Introduction

In many applications, individual data are not widely available, if at all, especially when small geographic areas or vulnerable data are involved. It is more frequent to find aggregated data instead. Two problems arise:

- Can these aggregated data be used to infer individual behavior (we will use the term of ecological inference)?
- How to manipulate categorical variables (which become compositional variables once aggregated)?

Compositional Data

A composition of D components is a vector x of the simplex S^D defined as

\[ S^D = \left\{ x = [x_1, x_2, \ldots, x_D] : x_i \geq 0, i = 1, 2, \ldots, D, \sum_{i=1}^{D} x_i = 1 \right\} \] (1)

This sample space is a (D – 1)-dimensional subset of \( \mathbb{R}^D \). For example, \( S^3 \) is a triangle (ternary diagram).

Aitchison (1986) [1] introduces operations on the simplex (sum and scalar multiplication). The inner product of two compositions is defined as:

\[ \langle x, y \rangle_a = \sum_{i=1}^{D} \frac{x_i}{g(x)} \log \frac{y_i}{g(y)} = \frac{1}{D} \sum_{i<j} \log \frac{x_i}{x_j} \log \frac{y_j}{y_i} \] (2)

The additive log-ratio transformation:

\[ \text{alr}(x) = [\log \frac{x_1}{x_D}, \log \frac{x_2}{x_D}, \ldots, \log \frac{x_{D-1}}{x_D}] \] (3)

The centered log-ratio transformation:

\[ \text{clr}(x) = [\log \frac{x_1}{g(x)}, \log \frac{x_2}{g(x)}, \ldots, \log \frac{x_D}{g(x)}] \] (4)

The isometric log-ratio transformation:

\[ \text{ilr}(x) = \langle [x, e_1], [x, e_2], \ldots, [x, e_{D-1}] \rangle \] (5)

Where \( e_1, e_2, \ldots, e_{D-1} \) represent an orthonormal basis of \( S^D \) and \( g(x) = (x_1 x_2 x_3)^{1/2} \) is the geometric mean.

Econometrics with Compositional Data

Regression analysis is generally used in statistics to study the relation between a response Y and a set of explanatory covariate \( x = (x_1, \ldots, x_D) \) as

\[ y = \beta_0 + \beta^T x + \epsilon \] (6)

The standard linear model is unreasonable with compositional data.

- Compositions contain only relative information: least squares methods examine continuous variables that are linked with an absolute relationship.
- When one part is altered, another is altered: the interpretation of the linear regressions coefficients assumes that "all other things being equal".
- Composition covariates involves multicollinearity problems.
- The simplex has particular geometric properties and operations: most common statistical procedures are developed in the usual Euclidean geometry.

The ilr transformation is an isometric transformation from the simplex to the real space with usual Euclidean geometry. It leads to apply the standard models to the compositions as

\[ y = \beta_0 + \beta^T \text{ilr}(x) + \epsilon \] (7)

Using the estimation of \( \beta \) is possible to find \( b \) with the ilr-inverse transformation. The composition \( b \) can be interpreted as the slope parameter of the standard regression. If \( x \) differs by \( \frac{1}{|b|} \) in the direction of \( b \), then \( y \) differs by \( |b| \).

An example

Analysis of median income according to diploma levels in Paris by iris.

Median income increases in the SUP direction and decreases in the BEP direction. In other words, when the proportion of SUP increases or the proportion of BEP decreases in an area, then median income increases.

Ecological Inference

Ecological inference consists of making inference of the individual behavior using aggregated group-level data. Group-level data are usually less reliable and subject to bias and imprecision that may involve mistake of inference. Robinson (1950) [5] pointed out to be careful with the use of aggregate data to study individuals ("ecological fallacy").

The aim is to find \( \beta_0^j \) and \( \beta_j^j \) which are the probability of being republican conditionally of being a man or a woman in the area \( j \).

Goodman regression (1953) [3] suggests these two probabilities are constant over area and could be estimated by least squares as

\[ p_j = \beta_0^j + \beta_j^j (1 - \phi_j) \] (8)

The method of bound (Duncan and Davis (1953) [2]) is to find the minimum and maximum of the probability with

\[ \max \left\{ 0, \frac{p_j - (1 - \phi_j)}{1 - \phi_j} \right\} \leq \beta_j^j \leq \min \left\{ \frac{p_j}{1 - \phi_j}, 1 \right\} \] (9)

\[ \max \left\{ 0, \frac{p_j - \phi_j}{1 - \phi_j} \right\} \leq \beta_0^j \leq \min \left\{ \frac{p_j}{1 - \phi_j}, 1 \right\} \] (10)

King’s solution (1997) [4] is an improvement in ecological inference by combining the Goodman method and the information of the bounds to improve inference. \( \beta_0^j \) and \( \beta_j^j \) are linked by tomography line within the unit square as

\[ \beta_0^j = \frac{p_j}{1 - \phi_j} - \frac{\phi_j}{1 - \phi_j} \beta_j^j \] (11)

King suggests three assumptions:

- \( \beta_0^j \) and \( \beta_j^j \) are in a single cluster which is generated by a truncated normal bivariate distribution conditional of \( \phi_j \).
- Absence of spatial autocorrelation: the number of exposed cases of an area not depends of the number exposed cases in the others area.
- Absence of aggregation bias: independence between the regressors \( \phi_j \) and the parameters \( \beta_0^j \) and \( \beta_j^j \).

References