Datorlaboration 5

Måns Magnusson

VT 2014

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**.
- Godkänd laboration ger **0.5 p** på tentamen. Extrapoängen är endast giltiga vid denna kursomgång.
- 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

1.1 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")
```

1.2 Läsa in `sweSCB` paketet

```
library(sweSCB)
```

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

```
install.packages("devtools") devtools::install_github("sweSCB","LCHansson")
library(sweSCB)
```

Detta paket finns i dagsläget inte på CRAN utan bara på GitHub. Mer information finns här: **sweSCB**.

1.3 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")
```

- Du ska nu ha läst in en fil med 1613 observationer och 70 variabler. Infomation om respektive variabler finns i dokumentet KodbokSurvey2010.pdf som finns på samma ställe som datamaterialet, dock finns inte alla variabler med i datasetet.

2 Bortfallsstudie med simulerade data

2.1 Simulera population

a) Vi ska nu visa hur det är möjligt att använda R för att bortfallskompensera för flera kategoriska variabler. Vi kommer inledningsvis att simulera en population (som bygger på befolkningsdata från Linköping). Kör följande kod för att generera en population som liknar Linköping.

```
sex <- sample(c("Man", "Kvinna"), size = 98462, replace = TRUE, prob = c(51144,
    47318)/98462)
age <- sample(c("15-24", "25-34", "35-44", "45-54", "55-64"), size = 98462,
    replace = TRUE, prob = c(23728, 21706, 19051, 18213, 15764)/98462)
simPop <- data.frame(sex = sex, age = age)
```

Vi har nu får population och baserat på denna skapar vi variabeln `assault` genom simulering för att efterlikna den utsatthet för brott som beskrivs i Nationella trygghetsundersökningen (Irlander and Hvitfeldt, 2012). Från denna undersökning vet vi att sannolikheten för att utsättas för misshandel är störst i de yngre åldrarna. Vi simulerar utsatthet för misshandel på följande sätt:

```
simPop$pAssault <- 0.001
simPop$pAssault[simPop$age == "15-24"] <- 0.05
simPop$pAssault[simPop$age == "25-34"] <- 0.02
simPop$pAssault[simPop$age == "35-44"] <- 0.005
simPop$pAssault <- simPop$pAssault + 1.5 * simPop$pAssault * as.numeric(simPop$sex ==
    "Man")
simPop$assault <- rbinom(n = length(sex), size = 1, prob = simPop$pAssault)
```

Vi ska också simulera en variabel som inte alls har med utsatthet för brott (eller bortfall) att göra. Skapa variabeln stjärntecken i populationen på följande sätt:

```
simPop$astroSign <- sample(c("Skorpion", "Skytt", "Tvilling", "Stenbock", "Jungfru",
    "Lejon", "Oxe"), size = length(sex), replace = TRUE)
```

Vi har nu en simulerad population som vi kan pröva att dra urval från.

Hur många har vart utsatt för brott i din population (vad är det sanna värdet vi försöker uppskatta), totalt och i de olika åldersgrupperna? [**Tips!** `aggregate()`] Hur stor är andelen som varit utsatta för misshandel?

b) Vi ska nu introducera en bortfallsmodell för vår population, d.v.s. sannolikheten för att en person skulle delta i studien eller inte. Vi gör detta genom att introducera denna som en sannolikhet som vi sedan kan använda för att simulera bortfall i studien. Vi vet sedan tidigare att kön och åldersgrupp spelar roll så vi skapar en bortfallsmodell baserat på dessa variabler.

```
simPop$response <- 0.95
simPop$response[simPop$age == "15-24"] <- 0.35
simPop$response[simPop$age == "25-34"] <- 0.5
simPop$response[simPop$age == "35-44"] <- 0.65
simPop$response[simPop$age == "45-54"] <- 0.85
simPop$response <- simPop$response - 0.15 * as.numeric(simPop$sex == "Man")
```

Vad har vi för antagande i vår bortfallsmodell, MCAR, MAR eller NMAR? Beskriv respektive antagande och varför/varför inte det är tillämpligt.

2.2 Urval och bortfallsanalys

a) Dra ett OSU av storlek $n = 4000$ från din population, skapa ett surveyobjekt och skatta totalen och proportionen i befolkningen avseende utsatthet för misshandel med tillhörande konfidensintervall. Skatta även utsatthet i de olika åldersgrupperna. Se föregående laboration för detaljer. Ignorera svarssannolikheten just nu (d.v.s. utgå från att alla svarar). Vad får du för resultat, täcker konfidensintervallen de sanna värdena i populationen?

b) Vi ska nu simulera bortfall i din studie. Skapa en variabel du kallar `respInd` som antar värdet `1` om personen svarar/deltar och `0` om personen inte deltar i studien. Det kan exempelvis göras på följande sätt (där `mittUrval` är det dataset jag skapat genom att dra ett urval från `simPop`):

Obs! Detta är ett vanligt dataset och inte ett surveyobjekt

```
mittUrval$respInd <- rbinom(length(mittUrval$response), 1, mittUrval$response)
```

Vi har nu skapat en variabel (`respInd`) som indikerar om personen deltar eller inte. Använd nu bara de svarande för att skapa ett surveyobjekt och skatta totalen och proportionen utsata för misshandel i populationen. För att välja ut de svarande kan du göra på följande sätt:

```
mittUrvalSvarande <- mittUrval[mittUrval$respInd == 1, ]
```

Vad händer? Vilken effekt har bortfallet på dina skattningar? Täcker konfidensintervallet det sanna värdet? Hur är det om du skattar befolkningstotalen i respektive åldersgrupp?

c) Vi ska nu göra en bortfallsanalys. Det finns två situationer antingen får vi bara datat för de som deltagit i undersökningen, eller så har vi data för de som ingått i urvalet. Detta ger två olika situationer för hur bortfallsanalysen kan genomföras. I detta exempel utgår vi från att vi känner till alla som deltog i studien (i del 3 på sidan 6 kommer vi pröva att göra en bortfallsanalys i den andra situationen).

Utgå från att du känner till variablerna `sex` och `age` för både de som svarat och de som inte svarat (vi kan bara analysera bortfallet efter de variabler vi har data både i undersökningen och på målpopulatiosnnivå). Anpassa en logistisk regressionsmodell med `respInd` som beroende variabel och `sex` och `age` som förklarande variabler. Detta kan göras på följande sätt i R:

```
bortfallsAnalys <- glm(respInd ~ age + sex + astroSign, family = binomial(logit),
    data = mittUrval)
summary(bortfallsanalys)
```

Vad drar du för slutsatser från denna bortfallsanalys. Vilka variabler kan förklara bortfallet / svarande? Stämmer det med den bortfallsmodell vi använde ovan?

2.3 Kalibrera för bortfallet

a) Vi ska nu kalibrera det surveyobjekt du skapade i uppgift 2.2 på föregående sida. Detta är lite klurigt att få till första gången och det kan framstå som lite underligt hur det görs. Dra dig till minnes att kalibrering egentligen är en form av regressionsskattning där vi känner till totalerna i befolkningen och kalibrerar efter dessa. Ett första steg vi behöver ta är därför att räkna ut totaler på ett sätt som R "förstår" ur ett "regressionsperspektiv". R behöver göra om kategorivariablerna till dummyvariabler (precis som vid vanlig regression) och sedan få information om totalerna för dessa dummyvariabler.

Pröva dig igenom koden nedan och försök förstå exakt vad för totaler som skapas i vektorn `popVector`. Observera att för att kalibrera är det viktigt att elementnamnen på `popVector` är samma namn som namnet för `modMat`.

Exempel:

```
modMat <- model.matrix(~as.factor(age) + as.factor(sex), simPop)
popVector <- colSums(modMat)
popVector
```

Vad får du för resultat i `popVector`? Vad innebär detta resultat?

Skapa nu en egen `popVector` beroende på vilka variabler som kan förklara bortfallet i enlighet med din bortfallsanalys.

b) Nu är vi klara för att kalibrera det surveyobjekt vi skapade i uppgift 2.2. För att kalibera använder vi funktionen `calibrate()` i surveypaketet och de argument som krävs är det surveydesignobjekt som ska kalibreras, den regressionformel som ska användas (men utan beroende variabel) och vektorn med populationsvärden, `population`.

Observera att du måste använda exakt samma formel i `calibrate()` som användes i `model.matrix()` ovan för att det ska fungera.

Exempel:

```
calMittUrvalSvarande <- calibrate(mittUrvalSvarande, formula = ~as.factor(age) +
    as.factor(sex), population = popVector)
```

Nu är surveyobjektet kalibrerat (och sparat som `calMittUrvalSvarande`) och vi kan producera skattningar av baserat på detta objekt med samma funktioner som tidigare (ex. `svymean()` och `svytotal()`). Skatta antalet och andelen utsatta för misshandel med tillhörande konfidensintervall med det kalibrerade surveyobjektet.

Vad får du för resultat? Täcker konfidensintervallet de sanna värdena i populationen?

c) Pröva nu att inkludera även `astroSign` i din kalibreringsmodell variabel (som uppenbart är fel). Pröva att estimera det totala antalet och andelen utsatta för misshandel. Vad får det för effekt att inkludera en felaktig variabel?

d) Pröva att inte inkludera någon variabel i din modell (d.v.s. bara intercepten) och estimera sedan totalen i för den simulerade population. Vad innebär detta? Vilket antagande om bortfallet innebär detta, MCAR, MAR eller NMAR?

3 Bortfall i undersökningen *Den svenska väljaren 2010*

Vi ska nu pröva att göra en bortfallsanalys på ett riktigt dataset. Vi ska nu undersöka det material som ligger till grund för analyserna i Hagevi (2011). Vi har redan laddat ned och läst in datamaterialet i R i del 1.3 på sidan 2. Vi ska nu göra en bortfallsanalys på detta material, kalibrera och göra en enkel analys. Undersökningen har gjorts genom ett obundet slumpmässigt urval (OSU) från befolkningen i Sverige 16 - 85 år. **Här** finns en kodbok som beskriver variablerna i datasetet.

a) Börja med att undersök vilka variabler du tycker verkar intressanta och du vill analysera i materialet, exempelvis en eller flera delar av Fråga 28 som handlar om befolkningens attityder i olika politiska frågor.

3.1 Bortfallsanalys

a) Som ett första steg ska vi genomföra en bortfallsanalys på det insamlade datamaterialet. Inte allt för sällan finns bara data för de individer vi samlat in att tillgå, utan data från alla som ingick i urvalet (till skillnad mot uppgift 2.2 ovan). Då är det inte möjligt att analysera bortfallet med en logistisk regression.

Som ett alternativ till logisktisk regression kan vi genomföra skattningar och jämföra vilka variabler som avviker från de "sanna" värdena i populationen. Ett exempel är skillnaden mellan hur många vi uppskattar (baserat på vårt urval) röstade på ett visst parti i vårt urval och hur många som faktiskt röstade på detta parti (baserat på data från SCB).

b) För att ta reda på de sanna värdena i populationen använder vi data från Statistiska centralbyrån. Det finns ett R-paket (`sweSCB`) för att få tillgång till data från Statistiska centralbyrån direkt från R. På följande sätt gör du för att komma åt data från SCB (precis som innan behöver du installera `sweSCB` om det inte är installerat):

```
library(sweSCB)
mittData <- findData()
```

Om detta paket inte fungerar för dig kan du hämta samma data från SCB:s webbplats.

c) Ta reda på hur stor den aktuella populationen var år 2010 och skapa ett surveyobjekt med denna data.

d) Använd surveyobjektet för att skatta totalvärdena på följande variabler. Jämför dina totalskattningar med data från SCB. Detta kan du hämta med paketet `sweSCB` eller från SCB:s hemsida.

- `Partival` (jmf. antal röstande i demokratistatistiken från SCB)
- `Kön` (jmf. antal kvinnor/män i befolkningsstatistiken från SCB)
- `Ålder9` : Åldersklasser (jmf. antal i olika åldersgrupper i befolkningsstatistiken)
- `romr`: Riksområde, se kodbeskrivningen för detaljer. (jmf. antal personer i respektive område i befolkningsstatistiken)

Det finns fler variabler du kan använda i din bortfallsanalys om du har lust. Vilka kommer du på?

Vad får du för resultat? Innebär bortfallet någon effekt på någon av dessa variabler? Vad kan bortfallet bero på? Vilka variabler bör du inkludera, vilka variabler tror du inte är aktuella? Sammanställ dina resultat till en bortfallsanalys.

3.2 Kalibrera materialet

Du har nu gjort din bortfallsanalys och nästa steg är att kalibrera ditt material efter de sanna värdena i populationen efter de variabler du har identifierat i 3.1 ovan.

Skapa ett kalibrera ditt surveyobjekt på samma sätt som i avsnitt 2.3 på sidan 4. Använd sedan detta kalibrerade surveyobjekt för att skatta de variabler du identifierat som intressanta i början av uppgiften.

Har skattningarna påverkats av att du kalibrerat ditt data? Varför / varför inte? Gör en mindre analys av de utvalda variablerna och resonera kring resultatet och hur bortfallet kan ha påverkat analysen du gör.

Nu är du klar!

Referenser

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

Irlander, Å., Hvitfeldt, T., 2012. NTU 2011 : om utsatthet, trygghet och förtroende. Brottsförebyggande rådet, Stockholm.

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