Analysing the impacts of socio-economic factors in French departmental elections with CODA

Nguyen, Laurent, Thomas-Agnan, Ruiz-Gazen

Toulouse School of Economics

12 juin 2018

Table of contents

Application Context

1. Modelling vote shares vectors in a multiparty system
2. Evaluate the impact of socio-economic factors on the votes shares in French departmental election (2015)

Distributions on the simplex

A composition is a vector of D parts of some whole which carries relative information. A D-composition $\mathbf{x}$ lies in the simplex $\mathbf{S}^D$.

$$\mathbf{S}^D = \left\{ \mathbf{x} = (x_1, ..., x_D) : x_j > 0, j = 1, ..., D; \sum_{j=1}^{D} x_j = 1 \right\}$$

The possible distributions on the simplex :

1. The **Dirichlet** distribution and its generalizations
2. The **logistic-normal** distribution (CODA approach)
3. The **logistic-Student** distribution
4. The **Aitchison** distribution
5. The **Compound** distributions (Normal-Multinomial and Dirichlet-Multinomial distributions)
6. The **hyperspherical** distribution

The CODA philosophy : working in coordinates

- Transform the parts into coordinates using a one-to-one mapping ϕ from $\mathbf{S}^D$ to $\mathbb{R}^{D-1}$
- Transport the vector space structure from $\mathbb{R}^{D-1}$ back to $\mathbf{S}^D$
- Transport the Euclidean structure of $\mathbb{R}^{D-1}$ back to $\mathbf{S}^D$

As of today, the only map ϕ that leads to a structure compatible with the principles of CODA (permutation invariance, interpretability) is the ILR transformation.

Vector space structure of the simplex

1. Perturbation as compositional sum
   $\mathbf{x} = (x_1, ..., x_D),\ \mathbf{y} = (y_1, ..., y_D),\ \mathbf{x}, \mathbf{y} \in \mathbf{S}^D$,

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, ..., x_D y_D] \text{ where } \mathcal{C}[\mathbf{x}] = \left[\frac{x_1}{\sum_{j=1}^D x_j}, \cdots, \frac{x_D}{\sum_{j=1}^D x_j}\right]$$

2. Powering as compositional scalar multiplication
   $\lambda \in \mathbb{R}$, $\mathbf{x} \in \mathbf{S}^D$

$$\lambda \odot \mathbf{x} = \mathcal{C}[x_1^\lambda, ..., x_D^\lambda]$$

Let $\boxdot$ be the compositional matrix product, which corresponds through the transformation to the matrix product in the Euclidean geometry

$$\mathbf{B} \boxdot \mathbf{x} = \mathcal{C}\left(\prod_{j=1}^D x_j^{B_{1j}}, \cdots, \prod_{j=1}^D x_j^{B_{Dj}}\right)^T$$

where $\mathbf{B}$ is a $D \times D$ matrix.

The simplex geometry : Aitchison geometry

1. The compositional inner product (C-inner product) of **x** and **y** in $\mathbf{S}^D$ is defined by

$$\langle \mathbf{x}, \mathbf{y} \rangle_c = \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \log \frac{x_i}{x_j} \cdot \log \frac{y_i}{y_j} = \sum_{i=1}^{D} \log \frac{x_i}{g(\mathbf{x})} \cdot \log \frac{y_i}{g(\mathbf{y})}$$

2. The compositional distance (C-distance) between **x** and **y** is defined by

$$\mathrm{d}_c(\mathbf{x}, \mathbf{y}) = \left(\frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right)^2 \right)^{1/2}$$

$$= \left(\sum_{i=1}^{D} \left(\log \frac{x_i}{g(\mathbf{x})} - \log \frac{y_i}{g(\mathbf{y})} \right)^2 \right)^{1/2} \quad \text{where} \quad g(\mathbf{x}) = \sqrt[D]{x_1 x_2 \ldots x_D}$$

Log-ratio approach : contrast matrix

In order to define the ILR transformation, we first need to introduce the contrast matrices.
A contrast matrix (dimension $D \times (D-1)$) is associated to any orthonormal basis $(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$ of $\mathbf{S}^D$ by

$$\mathbf{V}_D = \mathrm{clr}(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1}),$$

where

$$\mathrm{clr}(\mathbf{x}) = \left(\ln\left(\frac{x_i}{g(\mathbf{x})}\right)\right) \quad \text{where} \quad g(\mathbf{x}) = \sqrt[D]{x_1 x_2 \ldots x_D}, \; i = 1, \ldots, D$$

Log-ratio approach : ILR transformation

The Isometric Log-Ratio Transformation (ilr) associated to a contrast matrix $\mathbf{V}_D$ is defined by

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{V}_D^T \ln(\mathbf{x})$$

The ilr transformation is an isometry whereas alternative transformations such as clr or alr are not.

For $\mathbf{V}_D^T = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \end{bmatrix}$ we get

$$\mathbf{V}_D^T \log(x_1, x_2, x_3) = (\sqrt{\frac{2}{3}} \log(\frac{\sqrt{x_1 x_2}}{x_3}), \frac{1}{\sqrt{2}} \log(\frac{x_1}{x_2}))$$

Example : opposition between SUP and the rest, and between <BAC and BAC.

Expected value in the simplex : definition

The expected value $\mathbb{E}^{\oplus}\mathbf{Y}$ of a simplex-valued random composition $\mathbf{Y} \in \mathbf{S}^D$ (Pawlowsky) is defined by

$$\underset{\mathbf{z}\in\mathbf{S}^D}{\operatorname{argmin}} \mathbb{E}(d_c^2(\mathbf{Y},\mathbf{z}))$$

It is equal to

$$\mathbb{E}^{\oplus}\mathbf{Y} = \mathcal{C}(\exp(\mathbb{E}\log\mathbf{Y})) = \mathrm{clr}^{-1}(\mathbb{E}\mathrm{clr}(\mathbf{Y})) = \mathrm{ilr}^{-1}(\mathbb{E}\mathrm{ilr}(\mathbf{Y})) = \mathrm{ilr}^{-1}(\mathbb{E}\mathbf{Y}^*)$$

where $\mathbf{Y}^* = \mathrm{ilr}(\mathbf{Y})$.

The additive logistic normal (ALN) distribution, Aitchison 1980

Same principle : transport the gaussian distribution from $\mathbb{R}^{D-1}$ to $\mathbf{S}^D$
A random composition $\mathbf{x}$ is normally distributed on $\mathbf{S}^D$, with parameters $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}$, if it has the following density function with respect to Lebesgue measure

$$f(\mathbf{x}) = \frac{(2\pi)^{-(D-1)/2|\boldsymbol{\Sigma}|^{-1/2}}}{\sqrt{D}x_1 \cdots x_D} \exp\left[-\frac{1}{2}(\mathbf{x}^* - \boldsymbol{\mu}^*)\boldsymbol{\Sigma}^{-1}(\mathbf{x}^* - \boldsymbol{\mu}^*)^t\right]$$

where $\mathbf{x}^* = \text{ilr}(\mathbf{x}),\ \boldsymbol{\mu}^* = \text{ilr}(\boldsymbol{\mu})$, and $\boldsymbol{\mu} = \mathbb{E}^{\oplus}\mathbf{X}$.

This is equivalent to say that ilr($\mathbf{x}$) follows a $D-1$ normal distribution with mean $\boldsymbol{\mu}^*$ and variance-covariance matrix $\boldsymbol{\Sigma}$.
The ALN distribution can be estimated by OLS method with the packages 'compositions' and 'robCompositions' in R.

Data description

Vote share data for the French departmental election in 2015 is collected from the Cartelec website. **Vote shares Y** with three components :

- Left (L)
- Right (R)
- Extreme Right (XR)

$$\mathbb{E}^{\oplus}(\mathbf{Y_L}, \mathbf{Y_R}, \mathbf{Y_{XR}}) = (0.37, 0.388, 0.242)$$

Data description

Socio-economic data (2014) are from INSEE :

- **Age** with three categories : Age_1840, Age_4064, and Age_65.
- **Diploma** with three categories : <BAC (0.591), BAC (0.17), SUP (0.234).
- **Employment** with five categories : AZ (Agriculture, pêche), BE (Industrie manufacture, industrie extractive et autres), FZ (Construction), GU (Commerce, transport et service divers) and OQ (Administration publique, enseignement, santé humaine)
- **Unemployment rate** (unemp) is the rate of people who are unemployed
- **Employment evolution** (employ_evol).
- **Rate of people who own assets** (owner).
- **Rate of people who have a salary** (income).
- **Rate of foreigners** (foreign)

Data are collected at the department level in France (95 units).

Vote share description

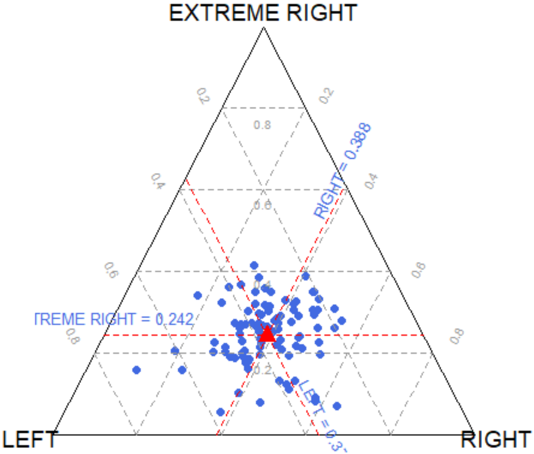
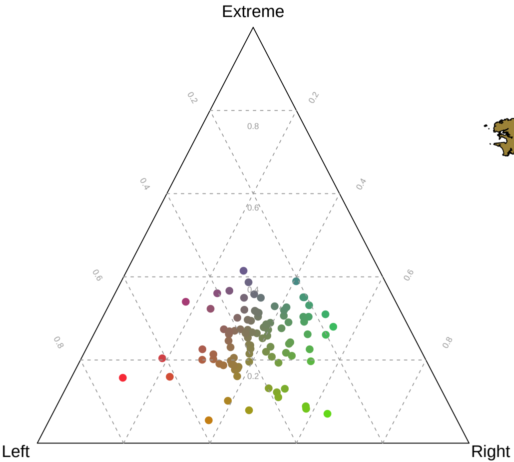
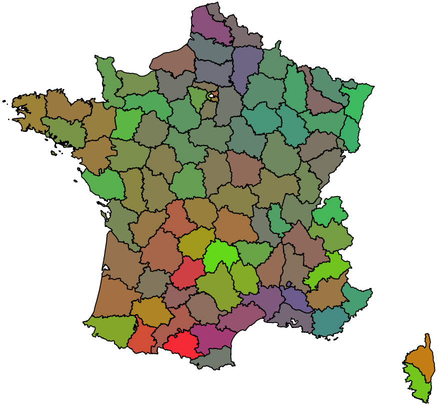


FIGURE – Ternary diagram of vote share data

Alternative vote share description





Logistic normal regression model in the ILR coordinate space

$$\mathbf{S}^L = \left\{ \mathbf{Y} = (Y_1, ..., Y_L) : Y_j > 0, j = 1, ..., L; \sum_{j=1}^{L} Y_j = 1 \right\}$$

$$\mathbf{S}^{D_q} = \left\{ \mathbf{X}_q = (X_{q1}, ..., X_{qD_q}) : X_{qp} > 0; \sum_{p=1}^{D_q} X_{qp} = 1 \right\}, \quad q = 1, ..., Q$$

The regression model in the ILR coordinate space is defined by

$$\mathrm{ilr}(\mathbf{Y_i}) = \mathbf{b_0}^* + \sum_{q=1}^{Q} \mathrm{ilr}(\mathbf{X_{qi}})\mathbf{B}_q^* + \sum_{k=1}^{K} Z_{ki}\mathbf{b}_k^* + \mathrm{ilr}(\boldsymbol{\epsilon}_i) \tag{1}$$

where $\mathbf{b_0}^*, \mathbf{B}_q^*, \mathbf{b}_k^*$ are parameters, and $\mathrm{ilr}(\boldsymbol{\epsilon}_i)$ are residuals which follow the multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$.

Writing the Logistic normal regression model in the simplex

The regression model in the simplex can also be written as

$$\mathbf{Y}_i = \mathbf{b_0} \bigoplus_{q=1}^{Q} \mathbf{B}_q \boxdot \mathbf{X}_{qi} \bigoplus_{k=1}^{K} Z_k \odot \mathbf{b}_k \oplus \boldsymbol{\epsilon}_i, \quad i = 1, ..., n \tag{2}$$

where
$\mathbf{Y}_i \in \mathbf{S}^L$ is the compositional response value of the ith observation ;
$\mathbf{X}_{qi} \in \mathbf{S}^{D_q},\ q = 1, ..., Q$ is the qth compositional covariate value of the ith observation ;
$Z_{ki},\ k = 1, ..., K$ is the kth continuous covariate value of the ith observation ;
$\mathbf{b_0}, \mathbf{B}_1, ..., \mathbf{B}_Q, \mathbf{b}_1, ..., \mathbf{b}_K$ are the parameters satisfying
$\mathbf{b_0},\ \mathbf{b}_k \in \mathbf{S}^L,\ \mathbf{B}_q \in \mathbf{S}^{D_q},\ \mathbf{j}_L^T \mathbf{B}_q = \mathbf{0}_{D_q},\ \mathbf{B}_q \mathbf{j}_{D_q} = \mathbf{0}_L,$
$\boldsymbol{\epsilon}_i \in \mathbf{S}^L$ follows the normal distribution on the simplex (ALN distribution).

Correspondence between parameters in the simplex and in coordinate space

Chen et al. (2016) prove that

$$\begin{cases} \mathbf{b_0} = \exp(\mathbf{b_0}^{*T}\mathbf{V}_L) = \mathrm{ilr}^{-1}(\mathbf{b_0}^*) \\ \mathbf{B}_q = \mathbf{V}_{D_q}^T \mathbf{B}_q^* V_L \end{cases}$$

We prove additionally that

$$\mathbf{b}_k = \exp(\mathbf{b}_k^* \mathbf{V}_L) = \mathrm{ilr}^{-1}(\mathbf{b}_k^*)$$

The parameters in the simplex do not depend on the chosen contrast matrix.

Prediction /Expected shares in the simplex

Two equivalent options :

1. Predict the values in the ILR coordinate space and then ILR inverse transform the predictions in the simplex space :

$$\hat{\mathbf{Y}}_{\mathbf{i}} = \widehat{\mathbb{E}^{\oplus}\mathbf{Y}} = \mathrm{ilr}^{-1}\left(\hat{\mathbf{b}}_0^* + \sum_{q=1}^{Q} \mathrm{ilr}(\mathbf{X}_{qi})\hat{\mathbf{B}}_q^* + \sum_{k=1}^{K} Z_{ki}\hat{\mathbf{b}}_k^*\right) \quad (3)$$

2. ILR inverse transform the estimated parameters in the ILR coordinate space and use the regression model in the simplex :

$$\hat{\mathbf{Y}}_{\mathbf{i}} = \hat{\mathbf{b}}_{\mathbf{0}} \bigoplus_{q=1}^{Q} \hat{\mathbf{B}}_q \boxdot \mathbf{X}_{qi} \bigoplus_{k=1}^{K} Z_{ki} \odot \hat{\mathbf{b}}_k \quad i = 1, \ldots, n \quad (4)$$

Prediction in the simplex

Equation (4) can also be written as follows

$$\hat{\mathbf{Y}}_{\mathbf{i}} = \mathcal{C}\left[\hat{\mathbf{b}}_{\mathbf{0}}.(\prod_{q=1}^{Q}\mathbf{X}_{\mathbf{qi}}{}^{\hat{B}_q}).(\prod_{k=1}^{K}\hat{\mathbf{b}}_k^{Z_{ki}})\right] \quad i = 1, ..., n$$

For example with a single classical variable Z_i

$$\begin{aligned}\hat{\mathbf{Y}}_i &= \mathcal{C}(\hat{\mathbf{b}}_{\mathbf{0}}\hat{\mathbf{b}}^{Z_i}) \\ &= \mathcal{C}\left(\hat{b}_{01}\hat{b}_1^{Z_i}, \hat{b}_{02}\hat{b}_2^{Z_i}, \hat{b}_{03}\hat{b}_3^{Z_i}\right)\end{aligned}$$

With $T = \hat{b}_{01}\hat{b}_1^{Z_i} + \hat{b}_{02}\hat{b}_2^{Z_i} + \hat{b}_{03}\hat{b}_3^{Z_i}$
we get

$$\hat{Y}_{i1} = \frac{\hat{b}_{01}\hat{b}_1^{Z_i}}{T}; \;\; \hat{Y}_{i2} = \frac{\hat{b}_{02}\hat{b}_2^{Z_i}}{T}; \;\; \hat{Y}_{i3} = \frac{\hat{b}_{03}\hat{b}_3^{Z_i}}{T}$$

A simple example to illustrate the impact of one variable on the prediction

- Vote share **Y** with three categories : Left (L), Right (R), and Extreme Right (XR).
- One classical explanatory variable Z : Unemp or income.
- Model estimated in the ILR coordinate space :

	Dependent variable :			
	y_ilr[, 1]	y_ilr[, 2]	y_ilr[, 1]	y_ilr[, 2]
unemp	-6.422^{***}	12.739^{***}		
	(1.956)	(1.977)		
income			1.350^{**}	-1.176
			(0.612)	(0.712)
Constant	0.787^{***}	-1.859^{***}	-0.712^{**}	0.285
	(0.232)	(0.234)	(0.340)	(0.396)
Observations	95	95	95	95
R^2	0.104	0.309	0.050	0.028
Adjusted R^2	0.094	0.301	0.039	0.018
Residual Std. Error (df = 93)	0.330	0.333	0.339	0.395
F Statistic (df = 1 ; 93)	10.782^{***}	41.540^{***}	4.862^{**}	2.726
Note :		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		

Estimated parameters in the simplex

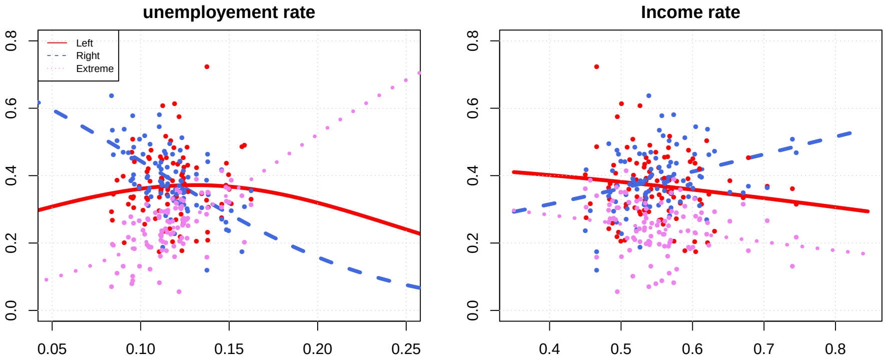
We get the estimated parameters in the simplex as :

	Left	Right	Extreme Right
Intercept	2.367e-01	7.208e-01	4.2e-02
Unemp	1.570e-05	1.786e-09	0.999

The predictions of the vote share for departments are

$$\hat{Y}_{Li} = 0.2367 * (1.570e^{-05})^{Z_i}$$
$$\hat{Y}_{Ri} = 0.7208 * (1.786e^{-09})^{Z_i}$$
$$\hat{Y}_{ERi} = 0.042 * (0.999)^{Z_i}$$

Impact of a classical explanatory variable



Note that at each value x of the covariate, we get $\hat{Y}_L + \hat{Y}_R + \hat{Y}_{XR} = 1$. In the simplex, the link between $\hat{Y}_j$, $j = L, R, XR$ and X is not linear neither monotone.

Impact of a compositional explanatory variable

We will consider the compositional variable Diploma

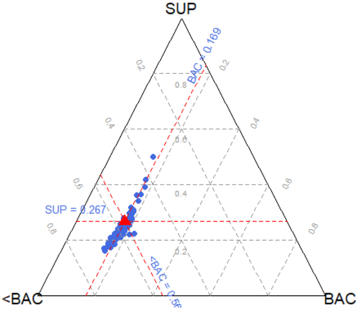


FIGURE – Observation of Diploma (blue points) and its geometric mean (red triangle)

Impact of a compositional explanatory variable

	Dependent variable	
	y_ilr[, 1]	y_ilr[, 2]
ilr1 (SUP/(BAC+<BAC))	$-0.47(0.27)^{\cdot}$	$-1.21(0.27)^{***}$
ilr2 (BAC/<BAC)	$1.26(0.50)^{*}$	$3.14(0.50)^{***}$
Constant	$-0.97(0.40)^{*}$	$-2.86(0.39)^{***}$
R^2	0.064	0.299
Adjusted R^2	0.044	0.284
Residual Std. Error (df = 92)	0.338	0.337
F Statistic (df = 2 ; 92)	3.168^{**}	19.67^{***}
Note :	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

TABLE – Regression results

Estimated parameters in the simplex

We get the estimated parameters as

	Left	Right	Extreme Right
Intercept	0.788	0.200	0.012
<BAC	-1.20	1.88	-1.68
BAC	-0.21	0.34	-0.13
SUP	1.41	-2.22	0.81

The predictions of the vote shares for departments are

$$\hat{Y}_{Li} = 0.788 * (<\mathrm{BAC}_i)^{-1.20} * (\mathrm{BAC}_i)^{-0.21} * (\mathrm{SUP}_i)^{1.41} / TD$$
$$\hat{Y}_{Ri} = 0.2 * (<\mathrm{BAC}_i)^{1.88} * (\mathrm{BAC}_i)^{0.34} * (\mathrm{SUP}_i)^{-2.22} / TD$$
$$\hat{Y}_{Xi} = 0.012 * (<\mathrm{BAC}_i)^{-1.68} * (\mathrm{BAC}_i)^{-0.13} * (\mathrm{SUP}_i)^{0.81} / TD$$

where

$$TD = \sum_{i=1}^{3} \hat{b}_{0i} (<\mathrm{BAC}_i)^{\hat{b}_{i1}} (\mathrm{BAC}_i)^{\hat{b}_{i2}} (\mathrm{SUP}_i)^{\hat{b}_{i3}}$$

Impact of a compositional explanatory variable

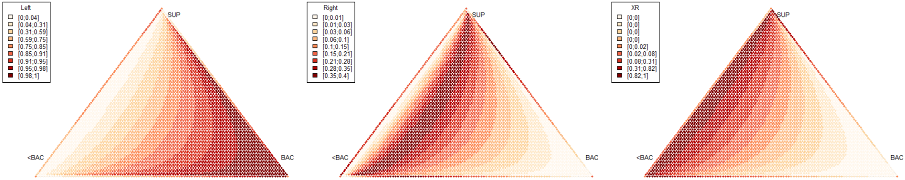


FIGURE – Predictions of vote shares according to Diploma

Impact of a compositional explanatory variable

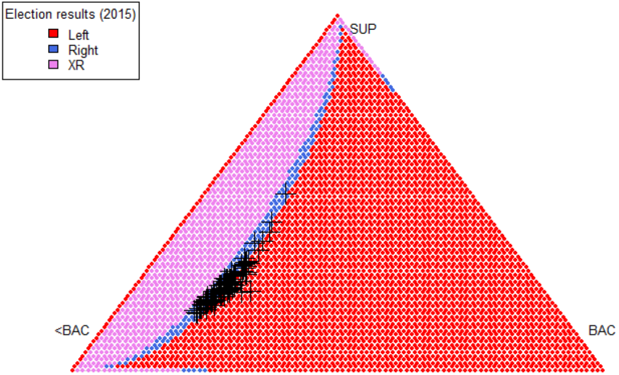


FIGURE – Predictions of vote shares according to Diploma : majority party

Full model with compositional and classical explanatory variables

	y_ilr[, 1]	y_ilr[, 2]
Diploma_ilr1	$-2.06(0.54)^{***}$	$-1.51(0.46)^{**}$
Diploma_ilr2	$-1.28(0.80)$	$-2.07(0.67)^{**}$
Employ_ilr1	$-0.05(0.30)$	$-2.12(0.34)$
Employ_ilr2	$0.12(0.37)$	$-2.62(0.46)^{**}$
Employ_ilr3	$0.30(0.30)$	$-2.12(0.34)$
Employ_ilr4	$0.13(0.11)$	$-2.62(0.46)^{\cdot}$
unemp	$-7.65(3.16)^{*}$	$-2.12(0.34)^{***}$
income	$2.04(1.37)$	$-2.62(0.46)^{***}$
Constant	$-2.324(1.15)^{*}$	$-4.80(0.97)^{***}$
R^2	0.30	0.62
Adjusted R^2	0.23	0.59
Residual Std. Error (df = 86)	0.30	0.26
F Statistic (df = 8 ; 86)	4.602^{***}	17.85^{***}

Perspectives

- CODA regression models can be useful in the context of political economy
- Introduce geographical dimension
- Use logistic Student distribution instead of logistic normal distribution
- Use elasticity to characterize impacts of covariates

Thank you for your attention !