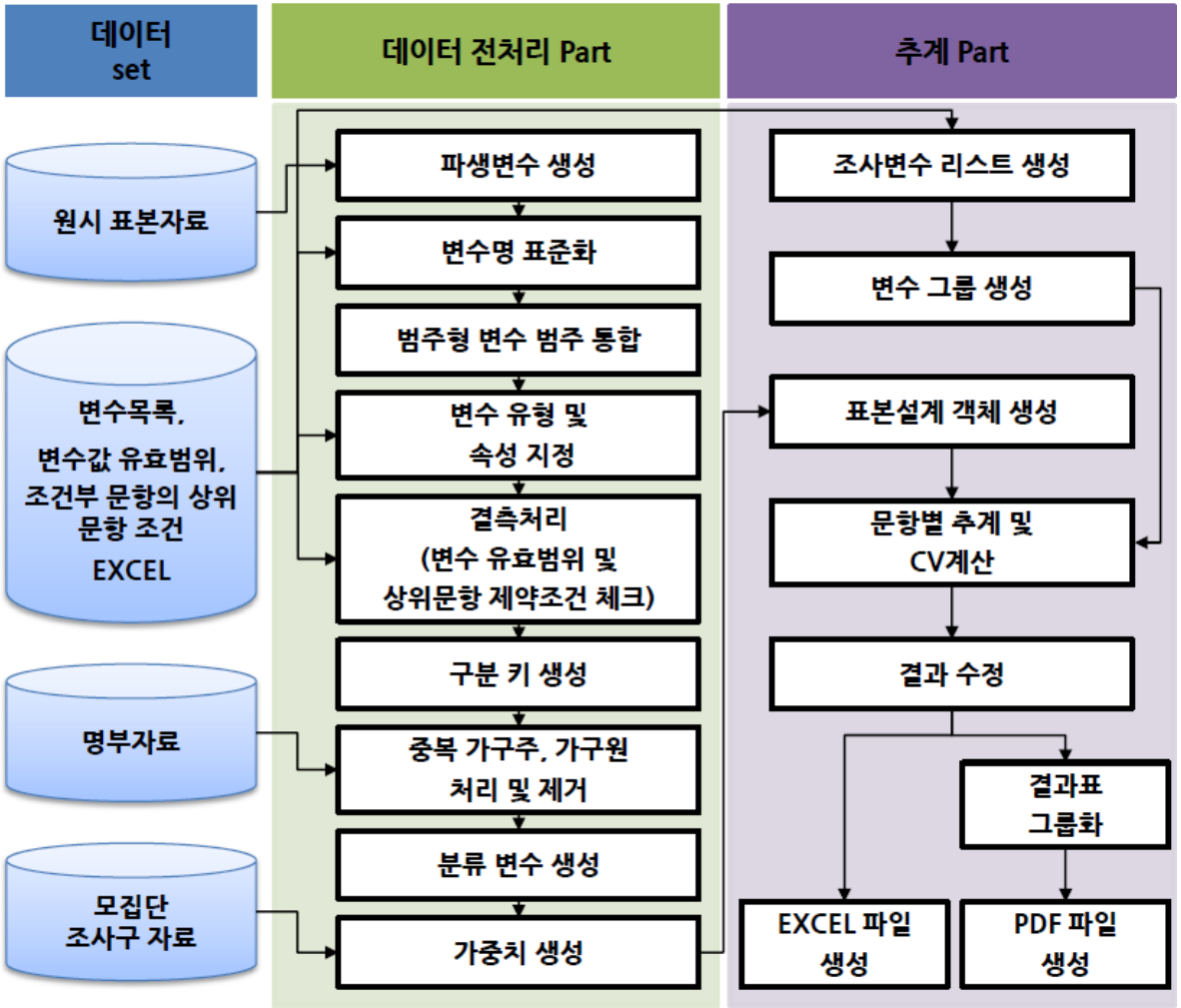
Figure 2: 단순 다중(복수) 응답

5-1. 올해(2014년 2학기) 대학교 등록금은 어떤 방법으로 마련하셨습니까? 해당되는 곳 모두에 각각 기입하여 주십시오. 현재 휴학 중이라면 최근 재학했던 학기를 기준으로 기입하여 주십시오. ( )

① 학생의 부모(가족) 도움 ····· ( )%
② 대출(학자금 대출, 일반대출, 마이너스통장 등) ····· ( )%
③ 자녀 스스로 벌어서 마련 ····· ( )%
④ 장학금 ····· ( )%
⑤ 기타( ) ····· ( )%

Figure 3: 조건부 다중(복수) 응답

분석의 절차


데이터 set
데이터 전처리 Part
추계 Part
원시 표본자료
변수목록,
변수값 유효범위,
조건부 문항의 상위 문항 조건
EXCEL
명부자료
모집단 조사구 자료
파생변수 생성
변수명 표준화
범주형 변수 범주 통합
변수 유형 및 속성 지정
결측처리 (변수 유효범위 및 상위문항 제약조건 체크)
구분 키 생성
중복 가구주, 가구원 처리 및 제거
분류 변수 생성
가중치 생성
조사변수 리스트 생성
변수 그룹 생성
표본설계 객체 생성
문항별 추계 및 CV계산
결과 수정
결과표 그룹화
EXCEL 파일 생성
PDF 파일 생성

프로그램에 사용된 R 패키지와 기능

R패키지	기능설명
data.table	데이터 객체 생성, 요약테이블의 작성
doParallel	멀티코어 병렬프로그램 지원
foreach	멀티코어 병렬프로그램 지원
survey	표본설계 객체 생성, 추계
HotDeckImputation	핫덱에 의한 결측치 대체
hot.deck	
XLConnect	엑셀데이터의 입출력
xtable	결과 테이블의 생성
rmarkdown	Document 생성
knitr	

문항별 추계표의 생성

- xtable이용: LaTeX table 생성

```
...
hlines.pos <- as.list(which(x[,1] != "") - 1)
cmd <- rep("\\hline", length(hlines.pos))
print(xtable(xx, caption=paste0(caption, "-(", i, ")"), digits=0,
             align=c("c", rep("l", 2), rep("r", ncol(xx)-2))),
      hline.after=c(-1, 0),
      include.rownames=F, type="latex",
      tabular.environment="longtable",
      caption.placement="top",
      format.args = list(big.mark = ","),
      ...
      add.to.row = list(
        pos = hlines.pos,
        command = cmd))
...
```

- knitr::kable, pander::pandoc.table 함수도 가능, but 튜닝하기가 어렵다 (경험상^^)

doParallel을 이용한 multicore 병렬연산

- 전체 문항 추계

```
# pre-processing steps...
library(doParallel)
library(foreach)
registerDoParallel(cores=35)
out_tb_list <- foreach(i=1:length(study_var_list)) %dopar%
  my_estimate_var(svar=study_var_list[[i]],
                  svar_tb=study_var_tb,
                  s_design=s_design,
                  ...,
                  condition_tb=condition_tb,
                  gvar = c("sigun_cd", "sex", "age", "edu", "eco", "income"))
stopImplicitCluster()
# post-rocessing steps...
```

Automatic report generation

- `knitr`: Elegant, flexibility and fast dynamic generation with R
  - Document templete + R code $\rightarrow$ single report with text, table, and figures
- `.Rmd`: R Markdown documents
- `rmarkdown`: R package to convert `.Rmd` into various documents

.Rmd file and report generation using rmarkdown

1. YAML header

```
---
title: "Document title.."
output:
  pdf_document:
    latex_engine: xelatex
---
```

2. R code chunks

````
```{r}
summary(iris);hist(iris[,1])
```
````

3. Compile .Rmd file as .html, .pdf, .doc

```
rmarkdown::render("test.Rmd", output_dir=output_path)
```

rmarkdown을 이용한 지역별 결과표 생성

- 하나의 프로그램으로 여러 지역별 결과표를 생성
  - for 루프를 이용, 지역에 따른 .Rmd파일 조립
  - .Rmd를 rendering
  - 지역별 결과표 생성

지역별 추계 결과표 생성 배치 프로그램의 구성

- generate_local_documents.R

```
# evaluate input arguments
args=(commandArgs(TRUE))
for(i in 1:length(args)) eval(parse(text=args[[i]]))
Sys.setenv(RSTUDIO_PANDOC="/usr/lib/rstudio-server/bin/pandoc")
for(rcode in rcodes){
  # generate Rmd file name
  rmd_file <- paste0(rcode, YEAR, "_v", rmd_version, ".Rmd")
  #make .Rmd file
  cat("......")
  #.Rmd file
  render(rmd_file, output_dir=output_path) # compile Rmd file
}
```

지역별 추계 결과표 생성 배치 프로그램의 실행

```
$ R CMD BATCH --no-save --no-restore '--args YEAR=2014 rmd_version="5.1"' \
generate_local_documents.R logs.txt &
```

- 결과파일 예

http://datamining.dongguk.ac.kr/ftp/gyeongbuk2017/2014/%EA%B3%B5%ED%86%B5_%EB%8F%84%EC%A0%84%EC%B2%B4_2014_v5.3.pdf

Summary

1. `doParallel` 병렬처리: 1~2H $\Rightarrow$ 5~7Min (빠른 디버깅)
2. `xtable` + `knitr` + `rmarkdown`: 47 $\times$ 3(개)의 문서 자동 생성

References

1. 경상북도 (2017). 사회조사 분석 및 R을 이용한 분석 프로그램 개발 결과 보고서.
2. https://www.google.co.kr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiou_favo7XAhVIebwKHfJ3CbUQFg
3. rstudio.com, R Markdown Reference Guide, https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf
4. Xie, Yihui. 2016. Knitr: A General-Purpose Package for Dynamic Report Generation in R. http://yihui.name/knitr/.
5. Yihui Xie's blog (knitr) http://yihui.name/en/categories/
6. R Bloggers: http://www.r-bloggers.com/ StackOverflow questions on R and knitr http://stackoverfow.com/questions/tagged/r+knitr