Computer Workshop on Using Multilevel Models for Small Area Estimation

NCRM Summer School, 5 July 2005

Introduction to R

In this workshop small area estimation will be illustrated using the statistical software R. R is free, powerful object oriented software for performing statistical analysis. It is almost perfectly compatible with S-plus. The only thing you need to do is download the software from the internet and use an editor to write your programme (e.g. Notepad). For this course no workshop knowledge of R is required. We have already written and documented the code for you. We will use the code step by step in order to illustrate some aspects of small area estimation.

[Don’t attempt the instructions in the rest of this page. You can use these when you download R at home]

How to download R

1 Go to

http://lib.stat.cmu.edu/R/CRAN/

2 On you right hand side you will see

Windows (95 and later)

Click there

3 Click on base

4 Click on rw2011.exe and save it in your hard disc

This is the last available version of the software. This is an executable file, which you can save in your hard disc. By double clicking on the name of this file R will be automatically installed and create a shortcut. All you need to do is follow the installation process.

How to open R

You can open an R working space by doing the following:

Start->All Programs->R->R 2.0.1

Leave the R working space open as we will use this for the purposes of our workshop.

The dataset

In this workshop small area estimation will be performed with multilevel models. There are two versions of multilevel models (a) random intercepts and (b) random slopes. Without loss of generality, we define a random intercepts and a random slopes model with one covariate.

A random intercepts model is defined below

$$y_{id} = \beta_0 + \beta_1 x + u_{0d} + \varepsilon_{id}$$

where β_0 is the intercept, β_1 shows the effect of x on y, u_{0d} is the area random effect and ε_{id} is the error term. You can see that each area has a different intercept. This is because each area has a different random effect. However, the slope β_1 is the same for all areas.

A random slopes model is defined below

$$y_{id} = \beta_0 + \beta_1 x + u_{0d} + u_{1d} x + \varepsilon_{id}$$

You can now see that each area has both a different intercept and a different slope.

The dataset we will use for our illustrations is a simulated one. A population is generated under a random intercepts model using the following steps

1. A continuous covariate *x* is generated under a normal distribution with mean and variance depending on the area
2. An area specific effect is generated under a normal distribution with mean 0 and standard deviation 10. The area specific effect is denoted by u
3. The error term is generated under a normal distribution with mean 0 and standard deviation 40. The error term is denoted by ε
4. The y values for individual *i* in area *d* are then constructed as follows

$$y_{id} = 5 + x_{id} + u_{0d} + \varepsilon_{id}$$

Having generated the population, a sample is selected from each area.

The Problem

We are interested in estimating the mean and the median of y in each area *d*.

Activate an internet browser (Mozilla Firefox is used on the lab PCs) and go to the website **http://www.ncrm.ac.uk/ss**. Right click on the file **Workshop for Summer School 1.txt** choose “Save Link As”, and save the file on the C drive. Repeat the same for the file with name **Workshop for Summer School 2.txt.** We will start working with the first file.

You can view the contents of this file using Notepad. Open Notepad by doing the following:

Start->All Programs->Accessories->Notepad

Having opened Notepad, find the file where you saved it earlier and open it using:

File->Open

The code is documented so that you can understand what each command is doing.

Loading the code in R

Go to R and do the following

File -> source R code

Change the type of the file so that you can view txt files

Locate the file with name **Workshop for Summer School 1.txt** and

Press open

Using the previous steps we read the source code that performs the required computations.

Viewing the data

To view the first 20 observations of the population

Type

pop[1:20,] and press **ENTER**

To view the first 20 sample observations

Type

data[1:20,] and press **ENTER**

Now type:

summary (sample.size) and press **ENTER**.

The previous command gives you summary statistics for the distribution of the area sample sizes i.e. the mean sample size, the area with the smallest sample size, the area with the largest sample size etc.

Let’s produce a graph of the area sample sizes

Type

plot(sample.size, xlab=“Area”) and press **ENTER**

This will give you a plot of the area sample sizes

Estimating small area means using Multilevel Models

In this workshop we will illustrate small area estimation using two versions of multilevel models (a) a random intercepts model and (b) a random slopes model.

Fitting a multilevel model with R

Multilevel models are implemented in R using the library nlme. Here is an example of how to fit a random intercepts model to our sample data

Type **library(nlme)** and press **ENTER**

Type

model.random.intercepts<-lme(y~x,random=~1|regioncode) and press **ENTER**

The results from fitting the multilevel model are saved in object **model. random.intercepts**. To see a summary of the fitted model

Type

```
summary (model.random.intercepts)
```

The random intercepts model can be extended to include random slopes.

Type

model.random.slopes<-lme(y~x,random=~1+x|regioncode) and press **ENTER**

The results from fitting the multilevel model are saved in object **model. random.slopes**

Type

```
summary (model.random.slopes)
```

The code we use computes estimates of small area means under the random intercepts and random slopes models.

To view the small area mean estimates under the random intercepts model

Type

means.intercept and press **ENTER**

To view the small area mean estimates under the random slopes model

Type

means.slopes and press **ENTER**

We can also compute the direct small area mean estimates. These are based only on the area specific data. To compute the direct small area mean estimates do

```
direct<-aggregate(data[,1],list(data[,3]),mean)
```

```
direct.mean<-direct[,2]
```

Type

direct.mean and press **ENTER**

Because we have generated the population, we also know the true small area means. Therefore, we can see how close the small area mean estimates are to the true small area means.

To view the true (population) small area means

Type

means.true and press **ENTER**

A plot will help us to study the results further

Type

plot(means.true,ylim=c(0,500), col="red", type="l", ylab="Small Area Mean Estimates”, xlab="Area") and press **ENTER**

points(direct.mean, pch="D") and press **ENTER**

lines(means.intercept, col="blue") and press **ENTER**

lines(means.slopes, col="black") and press **ENTER**

Small area estimation and the shrinkage effect

The small area estimator of the population mean in area d is defined as follows

$$\hat{\bar{Y}}_d = \hat{\gamma}_d \hat{\bar{y}}_d + (1-\hat{\gamma}_d)\bar{X}_d \hat{\beta},$$

where $\hat{\bar{y}}_d$ is the direct mean estimate for area d, $\bar{X}_d$ is the population mean of *X* in area d and $\hat{\gamma}_d$ is the shrinkage factor estimated as follows

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_\varepsilon^2}{n_d}}. \quad (1)$$

In (1) σ_u^2 is the between areas variance, σ_ε^2 is the within areas variance and n_d is the area sample size. It is straightforward to compute γ_d using (1). Our code does this automatically.

To see the shrinkage factors for each area under the random intercepts model

Type

shrinkage.factor and press **ENTER**

The shrinkage factor can be also estimated under the random slopes model but this involves a more complex expression for $\hat{\gamma}_d$.

What determines the shrinkage factor is the area sample size. To study the relationship between the shrinkage factor and the area sample sizes we use the following plot

Type

```
plot(shrinkage[,2],shrinkage[,1],type="l",ylab="ShrinkageFactor",xlab="Sample
Size",main="The relation between sample size and shrinkage factor")
```

Estimating small area medians

For deriving estimates of the small area medians we will use a different script. The population data are still generated under a random intercepts model but using chi-square instead of normal assumptions

1. A continuous covariate x is generated under a chi-square distribution with degrees of freedom depending on the area
2. An area specific effect is generated under a chi-square distribution with 1 degree of freedom
3. An error term is generated under a chi-square distribution with 3 degrees of freedom
4. The y values for individual *i* in area *d* are then constructed as follows

$$y_{id} = 5 + x_{id} + u_{0d} + \varepsilon_{id}$$

Go to R and do the following

File -> source R code

Change the type of the file so that you can view txt files

Locate the file with name **Workshop for Summer School 2.txt** (This file is in the location you saved it at the beginning of the workshop)

Press open

The code we provide computes estimates of small area medians under the random intercepts and random slopes models. The estimates of the small area medians are derived using the Chambers-Dunstan (1986) estimator of the population distribution function. Estimation under the Chambers-Dunstan estimator involves a numerical solution, which is implemented with our code.

To view the small area median estimates under the random intercepts model

Type

medians.intercepts and press **ENTER**

To view the small area median estimates under the random slopes model

Type

medians.slopes and press **ENTER**

Because we have generated the population, we also know the true small area medians. Therefore, we can see how close the small area median estimates obtained from the multilevel models are to the true small area means.

To view the true (population) small area medians

Type

medians.true and press **ENTER**

A plot will help us to study the results further

Type

plot(medians.true,ylim=c(0,330), col="red", type="l", ylab="Small Area Median Estimates”, xlab="Area") and press **ENTER**

lines(medians.intercepts, col="blue") and press **ENTER**

lines(medians.slopes, col="black") and press **ENTER**

References

Chambers R.L. and Dunstan R. (1986). Estimating Distribution Functions from Survey Data**,** *Biometrika***, 73 (3),** pp. 597-604 .