Generalized Akaike Information Criterion for small area models

M.J. Lombardía

In collaboration with
E. López-Vizcaíno, and C. Rueda

Research Group on Modeling, Optimization and Statistical Inference (MODES)
Departamento de Matemáticas
Universidade da Coruña

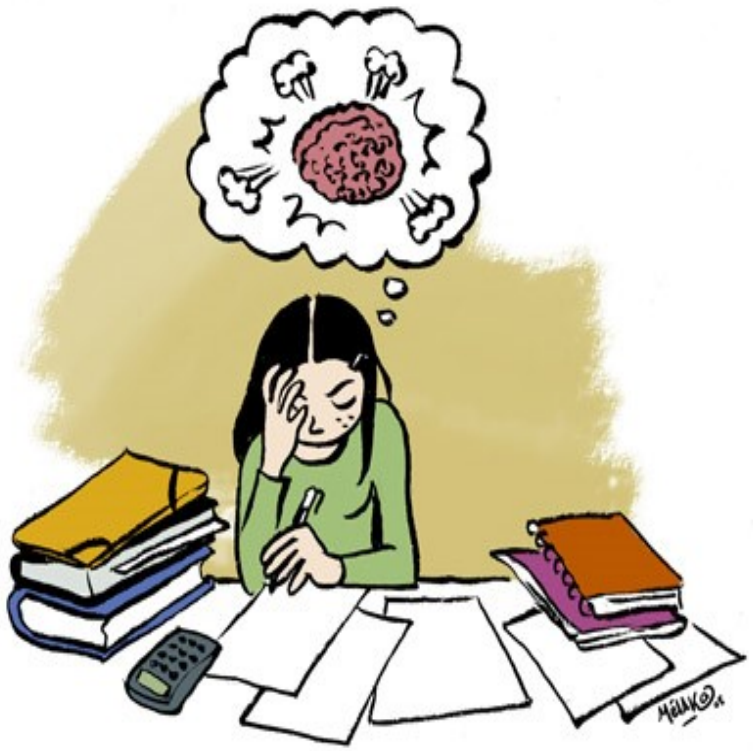
June 8-9, 2016
Galician Seminar of Nonparametric Statistical Inference (GSNSI)

UNIVERSIDADE DA CORUÑA

HAPPY BIRTHDAY

HAPPY BIRTHDAY

- **Inference in Finite Populations**



HAPPY BIRTHDAY

- **Small Area Estimation**

HAPPY BIRTHDAY

- **Mixed models**

Generalized Akaike Information Criterion for small area models

M.J. Lombardía

In collaboration with
E. López-Vizcaíno, and C. Rueda

Research Group on Modeling, Optimization and Statistical Inference (MODES)
Departamento de Matemáticas
Universidade da Coruña

June 8-9, 2016
Galician Seminar of Nonparametric Statistical Inference (GSNSI)

UNIVERSIDADE DA CORUÑA

Index

Section 1 | Introduction

Linear Models

- Linearity: the mean of the observation is a linear function of some covariates
- Normality: multivariate normal distribution for the vector of observed y-values
- Independence: observations are independent

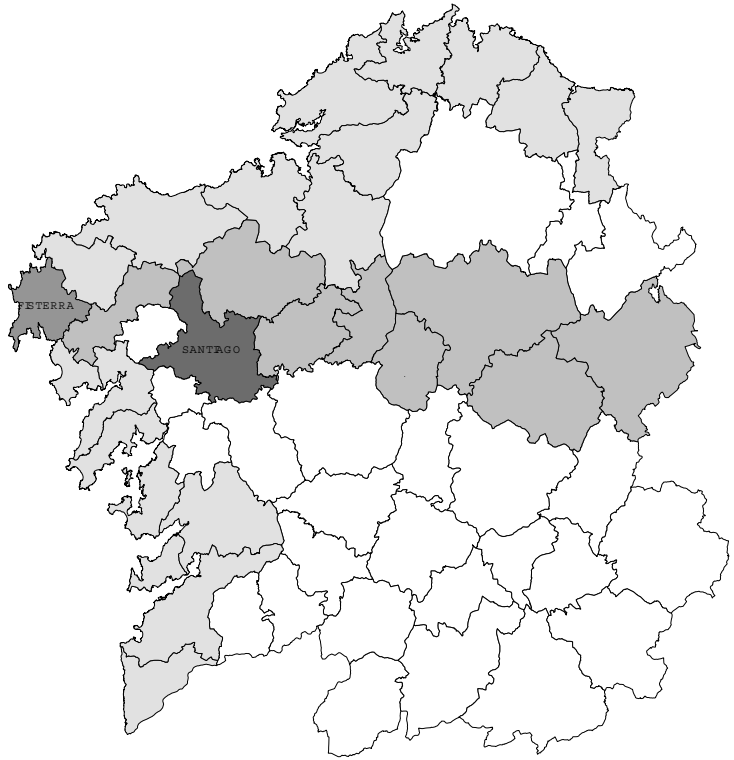
Linear Mixed Models

- Mixed models have a more complex multilevel or hierarchical structure. Observations in different levels or clusters are assumed to be independent, but observations within the same level or cluster are considered as dependent because they share common properties. Two sources of variation: between and within clusters. The possibility of modelling those sources of variation, commonly present in real data, gives a high flexibility, and therefore applicability, to mixed models.

What is a Small Area or Domain?

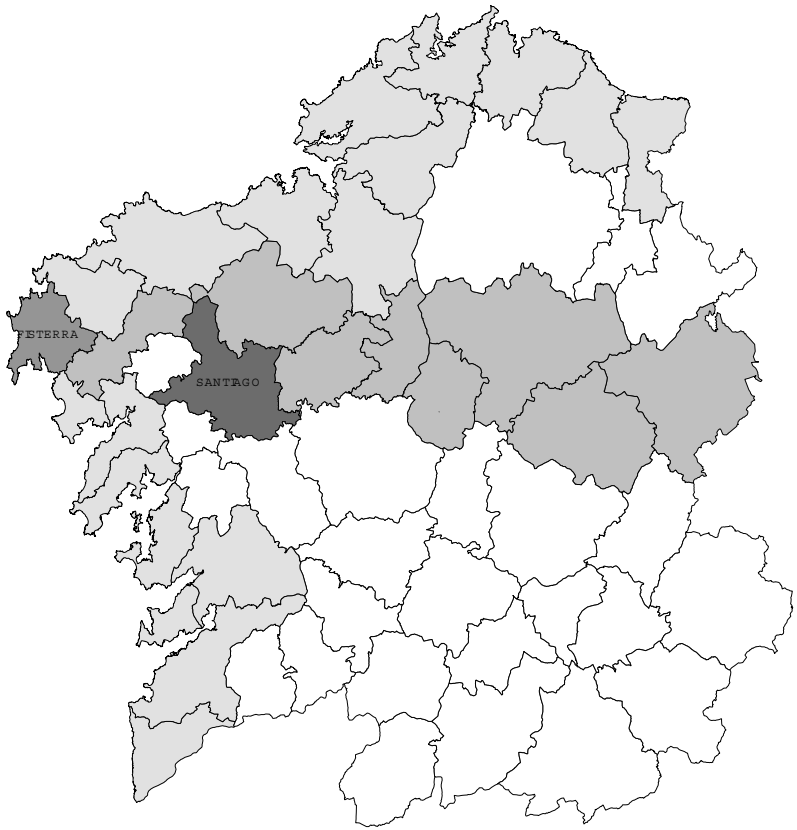
- Small Area: is commonly used to denote a small geographical area, such as a county, a municipality or a census division.

- Small Domain: is commonly used to denote a small subpopulation such as a specific age-sex-race group of people within a large geographical area. They may also describe a Small Area.



Why the inference problem?

Sample survey data can be used to derive reliable estimators of parameters (totals, means,....) for large areas or domains. The usual direct survey estimators for a small area, based on data only from the sample units in the area, are likely to yield unacceptably large standard errors due to the unduly small size of the sample in the area.



The aim

- Model selection and checking is one of major problems in SAE.
- Model selection for linear mixed models is different from model selection for linear regression models.
- Broad approaches: (Muller et al., 2013[1]).

Information Criteria

AIC(Akaike, 1973)
BIC (Schwarz, 1978).

Shrinkage Methods

LASSO (Tibshirani,1996).

Fence Methods

Jiang et al., 2008.

Others Methods

Others Bayesian methods, testing, etc.

[1] Muller, S., Scealy, J.L. and Welsh, A.H. (2013). Model Selection in linear Mixed Models *Statistical Science*, vol. 28, 135-167.

AIC

AIC

$$AIC(M) = -2\log(l(M)) + 2D$$

- AIC was designed to be an approximately unbiased estimator of the expected Kullback-Leibler information of a fitted model.
- $l(M)$ is the model likelihood $\Longleftarrow$ Loss function
- D measurement of model complexity $\Longleftarrow$ Penalty term
- The model with the lowest value of AIC is selected.

GAIC

GAIC

$$GAIC(M) = -2\log(l(M)) + GDF$$

- GDF is a measure of the sensitivity of each fitted value to perturbation in the corresponding observed value $\Longleftarrow$ Penalty term applicable to complex modeling procedures (Ye, 1998[2]).

[2] Ye, J. (1998). On measuring and correcting the effects of data mining and model selection, *J. Amer. Statist. Assoc.*, Vol. 93, 120-131.

GAIC

GAIC

$$GAIC(M) = -2\log(l(M)) + GDF$$

- This definition is vague because we can define different versions of the Akaike Information using different log density-like functions and we can consider various model estimators.
- $l(M)$ is the model likelihood $\Longleftarrow$ Loss function (Vaida and Blanchard, 2005[3]; Greven and Kneib, 2010[4]; Pfeffermann, 2013[5]; Han, 2013[6]).

[3] Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models, *Biometrika*, Vol. 92, 351-370.

[4] Greven,S. and Kneib, T. (2010). On The beahviour of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models, *Biometrika*, Vol. 97, 773-789.

[5] Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, Vol. 28, 40-68.

[6] Han, B (2013). Conditional Akaike information criterion in the Fay-Herriot model, *Statistical Methodology*, Vol. 11, 53-67.

Section 2 | xGAIC

The model

Notation:

- D is the number of the domains or small areas.
- μ_d is the characteristic of interest in the d-th area.
- $y_d = \bar{y}_d$ is the direct estimator of the characteristic μ_d.
- p auxiliary variables $(X_1, \ldots X_p)$.

The model

The model is composed in two-stages:

Fist stage:

$$y_d \sim N(\mu_d, \sigma_d^2) \quad \rightarrow \quad y_d = \mu_d + e_d, \quad d = 1, \ldots, D;$$

where $e_d \sim N(0, \sigma_d^2)$ are independent with σ_d^2 known, in practice we take the design-based variance of direct estimator y_d.

Second stage:

$$\mu_d \sim N(\theta_d, \sigma_u^2) \quad \rightarrow \quad \mu_d = \theta_d + u_d, \quad d = 1, \ldots, D;$$

where $\theta_d = f(x_{1d}, \ldots, x_{pd})$ is a linear or nonlinear function depending on the model considered, and $u_d \sim N(0, \sigma_u^2)$ are independent with the variance σ_u^2 unknown.

The final model can be expressed as a single model

$$y_d = \theta_d + u_d + e_d, \quad d = 1, \ldots, D.$$

The model

$$\mathbf{Y} = \boldsymbol{\theta} + \mathbf{u} + \mathbf{e}$$

Assumptions:

$\mathbf{u} \sim N(0, \boldsymbol{\Sigma}_u = \sigma_u^2 \boldsymbol{I}_D)$ is the small area random effect and independent of the model error $\mathbf{e} \sim N(0, \boldsymbol{\Sigma}_e)$, and $\boldsymbol{I}_D$ the identity matrix with dimension D. Note that the variability of $\mathbf{e}$ is known and different in each area, $\boldsymbol{\Sigma}_e = diag(\sigma_1^2, \ldots, \sigma_D^2)$.

Marginal approach

$$\begin{aligned} E(\mathbf{Y}) &= \boldsymbol{\theta} \\ Var(\mathbf{Y}) &= \mathbf{V}_y = \Sigma_u + \Sigma_e \end{aligned}$$

Conditional approach

$$\begin{aligned} E(\mathbf{Y}|\mathbf{u}) &= \boldsymbol{\mu} = \boldsymbol{\theta} + \mathbf{u} \\ Var(\mathbf{Y}|\mathbf{u}) &= \mathbf{V}_{y|u} = \Sigma_e \end{aligned}$$

The calculation of log-likelihood

$$\mathbf{Y} = \boldsymbol{\theta} + \mathbf{u} + \mathbf{e}$$

Marginal approach

$$\begin{aligned} E(\mathbf{Y}) &= \boldsymbol{\theta} \\ Var(\mathbf{Y}) &= \mathbf{V}_y = \Sigma_u + \Sigma_e \end{aligned}$$

Conditional approach

$$\begin{aligned} E(\mathbf{Y}|\mathbf{u}) &= \boldsymbol{\mu} = \boldsymbol{\theta} + \mathbf{u} \\ Var(\mathbf{Y}|\mathbf{u}) &= \mathbf{V}_{y|u} = \Sigma_e \end{aligned}$$

- **Marginal log-likelihood**:

$$\log(l_m(M)) = -\frac{1}{2} D \log(2\pi) - \frac{1}{2} \log |\mathbf{V}_y| - \frac{1}{2} (\mathbf{Y} - \boldsymbol{\theta})' \mathbf{V}_y^{-1} (\mathbf{Y} - \boldsymbol{\theta})$$

- **Conditional log-likelihood**:

$$\log(l_c(M)) = -\frac{1}{2} D \log(2\pi) - \frac{1}{2} \log |\mathbf{V}_{y|u}| - \frac{1}{2} (\mathbf{Y} - \boldsymbol{\mu})' \mathbf{V}_{y|u}^{-1} (\mathbf{Y} - \boldsymbol{\mu})$$

The calculation of GDF

GDF

GDF is a measure of the sensitivity of the expected estimated of the response with respect to the corresponding underlying means (Ye (1998)[a], You et al. (2016)[b]).

$$xGDF = \sum_{d=1}^{D} \frac{\partial E(\widehat{\mu}_i)}{\partial \mu_i} = \sum_{d=1}^{D} \sum_{i=1}^{D} V_{di}^{-1} Cov(\widehat{\mu}_d, y_i)$$

[a] Ye, J. (1998). On measuring and correcting the effects of data mining and model selection, *J. Amer. Statist. Assoc.*, Vol. 93, 120-131.

[b] You,C., Muller,S. and Ormerod,J.T. (2016). On generalized Degrees of Freedom with application in linear models selection, *Statistics and Computing*, Vol. 26, 199-210.

As alternative,

$$cGDF = \sum_{d=1}^{D} \sum_{i=1}^{D} V_{(y|u),di}^{-1} Cov(\widehat{\mu}_d, y_i)$$

The calculation of GDF: Parametric bootstrap

- **Mixed approach**

1. Fit the model $y_d = f(x_{1d}, \ldots, x_{pd}) + u_d + e_d$ with $u_d \sim N(0, \sigma_u^2)$ independent of $e_d \sim N(0, \sigma_d^2)$ and $V_d = Var(y_d)$. We calculate the estimators of the model parameters.

2. Repeat B times $(b = 1, \ldots, B)$

   1. Generate u_d^* and e_d^* as independents $N(0, \widehat{\sigma}_u^2)$ and $N(0, \sigma_d^2)$ respectively, $d = 1, \ldots, D$. Construct the bootstrap model $y_d^{*(b)} = \mu_d^{*(b)} + e_d^{*(b)}$, with $\mu_d^{*(b)} = \widehat{f}(x_{1d}, \ldots, x_{pd}) + u_d^{*(b)}$ and $\widehat{f}(x_{1d}, \ldots, x_{qd})$ the fitted model and $V_{y,d}^{*(b)} = Var(y_d^{*(b)})$.

   2. For each bootstrap sample, calculate $\widehat{\mu}_d^{*(b)} = \widehat{f}^{*(b)}(x_{1d}, \ldots, x_{pd}) + \hat{u}_d^{*(b)}$.

3. Calculate GDF as

$$\widehat{xGDF} = \sum_{d=1}^{D} \sum_{i=1}^{D} \frac{1}{B-1} \sum_{b=1}^{B} (V_{y,di}^{*(b)})^{-1} (\widehat{\mu}_d^{*(b)} - \bar{\widehat{\mu}}_d^*)(y_i^{*(b)} - \bar{y}_i^*)$$

where $\bar{\widehat{\mu}}_d^* = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mu}_d^{*(b)}$ and $\bar{y}_d^* = \frac{1}{B} \sum_{b=1}^{B} y_d^{*(b)}$.

The calculation of GDF: Parametric bootstrap

- **Conditional approach**

1. Fit the model $y_d = f(x_{1d}, \ldots, x_{pd}) + u_d + e_d$ where $u_d \sim N(0, \sigma_u^2)$ independent of $e_d \sim N(0, \sigma_d^2)$. With moments μ_d and $V_{y|u,d}$. Then, we calculate the estimators of the model parameters.

2. Repeat B times $(b = 1, \ldots, B)$
   1. Generate e_d^* as $N(0, \sigma_d^2)$, $d = 1, \ldots, D$. Construct the bootstrap model $y_d^{*(b)}|\hat{u}_d = \widehat{\mu}_d + e_d^{*(b)}$, with $\widehat{\mu}_d = \widehat{f}(x_{1d}, \ldots, x_{pd}) + \hat{u}_d$ and $\widehat{f}(x_{1d}, \ldots, x_{qd})$ the fitted model and $V_{y|u,d}^{*(b)} = Var(y_d^{*(b)}|\hat{u}_d)$.
   2. For each bootstrap sample, calculate $\widehat{\mu}_d^{*(b)} = \widehat{f}^{*(b)}(x_{1d}, \ldots, x_{pd}) + \hat{u}_d^{*(b)}$.

3. Calculate GDF as

$$\widehat{xGDF} = \sum_{d=1}^{D} \sum_{i=1}^{D} \frac{1}{B-1} \sum_{b=1}^{B} (V_{y|u,di}^{*(b)})^{-1} (\widehat{\mu}_d^{*(b)} - \bar{\widehat{\mu}}_d^*)(y_i^{*(b)} - \bar{y}_i^*)$$

where $\bar{\widehat{\mu}}_d^* = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mu}_d^{*(b)}$ and $\bar{y}_d^* = \frac{1}{B} \sum_{b=1}^{B} y_d^{*(b)}$.

xGAIC

xGAIC

- $cGAIC = -2\log(l_c(\widehat{M})) + c\widehat{GDF}$
- $yGAIC = -2\log(l_c(\widehat{M})) + x\widehat{GDF}$ (You et al. (2016)[7])
- $yGAIC = -2\log(l_m(\widehat{M})) + x\widehat{GDF}$ (You et al. (2016)

xGAIC

As alternative,

$$\log(l_x(M)) = -\frac{1}{2}D\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_y| - \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\mathbf{V}_y^{-1}(\mathbf{Y}-\boldsymbol{\mu}).$$

- $xGAIC = -2\log(l_x(\widehat{M})) + x\widehat{GDF}$

[7] You, C., Muller, S. and Ormerod, J.T. (2016). On generalized Degrees of Freedom with application in linear models selection, *Statistics and Computing*, Vol. 26, 199-210.

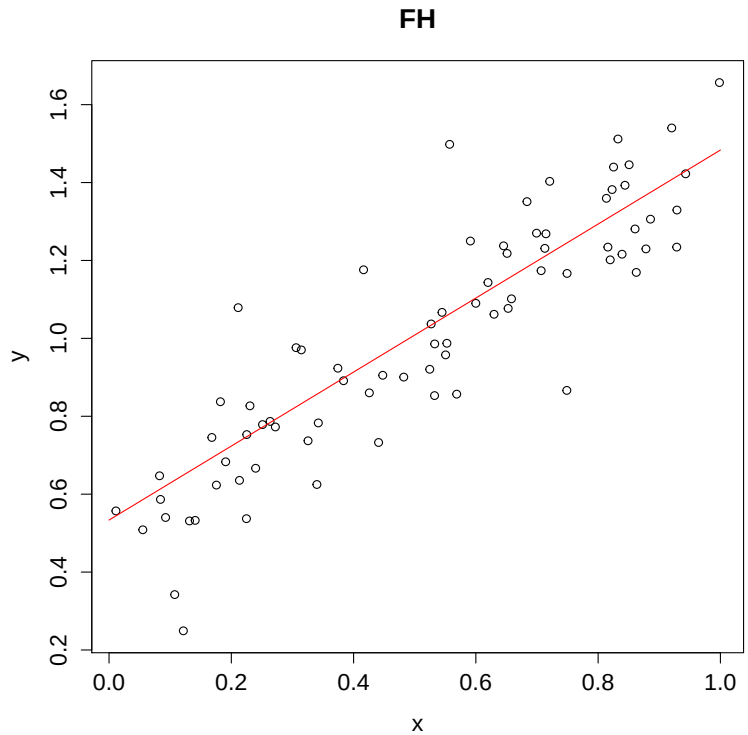
Section 3 | Particular models

The model

Fay-Herriot model:

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients.



The model

Fay-Herriot model:

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\mu} = \boldsymbol{\theta} + \mathbf{u}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients.

To fit the model we use Maximun Likelihood Estimation (MLE) and we use the functions available in package `sae` in R languaje (Molina and Marhuenda (2015)[8]).

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{Y} \quad \text{and} \quad \widetilde{\mathbf{u}} = \boldsymbol{\Sigma}_u\mathbf{V}_y^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

The variance components σ_u^2 are unknown, then well-known methods such MLE or restricted maximum likelihood (REML) can be used to estimate them, $\widehat{Var}(\mathbf{Y}) = \widehat{\mathbf{V}}_y = \widehat{\boldsymbol{\Sigma}}_u + \widehat{\boldsymbol{\Sigma}}_e$, you can see the details of the calculation in Rao and Molina (2015)[9] .

$$\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \quad \text{and} \quad \widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{u}}$$

[8] Molina I. and Marhuenda Y. (2015). sae: An R Package for Small Area Estimation. *The R Journal*, Vol. 7, 81-98.

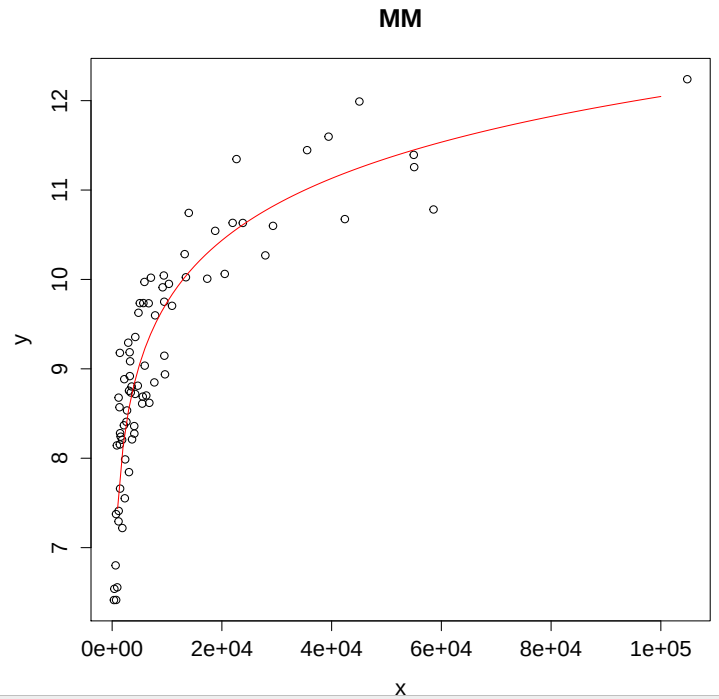
[9] Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation*, Wiley.

The model

Monotone model:

$$\theta_d = f(x_{1d}, \ldots, x_{pd}) = \sum_{j=1}^{p_1} \beta_j x_{jd} + \sum_{j=p_1+1}^{p} h_j(x_{jd}), \quad d = 1, \ldots, D;$$

where $h_j()$ are monotone functions.



The model

Monotone model:

$$\theta_d = f(x_{1d}, \ldots, x_{pd}) = \sum_{j=1}^{p_1} \beta_j x_{jd} + \sum_{j=p_1+1}^{p} h_j(x_{jd}), \quad d = 1, \ldots, D;$$

where $h_j()$ are monotone functions.

To obtain the MLE we use the methodology proposed in Rueda and Lombardía (2012)[10]

$$\widehat{\theta}_d = \sum_{j=1}^{p_1} \widehat{\beta}_j x_{jd} + \sum_{j=p_1+1}^{p} \widehat{h}_j(x_{jd}) = P_W(\mathbf{Y}|\mathbf{K})$$

$$\widehat{\mu}_d = \left(1 - \frac{\sigma_u^2}{\sigma_d^2 + \sigma_u^2}\right) \widehat{\theta}_d + \frac{\sigma_u^2}{\sigma_d^2 + \sigma_u^2} Y_d, \quad d = 1, \ldots, D.$$

In the case σ_u^2 unknown, we propose an iterative procedure to obtain $\widehat{\boldsymbol{\theta}} = P_W(\mathbf{Y}|\mathbf{K})$ and $\widehat{\sigma}_u^2$, which is based on Rueda et al. (2010)[11].

[10] Rueda, C. and Lombardía, M.J. (2012). Small Area Semiparametric Additive Isotone Models, *Statistical Modelling*, Vol. 12, 503-525

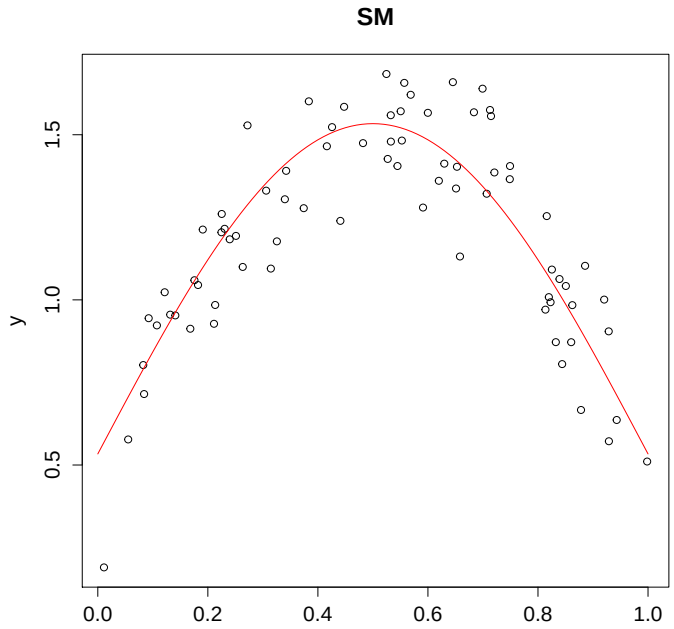
[11] Rueda, C. and Menéndez, J.A. and Gómez, F. (2010). Small area estimators based on restricted mixed models, *TEST*, Vol. 19, 558-568

The model

Penalized spline model:

$$\theta_d = f(x_{1d}, \ldots, x_{pd}) = \sum_{j=1}^{p_1} \beta_j x_{jd} + \sum_{j=p_1+1}^{p} f_j(x_{jd}), \quad d = 1, \ldots, D;$$

where $p = p_1 + p_2$ the number of area auxiliary variables, $f_j()$ are any smooth functions.



The model

Penalized spline model:

$$\theta_d = f(x_{1d}, \ldots, x_{pd}) = \sum_{j=1}^{p_1} \beta_j x_{jd} + \sum_{j=p_1+1}^{p} f_j(x_{jd}), \quad d = 1, \ldots, D;$$

where $p = p_1 + p_2$ the number of area auxiliary variables, $f_j()$ are any smooth functions.

Using P-splines we can write the model as the following mixed effects model (Opsomer et al. (2008))[12].

$$\mathbf{Y} = \boldsymbol{\theta} + \mathbf{u} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{u} + \mathbf{e},$$

where $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}$ represents the spline function. For fitting the model is suitable to treat $\mathbf{Z}\mathbf{v}$ as a random effect term, with $\mathbf{v} \sim N(0, \boldsymbol{\Sigma}_v = \sigma_v^2 \boldsymbol{I}_{c-2})$, where c is the dimension of $\mathbf{Z}$. Then, the covariance matrix of the variable $\mathbf{Y}$ is given by $Var(\mathbf{Y}) = \mathbf{V}_y = \mathbf{Z}\boldsymbol{\Sigma}_v\mathbf{Z}' + \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_e$, adding an additional term if we compare with the Fay-Herriot model.

$$\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{v}} \quad \text{and} \quad \widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{v}} + \widehat{\mathbf{u}}$$

Section 4 | Simulation study

Simulation study

$$y_d = f(x_d) + u_d + e_d$$

where $D = 77$, $x_d \sim U(0,1)$, $u_d \sim N(0, \sigma_u^2)$ and $e_d \sim N(0, \sigma_d^2)$.

12 scenarios are designed, based on different definitions for $f()$, and different σ_u and $\sigma_d, d = 1, ..., D$, values:

- (LM): $f(x_d) = \beta_0 + \beta_1 x_d$
- (MM): $f(x_d) = \beta_0 + \log(x_d)$
- (NM): $f(x_d) = \beta_0 + \sin(\pi x_d)$

- $\sigma_d^2 = \sigma_{de}^2, \qquad \sigma_d^2 = \sigma_{de}^2 * 10$
- $\sigma_u^2 = \sigma_{ue}^2, \qquad \sigma_u^2 = \sigma_{ue}^2/10$

Simulation study

Global statistics:

- Correct classification rates from using $xGAIC$, $cGAIC$ and $yGAIC$.
- Average values of $\widehat{\sigma}_u^2$, $\widehat{xGDF}$ and $\widehat{cGDF}$, for the Fay-Herriot, monotone and P-spline model.
- Relative root mean squared error (RRMSE) for the $\widehat{\sigma}_u^2$, corresponding to the model selected by $xGAIC$, $cGAIC$ and $yGAIC$:

$$RRMSE(\widehat{\sigma}_u^2) = \frac{\sqrt{\frac{1}{I}\sum_{i=1}^{I}(\widehat{\sigma}_u^{2(i)} - \sigma_u^2)^2}}{\sigma_u^2}.$$

Simulation study

Scenario	$xGAIC$			$cGAIC$		
	Fay-Herriot	Monotone	P-spline	Fay-Herriot	Monotone	P-spline
$\sigma^2_{ue}, \sigma^2_{de}$						
LM	36.98	61.85	1.17	47.04	44.07	8.89
MM	0	100	0	36.8	30.2	33
NM	1.35	14.35	84.3	34.98	33.63	31.39
$\sigma^2_{ue}/10, \sigma^2_{de}$						
LM	24.37	73.11	2.52	34.18	60.5	5.32
MM	0	100	0	13.6	46.4	40
NM	0	0	100	15.72	24.93	59.35
$\sigma^2_{ue}, \sigma^2_{de}$*10						
LM	36.57	56.12	7.31	38.10	47.96	13.95
MM	0	100	0	35.27	39.83	24.90
NM	3.60	2.80	93.60	10.80	54.80	34.40
$\sigma^2_{ue}/10, \sigma^2_{de}$*10						
LM	22.35	63.22	12.36	24.48	66.12	9.48
MM	0	100	0	7.00	56.60	36.40
NM	0	0	100	8.80	59.03	32.18

Table 1: Percentage of times Fay-Herriot, Monotone or P-Spline models are selected by $xGAIC$ and $cGAIC$ under different simulated scenarios.

Simulation study

Scenario	$xGAIC$	$cGAIC$	$yGAIC$
σ^2_{ue},σ^2_{de}			
LM	0.09	0.08	0.08
MM	0.12	0.59	0.53
NM	0.11	0.20	0.19
$\sigma^2_{ue}/10$, σ^2_{de}			
LM	0.12	0.11	0.11
MM	0.18	1.60	1.55
NM	0.10	0.91	0.92
σ^2_{ue}, $\sigma^2_{de} * 10$			
LM	0.10	0.10	0.09
MM	0.13	0.57	0.63
NM	0.11	0.20	0.20
$\sigma^2_{ue}/10$, $\sigma^2_{de} * 10$			
LM	0.19	0.16	0.15
MM	0.36	1.27	1.77
NM	0.15	0.97	0.98

Table 2: RRMSE of $\widehat{\sigma}^2_u$ using the model selected by $xGAIC$, $cGAIC$ and $yGAIC$ under different simulated scenarios.

Simulation study

Scenario	$\widehat{xGDF}$			$\widehat{cGDF}$		
	Fay-Herriot	Monotone	P-spline	Fay-Herriot	Monotone	P-spline
$\sigma^2_{ue}, \sigma^2_{de}$						
LM	74.83	75.05	74.86	74.89	74.99	74.86
MM	76.29	74.98	75.02	76.30	75.11	75.10
NM	75.44	75.45	74.98	75.47	75.41	74.96
$\sigma^2_{ue}/10, \sigma^2_{de}$						
LM	64.92	65.69	64.69	64.49	65.33	64.26
MM	76.12	67.15	69.42	76.12	67.23	69.95
NM	74.02	73.62	64.28	74.01	73.60	64.25
$\sigma^2_{ue}, \sigma^2_{de}$*10						
LM	65.09	65.80	64.79	64.70	65.25	64.37
MM	72.47	65.74	65.75	72.54	65.47	65.70
NM	67.61	68.27	65.22	67.28	67.60	64.64
$\sigma^2_{ue} * 10, \sigma^2_{de}$*10						
LM	40.17	42.79	39.39	38.75	40.36	38.16
MM	71.57	45.59	49.24	71.62	43.88	50.03
NM	60.14	58.36	38.26	59.60	57.22	37.70

Table 3: Average values of $\widehat{xGDF}$ and $\widehat{cGDF}$ under different simulated scenarios.

Section 5 | Real applications

Application to Labour Force Survey

- **Data set**: Labour Force Survey (LFS) of Galicia in the fourth quarter of 2013
- **Domains**: Economic activity ($D = 77$)

- **Objective**: Total employed people in each domain d, which includes people currently employed in the activity or unemployed people whose last job has been in such activity.

  Our goal is to estimate

$$Y_d = \sum_{j \in P_d} y_j,$$

  where $y_j = 1$ if the person j of domain d is employed and $y_j = 0$ in other case, and P_d is the population in the economic activity d.

Application to Labour Force Survey

Is it a small areas estimation problem?

- The LFS does not produce official estimates at the domain level, but the analogous direct estimates of the total Y_d, the mean $\bar{Y}_d = Y_d/N_d$ and the size N_d are

$$\hat{Y}_d^{dir} = \sum_{j \in S_d} w_j \, y_j, \; \hat{\bar{Y}}_d^{dir} = \hat{Y}_d^{dir} / \hat{N}_d^{dir}, \; \hat{N}_d^{dir} = \sum_{j \in S_d} w_j;$$

  where S_d is the sample domain and w_j's are the official calibrated sampling weights.

- The problem of the LFS is that when the domains are bellow the planned level we find very low sample sizes of domains and therefore very high sampling errors.

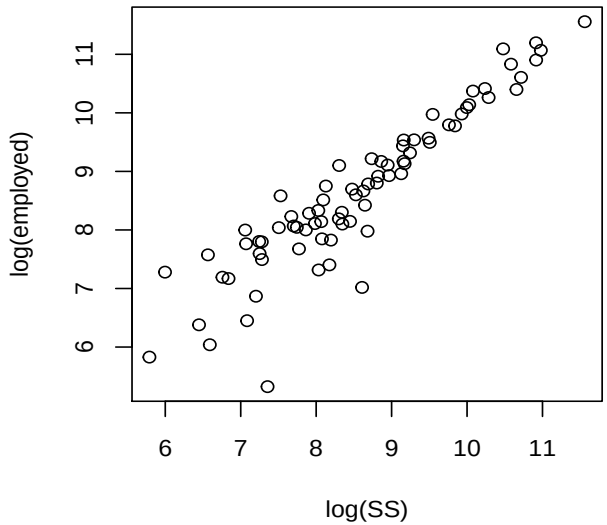
- For the fourth quarter of 2013
  - the minimum sample size in the domains is 1,
  - the first quartile is 12,
  - the median 31,

  therefore for some domains with the direct estimator can not get a reliable estimate of our objective.

Application to Labour Force Survey

- **Response variable (Y_d)**: $\hat{Y}_d^{dir}$.
- **Auxiliary variable (X_d)**: The people registered in the social security system (SS).
- **Model**: The models are formulated using the log transform to better fit the normality error assumption.

$$log(Y_d) = f(log(X_d)) + u_d + e_d$$



Application to Labour Force Survey

Model	$\widehat{cGDF}$	$cGAIC$	$\widehat{xGDF}$	$xGAIC$	$\hat{\sigma}_u^2$
Fay-Herriot	74.8	-296.2	74.5	99.8	0.21
Monotone	76.0	-297.5	75.8	109.0	0.24
P-spline	73.9	-296.1	74.7	100.1	0.21

Table 4: $\widehat{GDF}$, conditional and mixed $GAIC$ and $\hat{\sigma}_u^2$.

Fay-Herriot model:

$$log(\widehat{Y}_d^{dir}) = log(SS_d)\beta + u_d + e_d$$

Application to Health Data

- **Data set**: Surveys from the Behavioural Risk Factors Information System in Galicia (SICRI) for the period 2010-2011.

- **Domains**: $D = 41$ areas obtained from the 53 counties of Galicia.

- **Objective**: Prevalence of smokers by sex among the population aged 16 years and over, in the 41 areas of Galicia in the period 2010-2011.

  Our goal is to estimate

$$Y_d = \sum_{j \in P_d} y_j,$$

  where $y_j = 1$ if the person j of domain d is a smoker and $y_j = 0$ in other case, and P_d is the population in the area d.

Application to Health Data

Is it a small areas estimation problem?

- $\widehat{Y}^{dir}$ is the total direct estimator obtained from the SICRI. SICRI is designed to obtain precise estimates at province level.
- The problem is to get reliable estimates for domains below the planned level because of small sample sizes.
- For 2010-2011:

For men

- the minimum sample size in the domains is 44,
- the first quartile is 69,
- the median 93.

For women

- the minimum sample size in the domains is 48,
- the first quartile is 70,
- the median 88.

Application to Health Data

- **Response variable (Y_d)**: $\hat{Y}_d^{dir}$
- **Auxiliary variable (X_d)**:
  - **Age**: percentage of population under 15 years ($15age$), from 15 to 24 years ($15a24$) , from 25 to 44 years ($25a44$), from 45 to 64 years ($45a64$) and 65 and over ($65age$).
  - **Degree of urbanization**: percentage of population that live in densely-populated area (zdp), intermediate area (zip) and thinly-populated area (zpp).
  - **Activity**: proportion of employed (emp), unemployed ($unemp$) and inactive people ($inac$).
  - **Education level**: proportion of people with low education (low), secundary education (sec) and higher education ($higher.educ$).

Application to Health Data

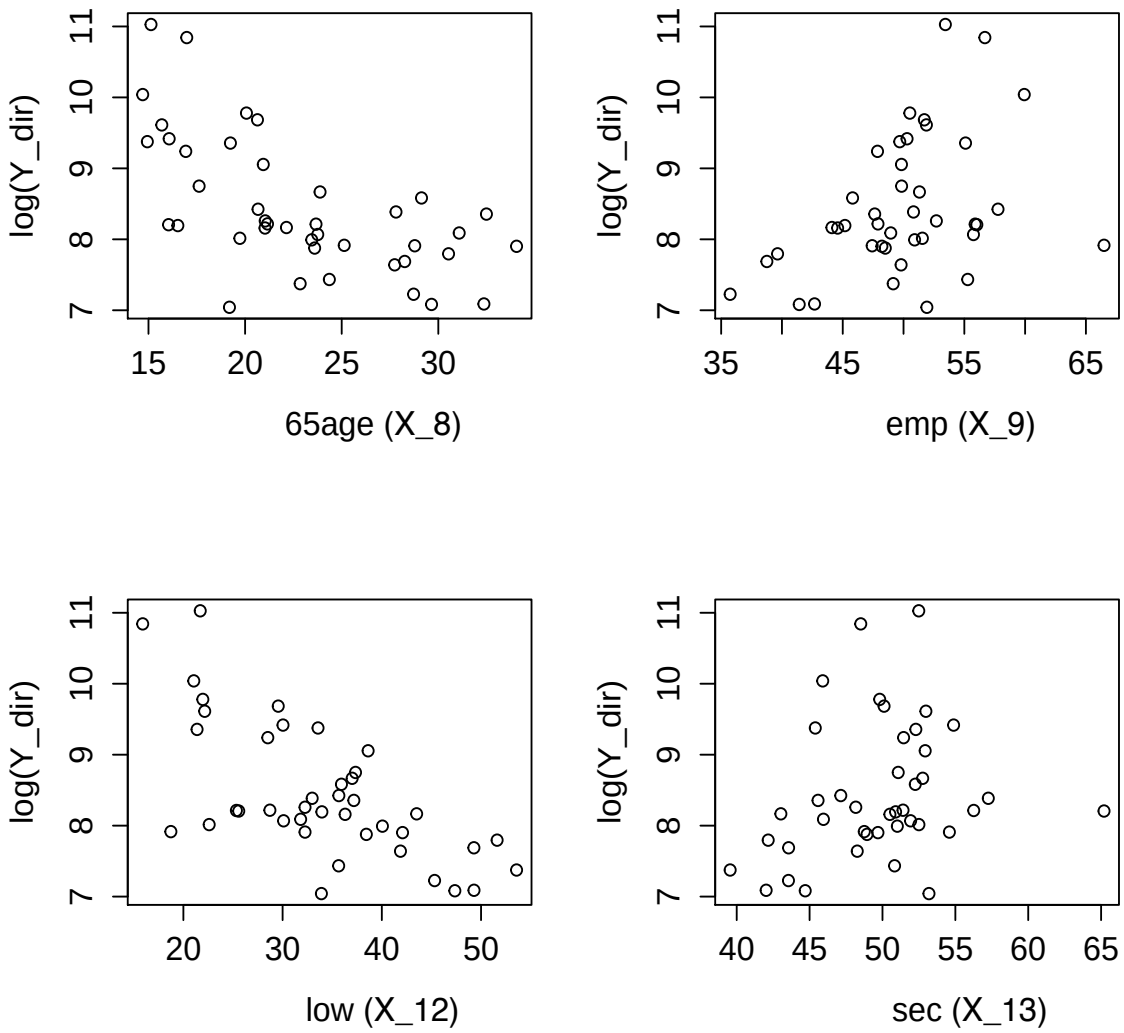


Figure 2: Relation between the auxiliary variables and the response variable ($\log(Y_{dir})$) in men.

Application to Health Data

Model Label	Linear Predictors	Monotone Predictors	P-spline Predictors	$\widehat{cGDF}$	$cGAIC$	$\widehat{xGDF}$	$xGAIC$	$\hat{\sigma}_u^2$
(M1)	X_{12}, X_8			37.2	-18.5	37.1	79.5	0.40
(M2)	X_{12}	X_8		40.9	-14.7	41.1	78.8	0.35
(M3)	X_{12}		X_8	36.5	-18.6	36.4	75.9	0.37
(M4)	X_{12}, X_8, X_{13}			37.1	-18.4	37.4	77.9	0.38
(M5)	X_{12}, X_{13}	X_8		41.0	-14.4	40.8	78.5	0,35
(M6)	X_{12}, X_{13}		X_8	36.7	-18.4	35.4	71.7	0.34
(M7)	X_{12}	X_8, X_{13}		40.9	-14.9	40.8	67.2	0.26
(M8)	X_{12}		X_8, X_{13}	36.2	-17.8	34.2	70.6	0.30
(M9)	X_{12},X_8, X_{13}, X_9			37.4	-18.1	36.7	76.1	0.38
(M10)	X_{12}, X_{13}, X_9	X_8		41.0	-12.9	40.4	69.9	0.28
(M11)	X_{12}, X_{13}, X_9		X_8	36.9	-17.5	34.6	68.9	0.31
(M12)	X_{12}, X_{13}	X_8,X_9		41.0	-14.1	40.8	78.4	0.34
(M13)	X_{12}, X_{13}		X_8,X_9	36.6	-20.0	35.6	68.5	0.31
(M14)	X_{12}	X_8,X_9,X_{13}		40.8	-13.0	40.4	64.9	0.24
(M15)	X_{12}		X_8,X_9,X_{13}	36.2	-17.3	34.0	68.5	0.26

Table 5: Models fitted to men data. $\widehat{GDF}$ and $GAIC$ conditional and mixed values, and $\hat{\sigma}_u^2$.

Application to Health Data

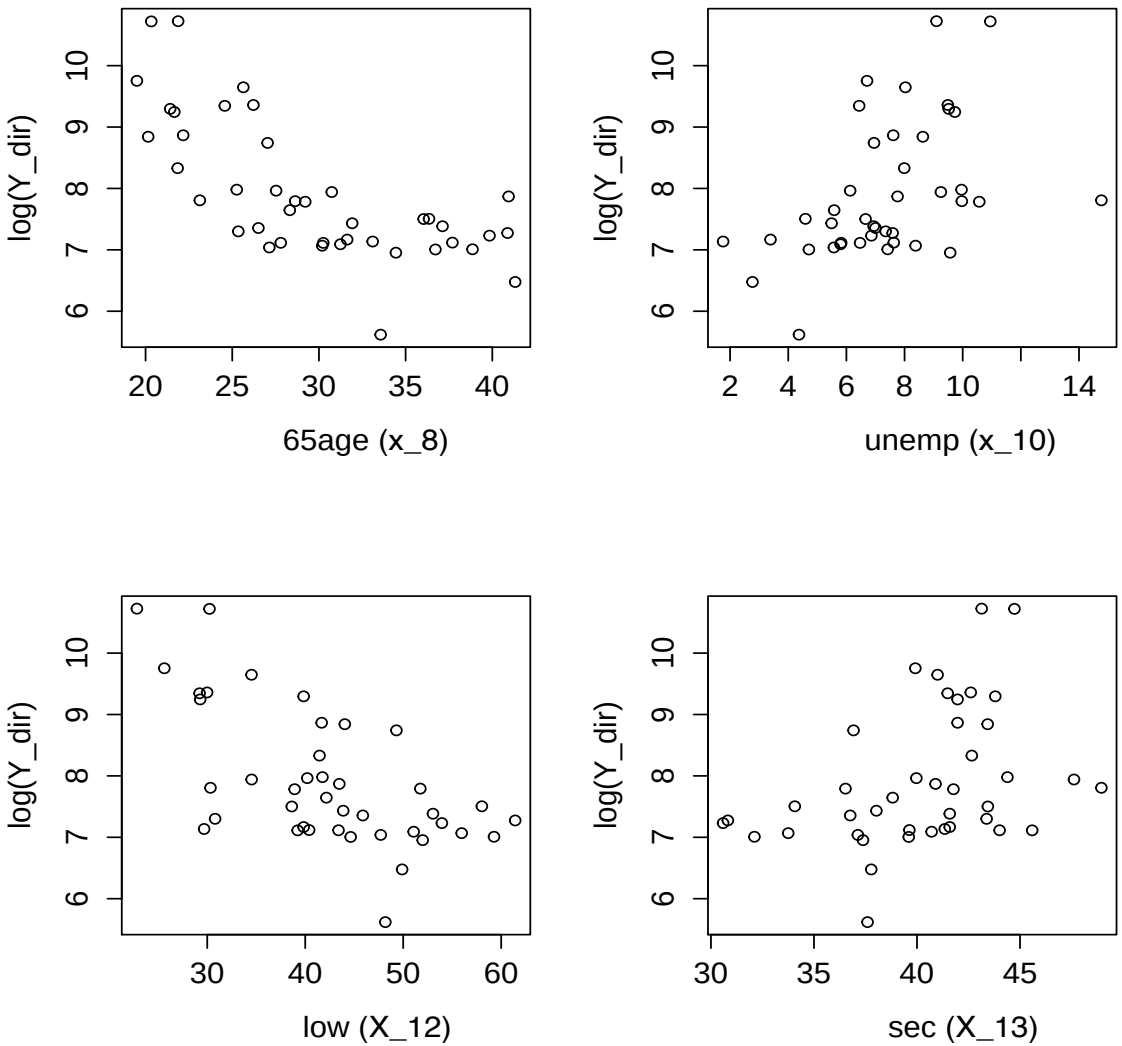


Figure 3: Relation between the auxiliary variables and the response variable ($\log(Y_{dir})$) in women.

Application to Health Data

Model Label	Linear Predictors	Monotone Predictors	P-spline Predictors	$\widehat{cGDF}$	$cGAIC$	$\widehat{xGDF}$	$xGAIC$	$\hat{\sigma}^2_u$
(W1)	X_{12}, X_8			34.7	10.5	35.2	89.5	0.48
(W2)	X_{12}	X_8		40.2	15.4	40.2	83.5	0.35
(W3)	X_{12}		X_8	32.9	10.5	31.5	74.7	0.31
(W4)	X_{12}, X_8, X_{13}			34.4	10.6	34.3	86.1	0.46
(W5)	X_{12}, X_{13}	X_8		40.0	15.7	39.8	82.5	0.35
(W6)	X_{12}, X_{13}		X_8	32.3	10.6	30.8	75.7	0.29
(W7)	X_{12}	X_8, X_{13}		39.8	15.6	40.0	78.7	0.30
(W8)	X_{12}		X_8, X_{13}	31.8	10.0	29.8	70.5	0.29
(W9)	X_{12},X_8, X_{13}, X_{10}			33.7	9.4	34.7	83.0	0.40
(W10)	X_{12}, X_{13}, X_{10}	X_8		39.9	15.8	40.3	79.5	0.30
(W11)	X_{12}, X_{13}, X_{10}		X_8	32.1	10.1	30.9	69.2	0.27
(W12)	X_{12}, X_{13}	X_8,X_{10}		39.4	15.9	39.1	70.5	0.23
(W13)	X_{12}, X_{13}		X_8,X_{10}	31.1	10.2	31.1	66.5	0.22
(W14)	X_{12}	X_8,X_{10},X_{13}		39.4	21.3	38.8	54.5	0.12
(W15)	X_{12}		X_8,X_{10},X_{13}	30.4	9.5	27.9	66.2	0.22

Table 6: Models fitted to women data. $\widehat{GDF}$ and $GAIC$ conditional and mixed values, and $\hat{\sigma}^2_u$.

Section 6 | Conclusions

Conclusions

- $xGAIC$ is a compromise solution derived from a mixed log-likelihood and an empirical estimator of a GDF.
- $xGAIC$ is easily obtained for complex models and it has a good behaviour in SAE applications.
- The simulations have shown that $xGAIC$ performs better than $cGAIC$, when the real model is not linear. This assertion is supported by a quite smaller classification error rate but also by a smaller RRMSE of the random effect variance parameter.
- In the socio-economic case, only one predictor is used being the assumption of linearity fair in this case.Then, the differences between the GAIC values from different candidate models are very small.
- In the health case, several predictors, which can hardly be assumed to have a linear relationship with the response, are considered. The differences between the $xGAIC$ and $cGAIC$ model selection are more important in this case. Being the $\widehat{\sigma_u}$ provided by the $cGAIC$ models, quite higher than that provided by the $xGAIC$ model.

Thank you and ...

Happy Birthday!!!