THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 11, Number 4/2010, pp. 303–311

REGIONAL ESTIMATION OF THE BODY MASS INDEX USING DATA FROM THE 2002/2003 FRENCH HEALTH SURVEY

Marius STEFAN[1], Adrian STREINU CERCEL[2], Daniela Adriana ION[2]

[1]University "Politehnica" of Bucharest
[2]"Carol Davila"University of Medicine and Pharmacy
E-mail: mastefan@gmail.com

In this paper we model one variable of the 2002/2003 French Health Survey (*The Body Mass Index-BMI*) in order to obtain regional estimations for the rate of overweight people. We construct the model, we derive the theoretical estimations for the parameters of interest . Then we test the fit of the model to the data and after deciding that the model is good enough we compute the estimations and their precisions. We conclude the paper with some directions for future research

Key words: Small area, Direct and indirect estimations, Borrow strength, Markov chains.

1. INTRODUCTION

The French Health Survey (FHS) is a large survey (almost 30,000 observations) taking place every ten years and collecting information on a large number of health variables (more than 200). The Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (DREES) is responsible for the statistical exploitation of the FHS data. Part of its job is to obtain national estimations for some parameters linked to a series of FHS variables. In doing this DREES is helped by INSEE, the French Statistical Institute. INSEE computes these estimations by using its well established methodology based on the classical survey sampling theory. The classical survey sampling theory centres its inference on the survey sampling distribution which is generated by the survey design, the way the sample is selected.

Recently there has been a growing demand for estimations at sub-national level. For instance, the French regional authorities are interested in estimating regional and county parameters (the French territory is divided into 22 regions, every region incorporating several counties, resulting in a total number of 96 counties). Generally, the national surveys like FHS are designed to insure an adequate level of precision at national level. When it comes to regional/county parameters one can still use the classical survey sampling theory resulting in the same formulas for the estimators and their precisions as at the national level but using the regional/county samples. These samples are composed of the observations from the national survey that come from the region or the county of interest. For a lot of such sub populations called areas or domains these observations are not numerous. This is why they are called *small areas* or *small domains*. As a consequence the estimators based on the classical survey sampling theory called *direct estimators* have not an adequate level of precision and alternative methods should be used.

The small area estimation is the new theory trying to improve the classical design-based survey sampling theory when it comes to estimating parameters at sub national levels. The key of the modern small area estimation is the modelling of the variable of interest population values and the use of the model to make inference. The model acts like a link between observations coming from different areas of the population. This is why when model-based an estimator for a sub population called *indirect estimator* uses the entire national sample, not only the sample coming from the sub population. Thus the indirect estimator is generally more precise than the direct estimator by *borrowing strength* from related areas. A detailed account of the small area estimation is given in Rao[3].

INSEE has not a methodology using small area estimation techniques. This is why DREES financed a research aimed at finding a small area methodology for regional and county parameters related to a number of variables in FHS. The results presented in this paper are part of this research which can be found in an unpublished manuscript in Stefan[4].

The present paper deals with the estimation of regional rates of overweight people aged 20 or more. As a consequence the study variable will be the *Body Mass Index* (*BMI*). In section 2 we show how to construct a model which will be used to estimate the parameters of interest. In section 3 we test the fit of the model. In section 4 we obtain the theoretical formulas of the estimators and their standard errors. Then we use the theoretical formulas to compute the estimations and their standard errors. Finally, in the last section we draw some conclusions and specify directions for future research.

2. COUNSTRUCTION OF THE MODEL

The *BMI* is a variable computed by the formula *BMI=Weight*/(*Height*^2). A person is considered as overweight if its *BMI* exceeds 25, otherwise he or she is a normal weighted individual. We computed the body mass index of all the 26930 individuals in the sample aged 20 or more and transformed this variable into a 0-1 variable where 1 indicates an overweight person and 0 a normal one. We did this given the fact that the parameters of interest are the rates of overweight people.

Then we undertook an exploratory analysis to see which are the variables that influence the *BMI*. Of all the variables in the date we retained four: the Region, the Sex and the Age indexed respectively by *i, s* and *k*. The Region has 22 values as the French territory is divided into 22 regions, the Sex has 2 values and the Age has 12 values because we transformed the Age from a continuous into a categorical variable with 12 values corresponding to the intervals [20, 24], [25, 29], [30, 34], …, [70, 74] and [75, 104], 104 being the age of the oldest person in the sample. As a result the population individuals are cross classified into $22\times 2\times 12$ cells with y_{iskl} being the value of the binary variable *BMI* for an individual *l* in cell $i\times s\times k$.

As *BMI* is binary in order to model its values we will use the Bernoulli distribution, that is $y_{iskl}\sim \text{Bernoulli}(\pi)$ for an individual *l*=1,…, N_{isk} , where N_{isk} is the population number of individuals in cell $i\times s\times k$. We mentioned that y_{iskl} depends on region *i*, sex *s* and age *k*, so it is natural to believe that taking a common π will result in a bad model. As a consequence we will consider a Bernoulli distribution of parameter π_{isk} depending on region, sex and age and the first line of the model will be.

$$y_{iskl}\sim \text{Bernoulli}(\pi_{isk}) ,$$

Next the probabilities π_{isk} will have to be modeled. A natural choice would be the logit function:

$$\text{logit}(\pi_{isk})=\beta_{1i}+\beta_{2s}+\beta_{3k} ,$$

where in order to avoid the redundancy we impose the usual corner constraints on the effects of the auxiliary variables $\beta_{21}=\beta_{31}=0$. In order to verify that the above additive specification is correct or that some interactions between the three auxiliary variables were not omitted or that some important explanatory variables were overlooked, we tested the fit of the model incorporating an error term following a logistic distribution of mean 0 and of distribution parameter σ (in this case we noticed that the logistic distribution is more appropriate than a normal one). Namely we replaced the above specification by:

$$\text{logit}(\pi_{isk})=\beta_{1i}+\beta_{2s}+\beta_{3k}+\varepsilon_{isk} ,$$
$$\varepsilon_{isk}\sim \text{Logistic}(0,\sigma) .$$

The new model takes more time to run due to the new parameters ε_{isk} and σ and provided a slightly better fit that doesn't offset for the longer estimation time. As a consequence we decided to keep the specification $\text{logit}(\pi_{isk})=\beta_{1i}+\beta_{2s}+\beta_{3k}$.

Due to the lack of information on the coefficients β_{1i},β_{2s} and β_{3k} , we conclude the hierarchy with non informative a priori distributions on these effects. We took normal laws of mean 0 and variance 1000 but a sensitivity analysis showed no influence of the a priori laws on the final estimations. As a consequence, a first hierarchical model will be:

Model 1

$$y_{iskl} \sim \text{Bernoulli}(\pi_{isk})$$
$$\text{logit}(\pi_{isk}) = \beta_{1i} + \beta_{2s} + \beta_{3k}$$
$$\beta_{1i} \sim N(0,1000), \beta_{2s} \sim N(0,1000), \beta_{3k} \sim N(0,1000)\,.$$

As in Stefan[5] or Stefan[6], we tried to improve the fit of Model 1 by taking individual probabilities for the Bernoulli laws, that is by taking π_{iskl} instead of π_{isk}. In doing this and given the results in Stefan[5] and Stefan[6] obtained for count variables we expect the same will hold true for the binary variable *BMI* meaning the new model will have a much better fit. The first line of the new model will be:

$$y_{iskl} \sim \text{Bernoulli}(\pi_{iskl})\,.$$

As for the second line, in this case we must incorporate an error term ε_{iskl} given the fact that π_{iskl} depends on the individual *l*:

$$\text{logit}(\pi_{iskl}) = \beta_{1i} + \beta_{2s} + \beta_{3k} + \varepsilon_{iskl}\,,$$
$$\varepsilon_{iskl} \sim \text{Logistic}(0,\sigma)\,.$$

We keep the same diffuse a priori distributions for the β coefficients and took a gamma distribution of parameters 0.001 for the positive parameter σ. The second model will be:

Model 2

$$y_{iskl} \sim \text{Bernoulli}(\pi_{iskl})$$
$$\text{logit}(\pi_{iskl}) = \beta_{1i} + \beta_{2s} + \beta_{3k} + \varepsilon_{iskl}$$
$$\varepsilon_{iskl} \sim \text{Logistic}(0,\sigma)$$
$$\beta_{1i} \sim N(0,1000), \beta_{2s} \sim N(0,1000), \beta_{3k} \sim N(0,1000)\,,\ \sigma \sim \text{G}(0.001, 0.001)\,.$$

Apart the logit function there is an alternative function for the parameter π of a Bernoulli distribution namely the probit function. We wanted to include in our analysis models using the probit function in order to compare the fit of these models with those of models 1 and 2 above. We kept the same specifications as in models 1 and 2 but replaced the logit by probit and the logistic distribution by the normal distribution that we found more appropriate in this case. Thus we obtain two new models denoted Model 3 and Model 4 which are given below:

Model 3

$$y_{iskl} \sim \text{Bernoulli}(\pi_{isk})$$
$$\text{probit}(\pi_{isk}) = \beta_{1i} + \beta_{2s} + \beta_{3k}$$
$$\beta_{1i} \sim N(0,1000), \beta_{2s} \sim N(0,1000), \beta_{3k} \sim N(0,1000)\,;$$

Model 4

$$y_{iskl} \sim \text{Bernoulli}(\pi_{iskl})$$
$$\text{probit}(\pi_{iskl}) = \beta_{1i} + \beta_{2s} + \beta_{3k} + \varepsilon_{iskl}$$
$$\varepsilon_{iskl} \sim N(0,\sigma^2)$$
$$\beta_{1i} \sim N(0,1000), \beta_{2s} \sim N(0,1000), \beta_{3k} \sim N(0,1000)\,,\ \sigma \sim \text{G}(0.001, 0.001)\,.$$

3. MODEL FIT

In the previous section we showed how to construct models for *BMI*. We now test the fit of these models to the data in the 2002/2003 French Health Survey. We will use the 26,930 individuals older than 20 I the sample and the tools in Stefan [5, 6] adapted to a binary variable. There are two categories of measures of fit: those that help selecting between several models and those telling if a model is good enough for making inference. In this section, for notation facility *i* designates an individual.

Let $\mathbf{y}_{obs}$ be the vector of all the observations. One way in which a model fit can be tested is to generate for every individual *i* in the sample new observations $y_{new,i}$ from the posterior predictive density $f(y_i \mid \mathbf{y}_{obs})$ and to compare the vector of these new observations $\mathbf{y}_{new}$ to $\mathbf{y}_{obs}$. It can be shown that a new value $y_{new,i}$ can be sampled from $f(y_i \mid \mathbf{y}_{obs})$ as follows: for each individual *i* we have the Markov chain $\{\pi_i^g\}$ corresponding to π_i obtained by estimating the model (see the next section). After the burn-in period the values $\{\pi_i^g\}$ come from $f(\pi_i \mid \mathbf{y}_{obs})$; we considered a burn-in period of 2000 iterations after which the chains reach convergence and used the next $G = 1{,}000$ iterations ; for each of the 1000 iterations we generated $y_{new,i}^g$ by sampling $\text{Bernoulli}(\pi_i^g)$. So $\mathbf{y}_{new}$ will be a $1000 \times 26{,}930$ matrix.

We will use three measures of discrepancy between $\mathbf{y}_{obs}$ and $\mathbf{y}_{new}$:

$$T(\mathbf{y}_{new}, \mathbf{y}_{obs}) = \sum_i \frac{(y_{new,i} - y_{obs,i})^2}{(y_{new,i} + 0.5)},$$

$$d(\mathbf{y}_{new}, \mathbf{y}_{obs}) = 2\sum_i \left[(y_{obs,i} + 0.5)\log\frac{(y_{obs,i} + 0.5)}{(y_{new,i} + 0.5)} + (1.5 - y_{obs,i})\log\frac{(1.5 - y_{obs,i})}{(1.5 - y_{new,i})} \right],$$

and

$$D(\mathbf{y}_{new}, \mathbf{y}_{obs}) = \sum_i (E(y_{new,i} \mid \mathbf{y}_{obs}) - y_{obs,i})^2 + \sum_i V(y_{new,i} \mid \mathbf{y}_{obs}).$$

Their a posteriori means will be given by:

$$E(T \mid \mathbf{y}_{obs}) = \frac{1}{G}\sum_g \sum_i \frac{(y_{new,i}^g - y_{obs,i})^2}{(y_{new,i}^g + 0.5)},$$

$$E(d(\mathbf{y}_{new}, \mathbf{y}_{obs}) \mid \mathbf{y}_{obs}) = 2\frac{1}{G}\sum_g \sum_i \left[(y_{obs,i} + 0.5)\log\frac{(y_{obs,i} + 0.5)}{(y_{new,i}^g + 0.5)} + (1.5 - y_{obs,i})\log\frac{(1.5 - y_{obs,i})}{(1.5 - y_{new,i}^g)} \right],$$

$$E(y_{new,i} \mid \mathbf{y}_{obs}) = \frac{1}{G}\sum_g y_{new,i}^g \text{ and } V(y_{new,i} \mid \mathbf{y}_{obs}) = \frac{1}{G}\sum_g (y_{new,i}^g - E(y_{new,i} \mid \mathbf{y}_{obs}))^2.$$

The numerical results can be found in Table 1. We can notice that Model 2 provides a much better fit than the other three models. The large improvement was achieved by incorporating individual Bernoulli probabilities π_{iskl} the price to pay being a model that takes longer to estimate but offering a much better fit to the data.

Another interesting feature in Table 1 is that a larger number of parameters in a model doesn't necessarily guaranty for a better fit. Model 4 like Model 2 has individual probabilities π_{iskl} but as a link function it uses probit instead of logit. Nonetheless the values of the measures of discrepancy are slightly less than those under Model 1.

Table 1

A posteriori means of the measures of discrepancy

Measure	Model 1	Model 2	Model 3	Model 4
$E(T(\mathbf{y}_{new}, \mathbf{y}_{obs}) \mid \mathbf{y}_{obs})$	6979.234	2777.125	7187.943	6939.522
$E(d(\mathbf{y}_{new}, \mathbf{y}_{obs}) \mid \mathbf{y}_{obs})$	11605.21	4589.38	12270.16	11457.77
$D(\mathbf{y}_{new}, \mathbf{y}_{obs})$	5281.76	2088.72	5584.39	5214.65

As a consequence we conclude that Model 2 is the best among the four models constructed in the previous section. We now have to check if Model 2 is well adapted to the data in the sample. In order to do this we introduce two measures of discrepancy between a vector of observations **y** and a vector of Bernoulli probabilities $\boldsymbol{\pi}$. The first is called Deviance and is given below:

$$\text{Deviance }(\mathbf{y},\boldsymbol{\pi}) = -2\sum_i \log(f(y_i \mid \pi_i)),$$

where $f(y_i \mid \pi_i)$ is the density function of the Bernoulli(π_i) distribution. The second denoted $\text{Dis}(\mathbf{y}_{obs},\boldsymbol{\pi})$ was already used in Stefan[5,6] but below its formula is adapted for the case when the variable under study is a 0-1 variable:

$$\text{Dis}(\mathbf{y},\boldsymbol{\pi}) = \sum_i \frac{(y_i - \pi_i)^2}{\pi_i(1-\pi_i)}.$$

Both $\text{Deviance}(\mathbf{y},\boldsymbol{\pi})$ and $\text{Dis}(\mathbf{y},\boldsymbol{\pi})$ allow one to verify if a model is good enough by estimating the probabilities that $\text{Deviance}(\mathbf{y}_{new},\boldsymbol{\pi}) \geq \text{Deviance}(\mathbf{y}_{obs},\boldsymbol{\pi})$ and $\text{Dis}(\mathbf{y}_{new},\boldsymbol{\pi}) \geq \text{Dis}(\mathbf{y}_{obs},\boldsymbol{\pi})$. The probabilities can be estimated by:

$$\hat{p}_1 = \frac{1}{G}\sum_g I[\text{Deviance}(\mathbf{y}^g_{new},\boldsymbol{\pi}^g) \geq \text{Deviance}(\mathbf{y}_{obs},\boldsymbol{\pi}^g)],$$

and

$$\hat{p}_2 = \frac{1}{G}\sum_g I[\text{Dis}(\mathbf{y}^g_{new},\boldsymbol{\pi}^g) \geq \text{Dis}(\mathbf{y}_{obs},\boldsymbol{\pi}^g)].$$

Values of $\hat{p}_1$ and $\hat{p}_2$ close to 0.5 indicates a good fit while values less than 0.1 or higher than 0.9 indicates a bad fit and a model that has to be rejected. In Table 2 we present the values of $\hat{p}_1$ and $\hat{p}_2$ under the four models. Table 2 shows that Model 2 is well fit to the data and can be used to estimate the regional rates of overweight people.

Table 2

Values of $\hat{p}_1$ and $\hat{p}_2$

Probability	Model 1	Model 2	Model 3	Model 4
$\hat{p}_1$	0.068	0.518	0.999	0.335
$\hat{p}_2$	0.034	0.5	0.994	0.248

Under each model we can plot $\text{Deviance}(\mathbf{y}_{obs},\boldsymbol{\pi})$ against $\text{Deviance}(\mathbf{y}_{new},\boldsymbol{\pi})$ represented in Fig. 1 and $\text{Dis}(\mathbf{y}_{obs},\boldsymbol{\pi})$ against $\text{Dis}(\mathbf{y}_{new},\boldsymbol{\pi})$ represented in Fig. 2. Under a good model half of the points are under the $y = x$ line and half are above the $y = x$ line. Figures 1 and 2 show that under Model 2 the points are equally distributed under and above the $y = x$ line.

4. PARAMETERS ESTIMATION

In this section we will estimate the parameters of interest which are the regional rates of overweight people denoted by p_i, $i = 1,...,22$. We will use the principles of Bayesian statistics which consist of estimating a parameter by its posterior mean and the precision of this estimation by its posterior variance or standard error. Thus, our estimators will be $\hat{p}_i = E(p_i \mid \mathbf{y}_{obs})$ and their variances $V(\hat{p}_i) = V(p_i \mid \mathbf{y}_{obs})$. So the focus of the Bayesian statistics is the posterior distribution of a parameter that is the distribution of the parameter after the sample was selected. Generally this distribution cannot be obtained in a closed form but in order to have $E(p_i \mid \mathbf{y}_{obs})$ and $V(p_i \mid \mathbf{y}_{obs})$ it is sufficient to get a sample from the posterior distribution. Then $E(p_i \mid \mathbf{y}_{obs})$ and $V(p_i \mid \mathbf{y}_{obs})$ will be approximated by the sample mean and the sample variance respectively.

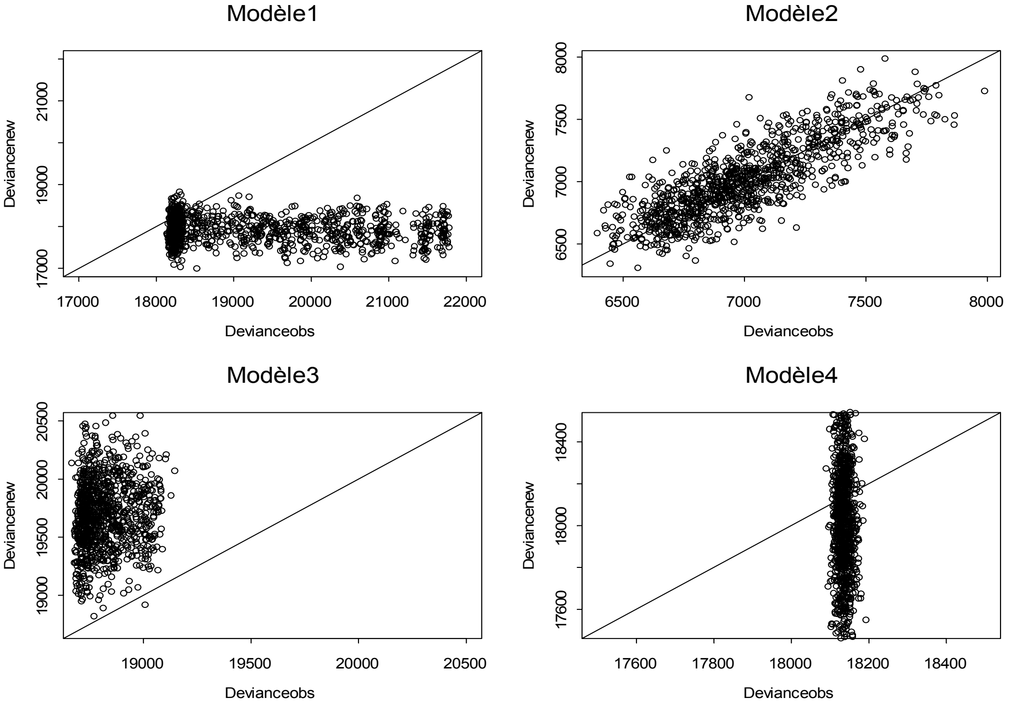

Fig. 1 – Deviance$(\mathbf{y}_{obs},\boldsymbol{\pi}^g)$ *vs.* Deviance$(\mathbf{y}^g_{new},\boldsymbol{\pi}^g)$.

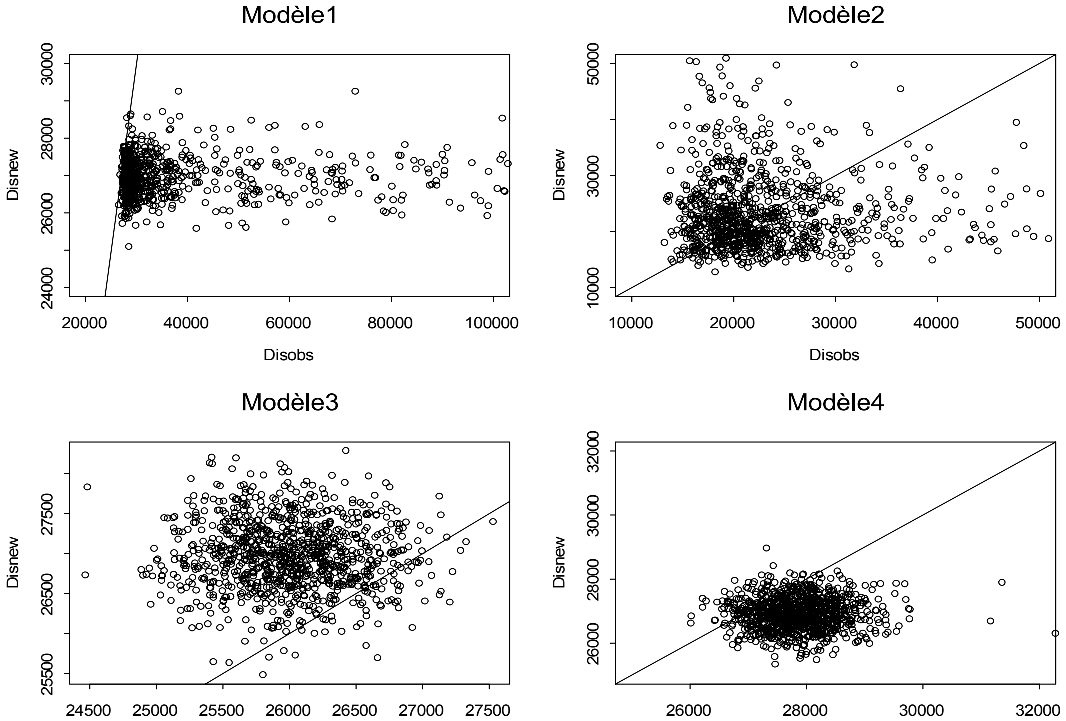

Fig. 2 – Dis$(\mathbf{y}_{obs},\boldsymbol{\pi}^g)$ *vs.* Dis$(\mathbf{y}^g_{new},\boldsymbol{\pi}^g)$.

Like in Stefan [5, 6] we divide the individuals in a region in two parts: the sampled obs_i and the non sampled $nobs_i$ individuals. Then the rate of overweight people in region *i* will be:

$$p_i = \frac{1}{N_i}\Big(\sum_s\sum_k\sum_{l\in obs_i} y_{iskl} + \sum_s\sum_k\sum_{l\in nobs_i} y_{iskl}\Big). \qquad (1)$$

Taking the a posteriori mean of (1), one will get:

$$\hat{p}_i = E(p_i \mid \mathbf{y}_{obs}) = \frac{1}{N_i}[\sum_s \sum_k \sum_{l \in obs_i} y_{iskl} + \sum_s \sum_k \sum_{l \in nobs_i} E(y_{iskl} \mid \mathbf{y}_{obs})]\,. \tag{2}$$

It can be proved (see Stefan[4, p. 95]) that for an individual $l_0 \in nobs_i$, $E(y_{iskl} \mid \mathbf{y}_{obs}) = \mathrm{pi}_{isk}$, where $\mathrm{pi}_{isk} = E(\pi_{iskl_0} \mid \mu_{isk}, \sigma)$. Denoting N_{isk} and n_{isk} the population and the sampled individuals in cell $i \times s \times k$ respectively (2) will become:

$$\hat{p}_i = \frac{1}{N_i}[\sum_s \sum_k \sum_{l \in obs_i} y_{iskl} + \sum_s \sum_k (N_{isk} - n_{isk}) E(\mathrm{pi}_{isk} \mid \mathbf{y}_{obs})]\,. \tag{3}$$

Let's denote $\mu_{isk} = \beta_{1i} + \beta_{2s} + \beta_{3k}$. By estimating Model 2 we get Markov chains for the parameters of the model denoted $\beta^g_{1i}, \beta^g_{2s}, \beta^g_{3s}$ and σ^g (the Markov chain for μ_{isk} will be $\mu^g_{isk} = \beta^g_{1i} + \beta^g_{2s} + \beta^g_{3s}$). The parameters pi_{isk} are not part of Model 2 so by estimating the model we will not get Markov chains for pi_{isk}. Nonetheless given how pi_{isk} were defined one can construct Markov chains for them. Let $f(x \mid \sigma)$ be the density function of a logistic distribution of mean 0 and dispersion parameter σ. Then:

$$\mathrm{pi}_{isk} = E(\pi_{iskl_o} \mid \mu_{isk}, \sigma) = \int_{-\infty}^{+\infty} \frac{e^{\mu_{isk}+x}}{1+e^{\mu_{isk}+x}} f(x \mid \sigma)\mathrm{d}x\,.$$

The integral cannot be computed exactly thus we will approximate it as follows: for each Markov chain iteration $g = 1,\ldots,G$ obtained by estimating Model 2 we generate $x^g_a, a = 1,\ldots,10{,}000$ values from the logistic distribution of mean 0 and distribution parameter σ^g; then we take the g iteration of pi_{isk} Markov chains as:

$$\mathrm{pi}^g_{isk} = \frac{1}{10{,}000} \sum_{a=1}^{10000} \frac{e^{\mu^g_{isk}+x^g_a}}{1+e^{\mu^g_{isk}+x^g_a}}\,.$$

Thus (3) will become:

$$\hat{p}_i = \frac{1}{N_i}[\sum_s \sum_k \sum_{l \in obs_i} y_{iskl} + \frac{1}{G} \sum_g \sum_s \sum_k (N_{isk} - n_{isk}) \mathrm{pi}^g_{isk}]\,. \tag{4}$$

In a similar manner as in Stefan [4, p. 96] by taking the a posteriori variance of (1) one will get

$$V(\hat{p}_i) = V(p_i \mid \mathbf{y}_{obs}) = \frac{1}{N_i^2}\left[\frac{1}{G}\sum_g \sum_s \sum_k (N_{isk} - n_{isk})\left(\mathrm{pi}^g_{isk} - \mathrm{pi}^{g\ 2}_{isk}\right) + \frac{1}{G}\sum_g \left(\sum_s \sum_k (N_{isk} - n_{isk})\mathrm{pi}^g_{isk}\right)^2 - \right. \\ \left. - \frac{1}{G}\sum_g \sum_s \sum_k (N_{isk} - n_{isk})\mathrm{pi}^g_{isk})^2\right]. \tag{5}$$

We estimated Model 2 and obtained the Markov chains of $\beta^g_{1i}, \beta^g_{2s}, \beta^g_{3s}$ and σ^g. For each parameter we run three Markov chains in order to monitor the convergence. We noticed that after a burn-in period of 2,000 iterations the three Markov chains converged. In order to save space in Fig. 3 we show the three Markov chains only for parameter σ of the model.

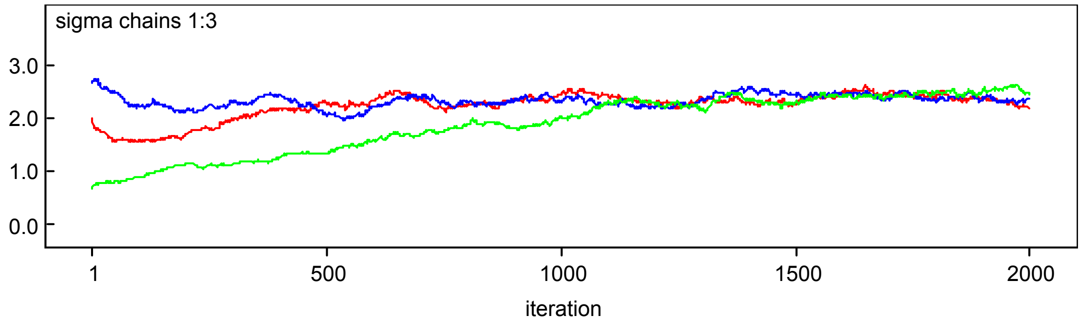

Fig. 3 – Markov chains of σ.

As a consequence for each parameter we dropped the first 2,000 iterations an run the chains for 2,000 more iterations. Thus we used 6000 iterations for computing $\hat{p}_i$ and $V(\hat{p}_i)$ based on (4) and (5). The results are showed in Table 3 below:

Table 3

Regional estimations of overweight people rates (%) age ≥20

Region	Estimation	Standard Error	Quantiles 0.025;0.5;0.975
Ile de France	9.05	0.35	8.37;9.04;9.75
Champagne-Ardenne	13.5	0.79	11.98;13.48;15.12
Picardie	14.38	0.82	12.76;14.37;16.03
Haute-Normandie	12.76	1.47	9.94;12.74;15.76
Centre	14.18	1.38	11.62;14.12;17.06
Basse-Normandie	13.94	1.51	11.16;13.90;17.13
Bourgogne	11.74	1.33	9.33;11.69;14.51
Nord Pas de Calais	15.47	0.70	14.13;15.48;16.84
Loraine	14.85	1.23	12.51;14.81;17.38
Alsace	15.28	1.46	12.54;15.25;18.22
Franche Comté	10.16	1.37	7.74;10.06;12.99
Pays de la Loire	9.90	0.83	8.34;9.88;11.61
Bretagne	10.12	0.98	8.30;10.08;12.15
Poitou Charente	11.32	1.34	8.86;11.25;14.11
Aquitaine	9.73	0.96	8.02;9.66;11.82
Midi-Pyrénées	9.27	1.00	7.45;9.21;11.39
Limousin	10.85	1.99	7.64;10.62;15.23
Rhône Alpes	9.27	0.81	7.86;9.21;11.02
Auvergne	11.87	1.73	8.76;11.76;15.54
Languedoc-Roussillon	9.05	1.09	7.12;8.98;11.38
PACA	7.76	0.49	6.83;7.76;8.75
Corse	18.57	6.10	8.35;18.08;31.43

In Fig. 4 we plotted the coefficients of variation versus regional sample size. In Stefan [5, 6] we could compare our results based on the small area theory with the values INSEE obtained by using their methodology based on the sampling design. Unfortunately for the *BMI* INSEE didn't provide any results so we couldn't plot in Fig. 4 the INSEE coefficients of variation.

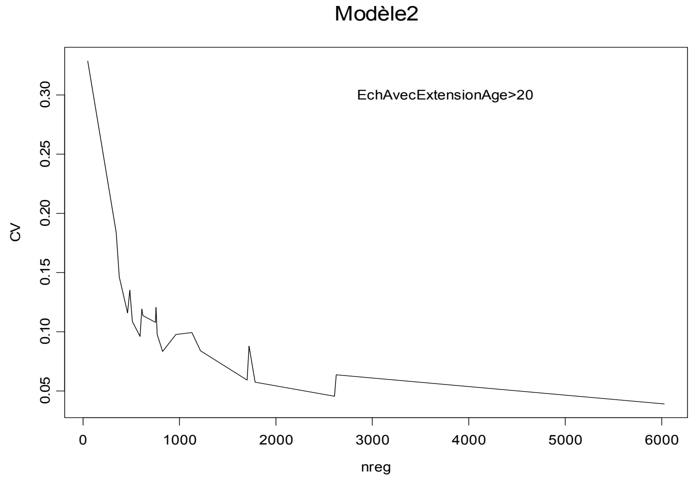

Fig. 4 – Coefficient of variation *versus* size of the regional sample size.

5. CONCLUSION

The objective of this paper was to estimate the regional rates of overweight people based on data in the 2002/2003 French Health Survey and using a small area methodology. The key of the small area theory is to model the data, test the fit of the model to the data and derive the numerical estimations. We showed how possible models can be constructed and how to choose between several options. Then we showed how a posteriori mean and variance can be derived for the parameters of interest. In this paper the variable under study is a binary variable so we underlined the necessary modifications to the methodology compared to Stefan[5, 6] where the interest variables were count variables.

In deriving the formulas (4) and (5) we supposed that the cell sample sizes n_{isk} are non-random. In practice this is generally not true. In the classical survey sampling theory computations using random n_{isk} are not feasible, that's why under such circumstances analyses are conditional on the realized sample sizes. In a full hierarchical Bayesian context Oleson and al.[2] proposed a model accounting for random sample sizes and also random population sample sizes. Based on their paper we will extend our present work.

Survey sampling are generally characterized by nonresponse and FHS is no exception. If not properly accounted for the nonresponse can lead to biased estimation. In our paper we supposed that there is complete response. In fact we removed the individuals or which we couldn't compute the *BMI* and performed our analysis on the remaining ones. Nandram et al. [1] and the references therein constitute a large literature to see how the nonresponse in FHS can be properly dealt with in a full hierarchical Bayesian context.

ACKNOWLEDGEMENTS

This work is supported by the Sectoral Operational Program for Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/89/1.5/S/64109.

REFERENCES

1. NANDRAM, B., COX, L., CHOI, W.J., *Bayesian analysis of nonignorable missing categorical data: An application to bone mineral density and family income*, Survey Methodology, **31**, *2*, pp. 213–225, 2005.
2. OLESON, J., HE, C., SUN, D., SHERIFF, S., *Bayesian estimation in small areas when sampling design strata differ from the study domains*, Survey Methodology, **33**, *2*, pp. 173–185, 2007.
3. RAO, J.N.K., *Small area estimation*, John-Wiley & Sons, Hoboken-New Jersey, 2003.
4. STEFAN, M., *Enquête décennale santé 2002/2003: Estimation petits domaines*, Rapport final de recherche, DREES, Paris, 2006.
5. STEFAN, M., *Hierarchical Bayesian estimation of the number of visits to the generalist in 2002/2003 French Health Survey*, Romanian Journal of Economic Forecasting, **53**, *2*, pp. 67–91, 2008.
6. STEFAN, M., *Small area estimation of the number of visits to the specialist in 2002/2003 French Health Survey,* Proceedings of the Romanian Academy, Series A, **54**, *2*, 2009.

Received August 11, 2010