Datorlaboration 7

Måns Magnusson

VT 2015

Instruktioner

- **Allmänt**
  Vid tidigare laborationer har vi använt SAS för att dra urval, studier av teoretiska egenskaper hos estimatorer och allokera urval. Nu ska vi fokusera på att med R göra analyser, bortfallhantering och estimation då vi fått in data från en surveyundersökning.
- **Datamaterial**
  Vilket datamaterial som ska användas framgår av respektive uppgift. Allt datamaterial finns att tillgå **här** om inte annat anges. För att ladda ned datan, klicka på den datafil du vill ladda ned och klicka sedan på "Raw" med högra musknappen och klicka "Spara länk som...".
- **Hjälpmaterial**
  Behöver ni hjälp med att använda R-paketet survey finns, utöver dokumentationen, extra material här. Det finns också en bok *Complex surveys : a guide to analysis using R* som behandlar analyser med surveypaketet i R.
- Det är tillåtet att diskutera med andra men att plagiera andra grupper är **inte tillåtet**.
- Utgå från mallen för laborationsrapporter som går att ladda ned som **LyX** eller **PDF**. Samtliga labbrapporter ska lämnas in **i PDF-format via LISAM**.
- Deadline för labben framgår på **kurshemsidan**.
- **Laborationsrapport**
  Rapporten ska innehålla den kod ni kört, eventuella resultat samt svara på de frågor som finns i laborationen.

Innehåll

1 Förberedelser **3**
1.1 Ladda ned och läs in Survey 2010 . . . 3
1.2 Läsa in `survey` paketet . . . 3
1.3 Ladda ned och läs in `agpop.dat` . . . 3

2 Analys **4**
2.1 t-test . . . 5
2.2 χ^2 - test . . . 5
2.3 Linjär regression . . . 7

Referenser **8**

1 Förberedelser

1.1 Ladda ned och läs in Survey 2010

- I denna laboration ska vi börja analysera riktiga surveydata (med alla problem det innebär). Vi har fått tillgång till det datamaterial som ligger till grund för boken *Den svenska väljaren* Hagevi (2011). Ett mindre urval av variablerna i studien har sparats som en R-fil. Ladda ned filen från kurshemsidan och läs in den i R med följande funktion

```
load("svy2010.Rdata")
```

- Det går också att ladda in filen direkt från webben med `repmis`-paketet:

```
library(repmis)
data_path <- "https://raw.github.com/MansMeg/KursSvyMeth/master/Labs/DataFiles/svy2010.Rdata"
source_data(data_path)

## Downloading data from: https://raw.github.com/MansMeg/KursSvyMeth/master/Labs/DataFiles/svy2010.R
##
## SHA-1 hash of the downloaded data file is:
## d11fe835c8306f4746b9f950bc128985049815e2

## [1] "svy2010"
```

- Du ska nu ha läst in en fil med 1613 observationer och 70 variabler. Information om respektive variabler finns i dokumentet **KodbokSurvey2010.pdf** som finns på samma ställe som datamaterialet, dock finns inte alla variabler med i datasetet.
- Vi ska nu skapa två kontinuerliga variabler i datasetet `fathAge` och `mothAge` på följande sätt:

```
svy2010$fathAge <- svy2010$FR37 - svy2010$FR40_1
svy2010$fathAge[abs(svy2010$fathAge) > 100 | svy2010$fathAge < 0 ] <- NA
svy2010$mothAge <- svy2010$FR37 - svy2010$FR40_2
svy2010$mothAge[abs(svy2010$mothAge) > 100 | svy2010$mothAge < 0] <- NA
```

1.2 Läsa in `survey` paketet

- Börja med att läsa in `survey`-paketet

```
library(survey)
```

Om det inte går att läsa in paketet (ex. du har en egen dator) behöver du installera paketet först. Det kräver internetanslutning och då använder du följande kod.

```
install.packages("survey")
```

1.3 Ladda ned och läs in `agpop.dat`

- Ladda ned filen (se instruktionen). Identifiera den mapp där du har sparat filen `agpop.dat`. Använd funktionen `setwd()` för att ställa in den korrekta sökvägen .
- Läs in `agpop.dat` i R, vilket kan göras med följande kod:

```
agpop<-read.table("agpop.dat",header=TRUE,sep=",")
```

- Även i detta fall kan vi självklart använda `repmis`-paketet.

```
agpop <- source_data(url = "https://raw.github.com/MansMeg/KursSvyMeth/master/Labs/DataFiles/agpop.da

## Downloading data from: https://raw.github.com/MansMeg/KursSvyMeth/master/Labs/DataFiles/agpop.dat
##
## SHA-1 hash of the downloaded data file is:
## fdd78ace764a7b61254f073564ab50968cdd9375
```

- Glöm inte bort att ta bort bortfallet för variablerna `ACRES92` och `ACRES87` i `agpop`. För att ta bort saknade värden kan följande kod användas:

```
agpop<-agpop[agpop$ACRES92>0 & agpop$ACRES87>0,]
```

- Nästa steg är att vi vill lägga till en till kategorisk variabel i agpop (för att göra analys av kategoriska data senare). Skapa följande variabel `FARMCAT`.

```
agpop$FARMCAT <- cut(agpop$FARMS92, c(0, 500, 1000, Inf), right = FALSE)
agpop$REGION_S <- agpop$REGION=="S"
```

2 Analys

Precis som när det gäller att skatta medelvärden, proportioner eller andra populationsparametrar innebär designperspektivet att vi betraktar populationsparametrarna som fixa "sanna" värden i den ändliga populationen. Vi drar sedan ett urval för att kunna dra statistiska slutsatser om denna ändliga population. Dessutom kan vi på grund av att urvalet är ett stratifierat urval eller ett klusterurval få en designeffect i våra studier som behöver hanteras när vi gör våra tester.

Denna laboration kommer endast att beröra det specifika med just surveysituationen, inte kring dessa tester och modeller i sin helhet.

Börja med att dra ett (Neymanallokerat) stratifierat urval från `agpop` och skapa ett surveyobjekt på samma sätt som gjordes i laboration D4. Jag döper mitt objekt till `agSTRAT`.

```
STRATAindex <- stratsample(agpop$REGION, c("NC"=69,"NE"=7,"S"=122,"W"=102))
agSTRATAdata <- agpop[STRATAindex,]
fpc.strata<-numeric(300)
fpc.strata[agSTRATAdata$REGION=="NC"] <- 1049
fpc.strata[agSTRATAdata$REGION=="NE"] <- 209
fpc.strata[agSTRATAdata$REGION=="S"] <- 1370
fpc.strata[agSTRATAdata$REGION=="W"] <- 414
agSTRAT <- svydesign(~1, strata=~REGION, fpc=fpc.strata, data=agSTRATAdata)
```

Vi ska nu också studera effekter på analyserna om vi har en totalundersökning. Vi kan skapa ett surveyobjekt som är en totalundersökning på följande sätt.

```
agTOT <- svydesign(~1,  fpc=rep(nrow(agpop),nrow(agpop)), data=agpop)
```

Skapa sedan ett surveyobjekt baserat på datamaterialet i Survey 2010, jag kommer kalla detta objekt `svy2010design`.

```
fpc.srs<-rep(7529673, nrow(svy2010))
svy2010design<-svydesign(ids=~1, data=svy2010, fpc=fpc.srs)
```

2.1 t-test

a) Många gånger finns det ett intressa av att jämföra olika redovisningegrupper efter en variabel som kan vara av intresse. Finns det skillnader i vår **ändliga population** avseende redovisningsgrupper som inte beror på slumpen vi skapat i vårt urval. Vill vi jämföra två grupper med varandra (eller med ett fast värde) avseende en kontinuerlig variabel använder vi funktionen `svyttest()`. För att göra testet behövs dels ange en formula, vilket består av den kontinuerliga variabeln vi vill testa skillnader avseende (ex. y) och en kategorisk variabel med två klasser (ex. `grupp`). Sedan krävs designobjektet vi vill testa.

Exempel:

```
svyttest(formula = y ~ grupp, design = mittDesignObjekt)
```

Nedan har jag gjort ett t-test. Vad har jag testat? Vad är slutsatsen från testet?

```
svyttest(formula=mothAge~Valdeltagande, design=svy2010design)


Design-based t-test

data:  mothAge ~ Valdeltagande
t = -0.2186, df = 1611, p-value = 0.827
alternative hypothesis: true difference in mean is not equal to 0
sample estimates:
difference in mean
           -0.1138
```

Använd surveyobjektet du skapat för Survey 2010 och testa om det finns skillnader mellan moderata och socialdemokratiska väljare när det gäller moderns ålder. Vad får du för resultat? Vad är din slutsats?

Obs! Först behöver du plocka ut de personer i undersökningen som röstade på moderaterna eller socialdemokraterna. Det enklaste är att göra detta med `subset()`. Är du osäker på hur du använder `subset()` - se laboration D4 och avsnittet om redovisningsgrupper.

b) Ändlighetskorrektionens betydelse för designbaserade test Pröva nu att genomföra följande t-test.

```
svyttest(formula=FARMS92~REGION_S, design=agSTRAT)
```

Förklara kort vad detta test har gjort. Jämför detta test med ett vanligt modellbaserat t-test (se nedan). Vad får du för skillnader och vad beror dessa skillnader på?

```
t.test(x = agSTRATAdata$FARMS92[agSTRATAdata$REGION_S],
       y = agSTRATAdata$FARMS92[!agSTRATAdata$REGION_S])
```

Gör nu om samma test för `agTOT`, d.v.s. för en totalundersökning. Jämför sedan med att göra ett icke-surveybaserat t-test på hela undersökningen (och jämför med totalundersökningen. Vad skiljer sig

Får du ett annat resultat? Vad beror detta på? Förklara.

2.2 χ^2 - test

a) Ett annat vanligt test för kategoriska variabler är det klassiska χ^2-testet. Precis som för t-testet behöver vi korrigera våra tester för att ta hänsyn till dels vår ändlighetskorrektionsfaktor och dels för att beakta potentiella designeffekter.

För att genomföra ett test om valdeltagande är oberoende av vilket parti en person föredrar kan ett χ^2-test användas på följande sätt.

Exempel:

```
svytable(formula = ~FR15 + Valdeltagande, design = svy2010design)

                                    Valdeltagande
FR15                                   R<U+00F6>stade inte   R<U+00F6>stade
  Centerpartiet                                     18672         322100
  Feministiskt initiativ                             9336          28009
  Folkpartiet                                       37345         606855
  Kristdemokraterna                                  9336         228738
  Milj<U+00F6>partiet                                102699         723558
  Moderaterna                                       168052        1969945
  Piratpartiet                                       28009          32677
  Socialdemokraterna                                121371        1708531
  Sverigedemokraterna                                37345         252078
  V<U+00E4>nsterpartiet                               18672          340773
  Annat parti (v.g. ange vilket)...?:                42013          65354

svychisq(formula = ~FR15 + Valdeltagande, design = svy2010design)


Pearson's X^2: Rao & Scott adjustment

data:  svychisq(formula = ~FR15 + Valdeltagande, design = svy2010design)
F = 6.497, ndf = 10, ddf = 16120, p-value = 4.321e-10
```

Vad är din slutsats av testet ovan?

Som standard används Rao-Scott-korrektion av testet. Vi kan välja andra korrektioner som exempelvis Wald-korrektion genom att sätta argumentet `statistic` till `''Wald''`.

Exempel:

```
svychisq(formula = ~FR15 + Valdeltagande, design = svy2010design, statistic="Wald")


Design-based Wald test of association

data:  svychisq(formula = ~FR15 + Valdeltagande, design = svy2010design,     statistic = "Wald")
F = 2.312, ndf = 10, ddf = 1612, p-value = 0.0107
```

Välj nu två kategoriska variabler du vill testa om de är oberoende av varandra och testa dem med ett designbaserat χ^2-test. Vad är dina slutsatser?

b) Pröva nu att göra följande χ^2 test.

```
svychisq(formula = ~REGION_S + FARMCAT, design = agSTRAT)
```

Jämför nu testet ovan med ett vanligt (modellbaserat) χ^2-test.

```
chisq.test(table(agSTRATAdata$REGION_S, agSTRATAdata$FARMCAT))
```

Vad är skillnaden mellan de olika metoderna? Skiljer sig dem åt och vad beror det på i så fall?

c) Upprepa testen ovan men med surveyobjektet för totalundersökningen istället. Vad får du för resultat? Hur skiljer det sig från samma test med `agSTRAT`?

Gör nu ett vanligt (modellbaserat) chitvåtest på `agpop`-datasetet med `chisq.test()`. Hur skiljer sig resultaten om vi tar hänsyn eller inte tar hänsyn till ändlighetskorrektionen i `agTOT`? Resonera kring varför det blir på detta sätt.

2.3 Linjär regression

a) Till sist ska vi pröva att använda linjär regression i surveysammanhang, d.v.s. att i en regression ta hänsyn till urvalsdesign och ändlighetskorrketion. Precis som vid testerna ovan innebär att detta att vi skattar dessa värden i den **ändliga populationen** vi vill undersöka.

Vi ska nu pröva att anpassa en regressionsmodell på materialet avseende åkerytor i USA.

```
my_svy_model <- svyglm(formula = ACRES92 ~ FARMS92 + REGION, design=agSTRAT)
my_svy_model

## Stratified Independent Sampling design
## svydesign(~1, strata = ~REGION, fpc = fpc.strata, data = agSTRATAdata)
##
## Call:  svyglm(formula = ACRES92 ~ FARMS92 + REGION, design = agSTRAT)
##
## Coefficients:
## (Intercept)      FARMS92     REGIONNE      REGIONS      REGIONW
##    226872.5         87.1    -162410.3     -83596.1     559943.5
##
## Degrees of Freedom: 299 Total (i.e. Null);  292 Residual
## Null Deviance:     6.39e+13
## Residual Deviance: 4.92e+13  AIC: 8660
```

Vad har jag gjort för analys ovan? Vad är dina slutsatser baserat på denna analys?

Precis som vid vanlig regression kan vi använda `summary()` för att få ytterligare information från vår modell (hypotestester m.m.).

Pröva sedan att anpassa en vanlig linjär regression på samma data (men utan att ta hänsyn till designen).

Exempel:

```
my_lm_model <- lm(ACRES92 ~ FARMS92 + REGION, agSTRATAdata)
```

Vad är skillnaden i resultat mellan de två metoderna? Vad beror denna skillnad på?

b) Gör om samma linjära regressionsanalys som vi gjorde på urvalet `agSTRAT` men med `agTOT`.

Vad får du medelfel på β-koefficienterna? Täcker konfidensintervallen du fick baserat på urvalet `agSTRAT` de sanna värdena i populationen?

Referenser

Hagevi, M., 2011. Den svenska väljaren, 1st Edition. Boréa, Umeå.

Lumley, T., 2010. Complex surveys : a guide to analysis using R. Wiley-Blackwell, Oxford.