Method used to perturb data for Religion and Fertility: The French Connection

Thomas Baudin*

December 17, 2014

*Centre de Recherche en Démographie et Sociétés - Université catholique de Louvain and EQUIPPE, Université de Lille.

Perturbation of variables

To preserve the confidentiality of the data used in this study, 3 percent of the observations have been randomly transformed. All the variables of the original sample are discrete variables. The transformed variables are also discrete. I have first generated a random variable T thanks to the stata command *rnormal*. I have then selected the observations in the 3 first percentiles of T and have operated transformations on these 3%.

Depending on their nature, the way variables can be perturbed is different. I have used two alternative methods:

1. *Discrete randomization*

   This technic is applied to the variables Fertility (dependent variable), Female_Income, Male_Income, Attendance_to_Religious_Services, Parental_Fertility, age, Family_Ties and Estimated_religiousness. For instance, the variable $DR\_Fertility$ is the dependent variable used in the file 'command.do'. For the observations which have been perturbed, I have added to the number of children a random term drawn from a normal distribution $\mathcal{N}(0, 0.5)$. I have then construct the Variable $DR\_Fertility$ as the closest integer to the result obtain previously, thanks to the command "round" in Stata. To prevent unrealistic values of $DR\_Fertility$, I have also used the following command on Stata:

   - replace $DR\_Fertility = 0$ if $DR\_Fertility < 0$
   - replace $DR\_Fertility = 10$ if $DR\_Fertility > 10$ (10 being the highest observed value in the original dataset)

   A similar technic has been used for the other variables. For DR_Below_28 and DR_Over_45, I have compared the perturbed age variable to 28 and 45.

2. *Dummy variables*

   Dummy variables of the original datafile could not be transformed using the previous technics. This is the case of $Never\_Married$, $Small\_Town$, $Live\_In\_Paris$, $Higher\_School$. For these variables, I have operated random permutations among the three first percentiles of T. I have operated the following transformation:

$$DR\_variable = \begin{cases} True\_variable & with\ \ probability\ \ 0.5 \\ Alternative\_value & with\ \ probability\ \ 0.5 \end{cases}$$

It means that, among the perturbed observations, each dummy variable has a probability equal to one half to change its value.

Some variables which have been transformed using discrete randomization can admit lower z-statistics and higher p-values. Intuitively, because the range of values these variables can admit is limited, the variance of these variables did not change significantly while permutations have been operated, the significance of these variables can only decrease. Nevertheless, the low proportion of observations which have been transformed makes this phenomenon a very limited one.