XI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 24–25–26 de outubro de 2013

mme: An R package for small area estimation with area level multinomial mixed models

López Vizcaíno Mª Esther[1], Lombardía Cortiña Mª José[2] e Morales González Domingo[3]

[1]Instituto Galego de Estatística
[2]Universidade da Coruña
[3]Universidade Miguel Hernández de Elche

RESUMO

The mme package for R implements three multinomial area level mixed models for small area estimation. The first model is based on the area level multinomial mixed model with independent random effects for each category of the response variable (López-Vizcaíno et al, 2013a). In the second model we take advantage from the availability of survey data from different time periods and we use a multinomial model with independent random effects for each category of the response variable and with independent time and domain random effects. The third model is similar to the second one, but with correlated time and domain random effects. In all the models the package use two approaches to estimate the mean square error (MSE), first through an analytical expression and second by bootstrap techniques.

Palabras e frases chave: small area, R package, multinomial mixed models.

1. INTRODUCTION

In Spain, like in other European countries, the estimation of some socioeconomic indicators (employed, unemployed, poverty,...) is made by means of surveys that use a stratified sampling design. Many times the stratification variable is the size of the municipality. As most municipalities and another local areas are not represented in the sample and many of them are present with a very small sample size, the estimates at the municipal or local level are not accurate enough. Small sample sizes and, in some cases, no sample at all is the main problem when performing local estimations. In this situation the sample size could be enlarged but this, in addition to cause an increase in the costs and in the denial by the respondents to answer the sample questionnaire, can lead to other kind of damages due to delays in obtaining results and to the impact of non-sampling errors. Therefore the increase of the sample size is not always advisable and even sometimes unfeasible from an economic point of view.

The interest in developing small area estimation techniques to solve these problems in a reasonable way is growing among statisticians. The term "small area" is often used to refer to geographic areas but it can also be applied to other interesting areas with non geographical boundaries (domains), like age groups, economic activity sectors and so on. It is the small sample size in the domain, and consequently the large variance of the "direct" estimators, the key point defining the concept of small area. It is not the actual size of the area. In the small area estimation context, an estimator of a parameter in a given domain is direct if it is based only on the sample data of the specific domain. A drawback of these estimators is that they can not be calculated when there is no sample observations in an area of interest.

Generally small area estimation techniques can be divided into design-based methods and model-based methods. The model-based methods make inference by taking into account the underlying model.The estimators based on these methods are useful because they give to practitioners an idea of how the data generation process is and how the different sources of information are incorporated. Mixed models are suitable for small area estimation due to its flexibility to make an effective combination of different sources of information and to its capacity to describe the various sources of error. These models incorporate

random area effects that explain the additional variability that is not explained by the fixed part of the model.

The objective of this work is to present a R package that implements three multinomial area level mixed models for small area estimation. The first model is based on the area level multinomial mixed model with independent random effects for each category of the response variable (López-Vizcaíno et al, 2013a). In the second model we take advantage from the availability of survey data from different time periods and we use a multinomial model with independent random effects for each category of the response variable and with independent time and domain random effects. The third model is similar to the second one, but with correlated time and domain random effects. In all the models the package use two approaches to estimate the mean square error (MSE), first through an analytical expression and second by bootstrap techniques.

2. MODELS

Let us start by giving some notation and assumptions. Let us use indexes $k = 1, \ldots, q-1$, $d = 1, \ldots, D$ and $t = 1, \ldots, T$ for the categories of the target variable, for the D domains and for the T time periods respectively. Let $u_{1,dk}$ and $u_{2,dkt}$ be the random effects associated to the domain d and the category k and to the domain d, the category k and the time instant t respectively. In the third model (Model 3) of this work we write the random effects in the form

$$\begin{aligned} \boldsymbol{u}_1 &= \underset{1\le d\le D}{\text{col}}(\boldsymbol{u}_{1,d}), \quad \boldsymbol{u}_{1,d} = \underset{1\le k\le q-1}{\text{col}}(u_{1,dk}), \quad \boldsymbol{u}_2 = \underset{1\le d\le D}{\text{col}}(\boldsymbol{u}_{2,d}) \\ \boldsymbol{u}_{2,d} &= \underset{1\le k\le q-1}{\text{col}}(\boldsymbol{u}_{2,dk}), \quad \boldsymbol{u}_{2,dk} = \underset{1\le t\le T}{\text{col}}(u_{2,dkt}), \quad \boldsymbol{u}_{2,dt} = \underset{1\le k\le q-1}{\text{col}}(u_{2,dkt}), \end{aligned}$$

and we suppose that

1. $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are independent,
2. $\boldsymbol{u}_1 \sim N(\boldsymbol{0}, \boldsymbol{V}_{u_1})$, where $\boldsymbol{V}_{u_1} = \underset{1\le d\le D}{\text{diag}}\,(\underset{1\le k\le q-1}{\text{diag}}\,(\varphi_{1k}))$, $k = 1, \ldots, q-1$.
3. $\boldsymbol{u}_{2,dk} \sim N(\boldsymbol{0}, \boldsymbol{V}_{u_{2,dk}})$, $d = 1, \ldots, D$, $k = 1, \ldots, q-1$, are independent with covariance matrix AR(1), i.e. $\boldsymbol{V}_{u_{2,dk}} = \varphi_{2k}\Omega_d(\phi_k)$ and

$$\Omega_d(\phi_k) = \Omega_{d,k} = \frac{1}{1-\phi_k^2}\begin{pmatrix} 1 & \phi_k & \ldots & \phi_k^{T-2} & \phi_k^{T-1} \\ \phi_k & 1 & \ddots & & \phi_k^{T-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \phi_k^{T-2} & & \ddots & 1 & \phi_k \\ \phi_k^{T-1} & \phi_k^{T-2} & \ldots & \phi_k & 1 \end{pmatrix}_{T\times T}.$$

It holds that $\boldsymbol{V}_u = \text{var}(\boldsymbol{u}) = \text{diag}(\boldsymbol{V}_{u_1}, \boldsymbol{V}_{u_2})$, where $\boldsymbol{V}_{u_2} = \text{var}(\boldsymbol{u}_2) = \underset{1\le d\le D}{\text{diag}}\,(\underset{1\le k\le q-1}{\text{diag}}\,(\boldsymbol{V}_{u_{2,dk}}))$. We also assume that the response vectors $\boldsymbol{y}_{dt} = \underset{1\le k\le q-1}{\text{col}}(y_{dkt})$, conditioned to $\boldsymbol{u}_{1,d}$ and $\boldsymbol{u}_{2,dt}$, are independent with multinomial distributions

$$\boldsymbol{y}_{dt}|_{\boldsymbol{u}_{1,d},\boldsymbol{u}_{2,dt}} \sim \text{M}(\nu_{dt}, p_{d1t}, \ldots, p_{dq-1t}),\ d = 1, \ldots, D,\ t = 1, \ldots, T. \tag{0.1}$$

where the ν_{dt}'s are known integer numbers which are equal to n_{dt} in the considered real data case. The covariance matrix of $\boldsymbol{y}_{dt}$ conditioned to $\boldsymbol{u}_{1,d}$ and $\boldsymbol{u}_{2,dt}$ is $\text{var}(\boldsymbol{y}_{dt}|\boldsymbol{u}_{1,d}, \boldsymbol{u}_{2,dt}) = \boldsymbol{W}_{dt} = \nu_{dt}[\text{diag}(\boldsymbol{p}_{dt}) - \boldsymbol{p}_{dt}\boldsymbol{p}'_{dt}]$, where $\boldsymbol{p}_{dt} = \underset{1\le k\le q-1}{\text{col}}(p_{dkt})$ and $\text{diag}(\boldsymbol{p}_{dt}) = \underset{1\le k\le q-1}{\text{diag}}\,(p_{dkt})$. For the natural parameters $\eta_{dkt} = \log\frac{p_{dkt}}{p_{dqt}}$, we assume the model

$$\eta_{dkt} = \boldsymbol{x}_{dkt}\boldsymbol{\beta}_k + u_{1,dk} + u_{2,dkt}, \quad d = 1, \ldots, D,\ k = 1, \ldots, q-1,\ t = 1, \ldots, T, \tag{0.2}$$

where $\boldsymbol{x}_{dkt} = \underset{1\le r\le p_r}{\text{col}'}\,(x_{dktr})$, $\boldsymbol{\beta}_k = \underset{1\le r\le p_k}{\text{col}}\,(\beta_{kr})$ and $p = \sum_{k=1}^{q-1} p_k$.

Along the work we also consider two simpler models. Model 2 is obtained by restricting Model 3 to $\phi_1 = \ldots = \phi_{q-1} = 0$. Model 1 is obtained by restricting Model 2 to one time period ($T = 1$) and by considering only the random effect $\boldsymbol{u}_1$. This is the model studied by López-Vizcaino et al. (2013a). For the sake of brevity we skip formulas for Models 1-2. In matrix notation, Model 3 is

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \boldsymbol{Z}_2\boldsymbol{u}_2 = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},$$

where $\boldsymbol{Z} = (\boldsymbol{Z}_1', \boldsymbol{Z}_2')'$, $\boldsymbol{\eta} = \underset{1\leq d\leq D}{\text{col}}(\boldsymbol{\eta}_d)$, $\boldsymbol{X} = \underset{1\leq d\leq D}{\text{col}}(\boldsymbol{X}_d)$, $\boldsymbol{Z}_1 = \underset{1\leq d\leq D}{\text{diag}}(\boldsymbol{Z}_{1d})$, $\boldsymbol{Z}_2 = \underset{1\leq d\leq D}{\text{diag}}(\boldsymbol{Z}_{2d})$,

$$\begin{aligned}
\boldsymbol{\eta}_d &= \underset{1\leq k\leq q-1}{\text{col}}\left(\underset{1\leq t\leq T}{\text{col}}(\eta_{dkt})\right), \quad \boldsymbol{X}_d = \underset{1\leq k\leq q-1}{\text{diag}}\left(\underset{1\leq t\leq T}{\text{col}}(\boldsymbol{x}_{dkt})\right), \quad \boldsymbol{\beta} = \underset{1\leq k\leq q-1}{\text{col}}(\boldsymbol{\beta}_k), \\
\boldsymbol{Z}_{1d} &= \underset{1\leq k\leq q-1}{\text{diag}}(\boldsymbol{1}_T), \quad \boldsymbol{Z}_{2d} = \underset{1\leq k\leq q-1}{\text{diag}}\left(\underset{1\leq t\leq T}{\text{diag}}(1)\right) = \boldsymbol{I}_{T(q-1)}, \quad \boldsymbol{1}_T = \underset{1\leq t\leq T}{\text{col}}(1).
\end{aligned}$$

To fit the model we combine the PQL method, introduced by Breslow and Clayton (1996) for estimating and predicting the β_{kr}'s, the $u_{1,dk}$'s and the $u_{2,dkt}$'s, with the REML method for estimating the variance components φ_{1k}, φ_{2k} and ϕ_k, $k = 1, \ldots, q-1$. The presented method is based on a normal approximation to the joint probability distribution of the vector $(\boldsymbol{y}, \boldsymbol{u})$. The combined algorithm was first introduced by Schall (1991) and later used by Saei and Chambers (2003), Molina et al. (2007) and Herrador et al. (2009) in applications of generalized linear mixed models to small area estimation problems. In this work, we adapt the combined algorithm to Model 3. The algorithm has two parts. In the first part the algorithm updates the values of $\boldsymbol{\beta}$, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. In the second part it updates the variance components.

3. THE mme PACKAGE

In the mme package we introduced a range of new functions that may be of interest to those conducting applied research. The nine principal new functions are summarized in the table bellow.

Function	**Description**	**Reference**
data.mme	Based on the input data this function generates some matrices that are required in subsequent calculations	Lopez-Vizcaino et al (2013a)
initial.values	Initial values for fitting algorithm to estimate the fixed and random effects and the variance components	Lopez-Vizcaino et al (2013a)
fitmodel1	Function used to fit the multinomial mixed model with one independent random effect per category of the response variable (Model 1)	Lopez-Vizcaino et al (2013a)
fitmodel2	Function used to fit the multinomial mixed model with two independent random effects for each category of the response variable: one domain random effect and another independent time and domain random effect (Model 2)	Lopez-Vizcaino et al (2013b)
fitmodel3	Function used to fit the multinomial mixed model with two independent random effects for each category of the response variable: one domain random effect and another correlated time and domain random effect (Model 3)	Lopez-Vizcaino et al (2013b)
msef	This function is used to calculate the analytic MSE for Model 1	Lopez-Vizcaino et al (2013a)
msef.it	This function is used to calculate the analytic MSE for Model 2	Lopez-Vizcaino et al (2013b)
msef.ct	This function is used to calculate the analytic MSE for Model 3	Lopez-Vizcaino et al (2013b)
mseb	Function used to calculate the bias and the MSE for the multinomial mixed effects models using parametric bootstrap	Lopez-Vizcaino et al (2013a) and Lopez-Vizcaino et al (2013b)

4. EXAMPLE TO FIT MODEL 1

The following code provides and example to fit the model 1. Note that its necesary to use a a data frame with this variables: area indicator, time indicator, sample, population, the categories of the response variable and the covariates of each category of the response variable, in this order. Indeed the package needs, also, two imput parameters, pp, that is a vector with the number of auxiliary variables in each category and k, the number of categories of the response variable. Now we use a data frame of 50 small areas and with 10 periods. In this example we use the last period. The response variable has three categories ($k = 3$), and we use one covariate for each category, then $pp = c(1, 1)$. The last three columns of the data frame are de direct estimators of the response variable.

```
> library(mme)
> datos=as.data.frame(datos)
> names(datos)

 [1] "area"       "time"       "sample"     "population" "y1"
 [6] "y2"         "y3"         "x1"         "x2"         "y11"
[11] "y22"        "y33"


> datos1=subset(datos,datos$time==10)
> k=3 #number of categories of the response variable
> pp=c(1,1) #vector with the number of auxiliary variables in each category #data
> D=nrow(datos1)
>
> #Needed matrix
> datar=data.mme(datos1,k,pp)
> mod=1 #Model 1
>
> #Initial values
> initial=initial.values(D,pp,datar,mod)
>
> #Model fit
> result=modelfit1(pp,datar$Xk,datar$X,datar$Z,initial,datar$y[,1:(k-1)],datar$n,datar$N)
>
> #Estimates of the regression parameters, their standard deviations and the p-values
> result$beta.Stddev.p.value

       Beta  Std.dev   p.value
1  1.449746 3.268338 0.6573515
2 -1.096560 2.431285 0.6519745
1 -3.001553 2.850849 0.2924040
2  1.527230 1.650165 0.3547051

>
> #Estimates the random effects, their standard deviations and the p-values
> result$phi.Stddev.p.value

      phi.est   Std.dev      p.value
[1,] 2.493575 0.5382600 3.609986e-06
[2,] 2.173891 0.4824172 6.598418e-06

>
> #Direct estimators
> dir1=datos1$y11
> dir2=datos1$y22
>
```

The following code will generate Figure 0.1 that plots direct estimator in front of the model-based estimator.

```
> #Plot direct estimator in front of model estimator
> dos.ver<-matrix(1:2,1,2)
> layout(dos.ver)
> plot(dir1,result$mean[,1],main="Small area estimator Y1",xlab="Direct estimate",
> ylab="model estimate",font.main=2,cex.main=1.5,cex.lab=1.3)
> abline(a=0,b=1)
> plot(dir2,result$mean[,2],main="Small area estimator Y2",xlab="Direct estimate",
> ylab="model estimate",font.main=2,cex.main=1.5,cex.lab=1.3)
> abline(a=0,b=1)
>
```

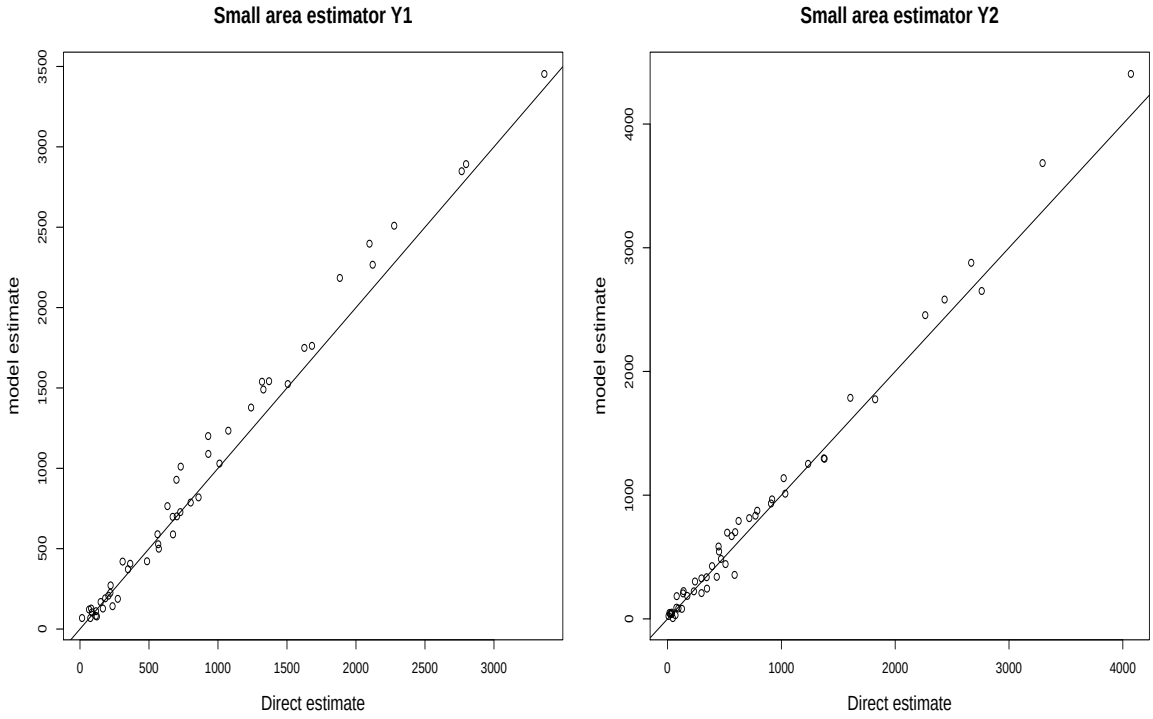

Figura 0.1: Model estimator in front of direct estimator.

```
> #Model estimator
> datos1$yest1=result$mean[,1]
> datos1$yest2=result$mean[,2]
```

The following code will generate Figure 0.2 that presents the direct and the model estimator with the areas sorted by sample size.

```
> #Plot direct estimator and model estimator with the areas sorted by sample size
> dos.ver<-matrix(1:2,1,2)
> layout(dos.ver)
> a=datos1[order(datos1[,3]),]
> g_range <- range(0,45)
```

```
> plot(a$y11/1000,type="b", col="blue",axes=FALSE, ann=FALSE)
> lines(a$yest1/1000,type="b",pch=4, lty=2, col="red")
> title(xlab="Sample size")
> axis(1,at=c(1,10,20,30,40,50),lab=c(a$muestra[1],a$muestra[10],
> a$muestra[20],a$muestra[30],a$muestra[40],a$muestra[50]))
> axis(2, las=1, at=1*0:g_range[2])
> legend("topleft", c("Direct","Model"), cex=1, col=c("blue","red"),
+    lty=1:2,pch=c(1,4), bty="n")
> title(main="Small area estimator Y1", font.main=1.2,cex.main=1)
> plot(a$y22/1000,type="b",col="blue",axes=FALSE, ann=FALSE)
> lines(a$yest2/1000,type="b",pch=4, lty=2, col="red")
> title(xlab="Sample size")
> axis(1,at=c(1,10,20,30,40,50),lab=c(a$muestra[1],a$muestra[10],
> a$muestra[20],a$muestra[30],a$muestra[40],a$muestra[50]))
> axis(2, las=1, at=1*0:g_range[2])
> legend("topleft", c("Direct","Model"), cex=1, col=c("blue","red"),
+    lty=1:2,pch=c(1,4), bty="n")
> title(main="Small area estimator Y2", font.main=1.2,cex.main=1)
```

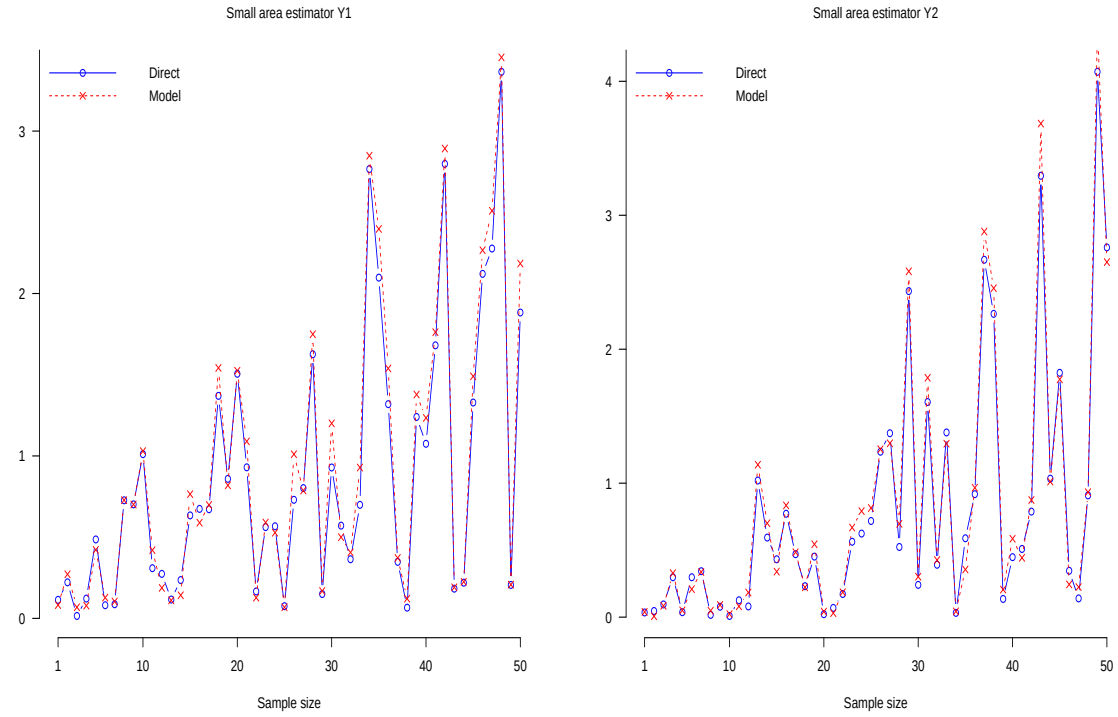

Figura 0.2: Model estimator and direct estimator with the areas sorted by sample size.

```
> #Bootstrap parametric MSE
> B=500 #Bootstrap iterations
> ss=12345 #SEED
> set.seed(ss)
> mse.pboot=mseb(pp,datar$Xk,datar$X,datar$Z,datar$n,datar$N,result,B,mod)
> cv=mse.pboot$cv
```

The following code will generate Figure 0.3 that plots the parametric bootstrap estimates of the RMSE of model-based estimates.

```
> dos.ver<-matrix(1:2,1,2)
> layout(dos.ver)
> g_range <- range(0,45)
> plot(cv[,1],type="b", col="blue",axes=FALSE, ann=FALSE)
> title(xlab="Sample size")
> axis(1,at=c(1,10,20,30,40,50),lab=c(a$muestra[1],a$muestra[10],
> a$muestra[20],a$muestra[30],a$muestra[40],a$muestra[50]))
> axis(2, las=1, at=10*0:g_range[2])
> title(main="RMSE for the estimator of Y1", font.main=1.2,cex.main=1)
> g_range <- range(0,45)
> plot(cv[,2],type="b",col="blue",axes=FALSE, ann=FALSE)
> title(xlab="Sample size")
> axis(1,at=c(1,10,20,30,40,50),lab=c(a$muestra[1],a$muestra[10],
> a$muestra[20],a$muestra[30],a$muestra[40],a$muestra[50]))
> axis(2, las=1, at=10*0:g_range[2])
> title(main="RMSE for the estimator of Y2", font.main=1.2,cex.main=1)
>
```

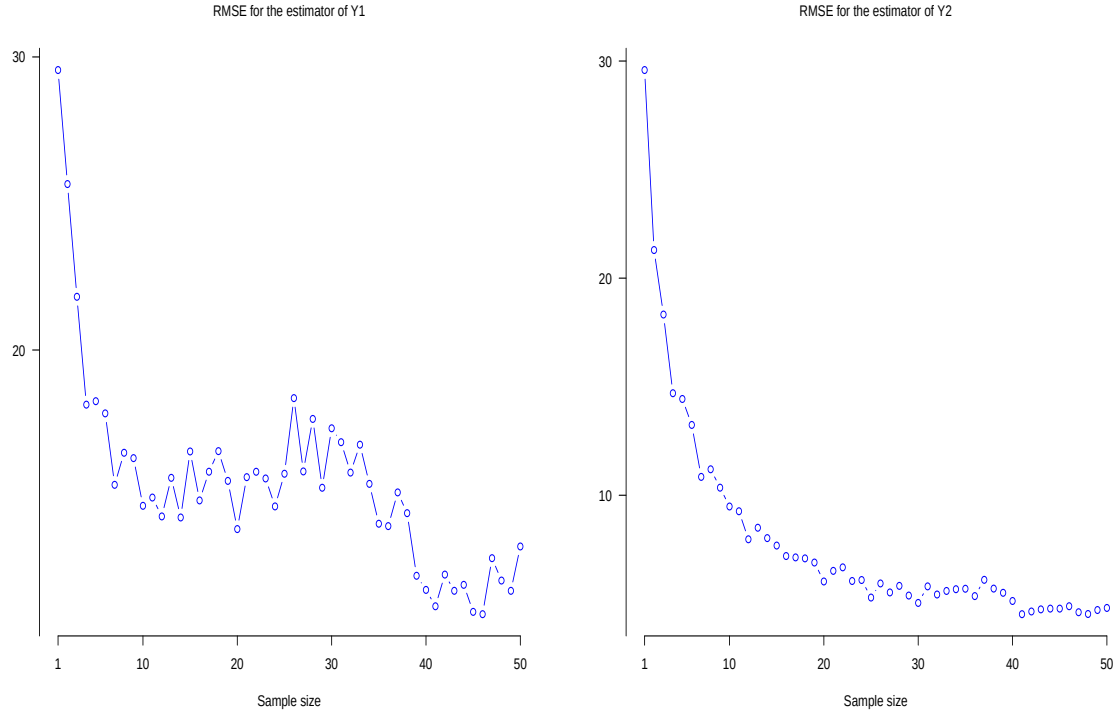

Figura 0.3: The parametric bootstrap estimates of the RMSE of model-based estimates

CONCLUSIONS

The obtained model-based estimates for all the models have low mean squared errors, especially for counties with small sample size.

REFERENCES

Breslow, N. and Clayton, D. (1996) Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88, 9-25.

López-Vizcaíno, M.E., Lombardía, M.J. and Morales, D. (2013a) Multinomial-based small area estimation of labour force indicators. Statistical Modelling, 13, 153-178.

López-Vizcaíno, M.E., Lombardía, M.J. and Morales, D. (2013b) Small area estimation of labour force indicator under a multinomial mixed model with correlated time and area effects. Submited for review.

Herrador, M., Morales, D., Esteban, M.D., Sánchez, A., Santamaría, L., Marhuenda, Y., Pérez, A. and Molina, I. (2009). Estimadores de áreas pequeñas basados en modelos para la Encuesta de Población Activa (in Spanish). Estadística Española, 51, n. 170, 133-172.

Molina, I., Saei, A. and Lombardía, M. J. (2007) Small area estimates of labour force participation under multinomial logit mixed model. The Journal of the Royal Statistical Society, series A, 170, 975-1000.

Saei, A. and Chambers, R. (2003). Small area estimation under linear an generalized linear mixed models with time and area effects. S3RI Methodology Working Paper M03/15, Southampton Statistical Sciences Research Institute.

Schall, R. (1991). Estimation in generalized linear models with random effects. Biometrika, 78, 719-727.