Forecasting – 7.1 Aggregation

Michel Bierlaire

Stratified sampling and sample enumeration

Given a choice model $P(i|x_n)$ that has been estimated from data, the predicted share of the population of N individuals choosing alternative i is given by

$$W(i) = \frac{1}{N}\sum_{n=1}^{N} P(i|x_n;\theta) = \mathrm{E}\left[P(i|x_n;\theta)\right]. \tag{1}$$

In practice, the population is often too large for the analyst to have access to each x_n vector, or even to their distribution. We introduce here a practical method to estimate $W(i)$ called *sample enumeration*.

The idea is to draw a sample from the population. It is actually possible to use the same sample used for the estimation of the parameters, but only if it consists of revealed preference data, that is data where the actual choice has been observed. Stated preferences data, where respondents are exposed to hypothetical scenarios, cannot be used for aggregation and prediction.

It is usually infeasible in practice to collect a purely random sample, where each individual in the population has exactly the same probability to be considered. A method called *stratified random sampling* is more realistic to implement.

It consists in partitioning the population into G mutually exclusive and collectively exhaustive groups, each called a stratum. This segmentation is motivated by the logistic of the data collection, and by the objectives of the survey. For instance, each stratum can be a geographical territory (a city, a county, etc.), where a local coordinator can be assigned. Or the partition can be organized by age, because we are interested in the impact of age on the choice behavior, like in the simple example presented in the beginning of the course.

Once the partitioning is defined, we sample S_g observations in each stratum g, using simple random sampling. The total size of the sample is $S = \sum_{g=1}^{G} S_g$.

Contrarily to simple random sampling, stratified sampling generates samples where some groups are proportionally more represented in the sample than they are in the population. This has to be taken into account when infering quantities related to the population from the same quantities calculated with the sample.

To do that, each group is associated with a weight:

$$\omega_g = \frac{N_g}{N}\frac{S}{S_g} = \frac{\text{share of persons in segment } g \text{ in the population}}{\text{share persons in segment } g \text{ in the sample}}. \quad (2)$$

As each individual n belongs to exactly one stratum g, we define

$$\omega_n = \sum_{g=1}^{G} \delta_{ng}\omega_g, \quad (3)$$

where $\delta_{ng} = 1$ if individual n belongs to stratum g, and 0 otherwise.

Therefore, an estimate of the predicted share (1) of the population choosing alternative i is given by

$$\widehat{W}(i) = \frac{1}{S}\sum_{n=1}^{S} \omega_n P(i|x_n;\theta). \quad (4)$$