

Factor-Based GLM Model with Socio-Economic Inputs

Jie Wen

Heriot-Watt University, Edinburgh

School of Mathematical and Computer Sciences

IFoA ARC Workshop

2 December 2019

- 1 Data Overview
- 2 Model Outlook
- 3 Parameter Estimation and Model Selection Criteria
- 4 Variable Selection
- 5 Region/Urban-Rural Pattern
- 6 Summary and Key Takeaways
- 7 Q&A

Multi-population mortality data for England:

- Death and exposure counts at granular level - D_{itx} and E_{itx} available at each individual LSOA i , in year t at age x .
- Overall age range is 0-90 and for years 2001-2016. We concentrate on the 'pensionable' age 65-89.
- Set of socio-economic factors available at LSOA-level, i.e. indices and numerical measures related to income, education, housing affordability, etc.
- Categorical variables indicating certain features of each individual LSOA, i.e. which region is it located in, which urban-rural class does it belong to, etc.

Data Overview (cont.)

Actual-vs-Expected Death Ratio (A-E Ratio):

$$D_i = \sum_{tx} D_{itx} \quad \text{actual deaths at LSOA } i$$
$$\hat{D}_i^0 = \sum_{tx} m_{tx}^b E_{itx} \quad \text{'expected' deaths at LSOA } i$$
$$m_{tx}^b = \frac{\sum_i D_{itx}}{\sum_i E_{itx}} \quad \text{national-level death rate}$$
$$y_i = \frac{D_i}{\hat{D}_i^0} \quad \text{A-E Ratio of LSOA } i$$

It measures the relative mortality at individual LSOAs to national level. A-E ratio helps on lots of analysis s it is a scalar LSOA-level mortality measure, which makes analysis more straight forward.

Data Overview (cont.)

Considerations:

- Individual LSOAs have limited exposure size and are likely to make modelling volatile.
- Deaths and exposures can be combined over all LSOAs. National-level mortality is modelled by the combined data.
- LSOA-level mortalities are modelled via top-down approach starting with national mortality.
- Factors at LSOA-level are used as predictors (input variables) to model the LSOA-specific mortality shift from the national level.
- It limits the volatilities from small death/exposure sizes at individual LSOAs.

Data Overview (cont.)

Some of the LSOA-level socio-economic factors are listed as example. Most of them are dated in 2011 and gender neutral.

Predictive variables / Risk factors	Code
IMD income score	inc
IMD income-old score	incOld
IMD employment score	emp
IMD education score	edu
IMD crime score	cri
IMD barriers to housing and services score	bHS
IMD wider barriers score	widerB
Average number of bedrooms	numb
Prportion of population above age 65 with no qualification	pNoqual65
Ratio of UK-born population	rUK
Proportion of population living in care home with nursing	pCHwithN
Proportion of population living in care home without nursing	pCHnoN
Proportion of population working 49+ hours weekly	work49h

Data Overview (cont.)

- Most of the variables studied have been standardized to be $N(0, 1)$ distributed.
- Deprivation scores are constructed from quite different indicators and some other factors are taken as different measurements, they are at quite different scales from each other.
- Standardization transforms different factors to consistent scale, which avoids the modelling from being significantly distorted.
- Standardization follows uniform and normal transformation step-by-step. Starting from the original factors S :

For certain variable F_i at LSOA $i = 1, 2, 3, \dots, N$:

$$U_i = \frac{\text{rank}(F_i)}{N + 1}$$

$$X_i = \Phi^{-1}(U_i) \quad \Phi \text{ is the probability function of } N(0,1)$$

Model Outlook

- m_{itx} is death rate at LSOA i ;
- m_{tx}^b is national level death rate ("base mortality");
- RR_{itx} is the relative risk factor of LSOA i .

The model has its overall structure as:

$$\log m_{itx} = \log m_{tx}^b + RR_{itx}$$

The national-level base mortality m_{tx}^b can be evaluated by different methods

$$m_{tx}^b = \begin{cases} \text{crude national death rates} \\ \text{OR} \\ \kappa_t^1 + \kappa_t^2(x - \bar{x}) \quad \text{or other mortality models} \end{cases}$$

Relative risk factors RR_{itx} :

Evaluated as linear combination of risk factors at LSOA-level as a typical GLM does.

Model Outlook (cont.)

A simple model structure is:

$$\log m_{itx} = \kappa_t^1 + \kappa_t^2(x - \bar{x}) + \sum_{j=1}^p \beta_j X_{ji}$$

- κ 's capture the overall period effect of mortality at national level.
- β 's are coefficients of the socio-economic risk factors mapping their effects onto mortality of individual LSOAs.
- The national-level mortality is modelled by the CBD structure, while the predictors shift individual LSOAs up and down from national.
- This structure assumes all LSOAs follow the same age pattern as national-level mortalities do.
- This structure assumes factors X have constant effects onto mortality over time.

Model Outlook (cont.)

Now age interaction and time dependency are added to capture more complex patterns.

$$\log m_{itx} = \underbrace{\kappa_t^1 + \kappa_t^2(x - \bar{x})}_{m_{tx}^b} + \underbrace{\sum_{j=1}^P \left(\beta_{j0,t} + \beta_{j1,t}(x - \bar{x}) \right) X_{ji}}_{RR_{itx}}$$

- For each individual predictor j , there are **intercept** β_{j0} and **age interaction** β_{j1} .
- Coefficients vary over time, which capture timely movements of mortality pattern over the socio-economic factors.
- The estimated β 's indicate for how much do their corresponding factors contribute to inequality of mortality over different LSOAs.

Model Outlook (cont.)

The GLM coefficients are estimated by the Maximum Likelihood Estimation (MLE) approach with the deaths and exposures data, D_{itx} and E_{itx} . Likelihood functions are constructed based on Poisson assumption for the number of deaths (death frequency).

$$D_{itx} \sim \text{Poisson}(\hat{D}_{itx})$$
$$\hat{D}_{itx} = E_{itx} \hat{m}_{itx}$$

The estimated death rate \hat{m}_{itx} is modelled by the GLM-type models with deaths, exposures and socio-economic factors as input data. The Poisson likelihood function is therefore constructed as:

$$L(\theta|D, E) = \prod_{itx} \frac{e^{-\hat{D}_{itx}} \hat{D}_{itx}^{D_{itx}}}{D_{itx}!}$$

Parameter Estimation and Model Selection Criteria

- The main model selection criteria is BIC, defined as $k \log N - 2 \log \hat{L}$, with number of unknown parameters k and sample size N (smaller BIC is better following the definition).
- BIC penalizes overfitting if too many redundant factors are included.
- Many different sets of factors are tested and those giving most preferable BICs are picked up for further study.

List of combinations with most preferable BIC:

<u>X</u>	LL	BIC
incOld+emp+pNoqual65+widerb	-6,569,457	13,141,537
incOld+pNoqual65+numb+rUK	-6,569,646	13,141,914
incOld+pNoqual65+numb+widerb	-6,569,364	13,141,351
incOld+emp+pNoqual65+numb+rUK	-6,569,263	13,141,674

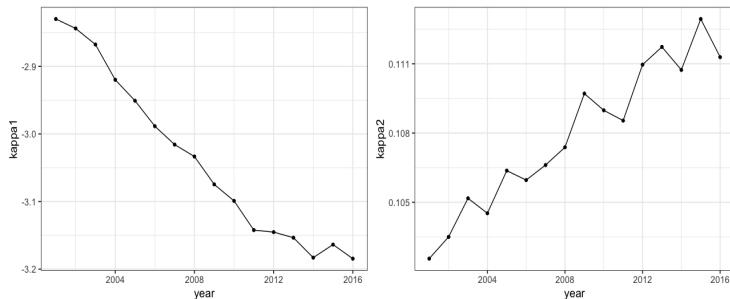
Parameter Estimation and Model Selection Criteria (cont.)

The estimated parameters under the BIC-optimized profile:
incOld+pNoqual65+numb+widerb

$$\begin{aligned}\log m_{itx} = & \kappa_t^1 + \kappa_t^2(x - \bar{x}) + (\beta_{10,t} + \beta_{11,t}(x - \bar{x}))X_1 \\ & + (\beta_{20,t} + \beta_{21,t}(x - \bar{x}))X_2 \\ & + (\beta_{30,t} + \beta_{31,t}(x - \bar{x}))X_3 \\ & + (\beta_{40,t} + \beta_{41,t}(x - \bar{x}))X_4\end{aligned}$$

Parameter Estimation and Model Selection Criteria (cont.)

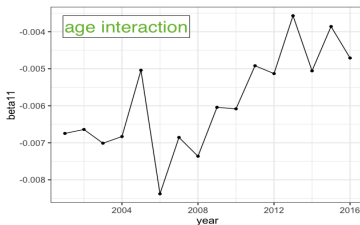
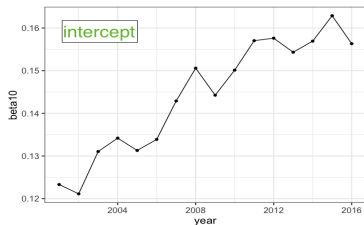
The common κ 's at national level:



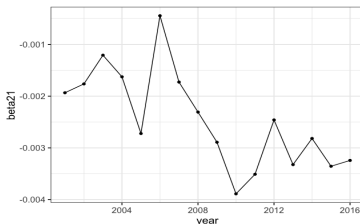
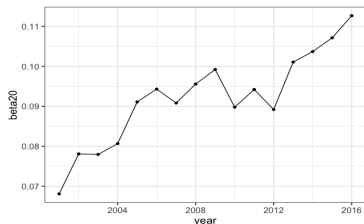
- κ^1 shows the declining mortality at national level over 2001-2016, with improvement slow down slightly around 2011.
- κ^2 being positive shows there has been bigger improvement at young ages (below the mean age).

Parameter Estimation and Model Selection Criteria (cont.)

IMD income-old score (*incOld*):

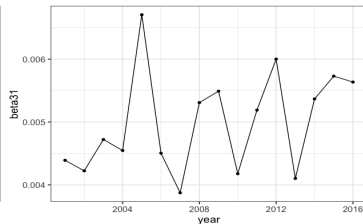
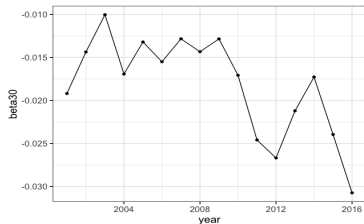


Proportion of no-qualification population above 65 (*pNoqual65*):

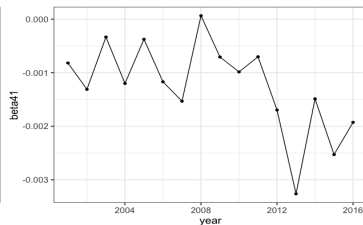
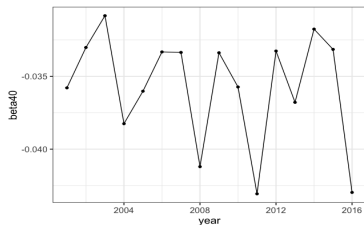


Parameter Estimation and Model Selection Criteria (cont.)

Average number of bedrooms (*numb*):



IMD wider barriers score (*widerb*):



Parameter Estimation and Model Selection Criteria (cont.)

Conclusions from estimated coefficients

Reminder of the 4 underlying factors:

- incOld is IMD income-old deprivation score indicating proportion of population receiving relevant type of benefit due to income shortage.
- pNoqual65 is proportion of population over age 65 with no degree qualification.
- numb is the average number of bedrooms per household.
- widerb is the IMD wider-barriers deprivation score indicating some housing related facts: overcrowding, affordability, etc.

Parameter Estimation and Model Selection Criteria (cont.)

- β_{10} and β_{20} have increasing trend indicating increasing contribution from *incOld* and *pNoqual65* to inequality of mortality across different LSOAs over 2001-2016.
- β_{30} and β_{40} do not have clear pattern and their fluctuations are more likely to be from sampling variation.
- LSOAs with large value of *incOld* and *pNoqual65* tend to have higher mortality.
- *widerb* has negative sign over all years, which is counterintuitive. Few reasons drive its behavior: effects from London-based LSOAs, dependency between different risk factors (e.g. with *incOld*), etc.

Parameter Estimation and Model Selection Criteria (cont.)

Considerations on the estimated coefficients β :

- Except for β_{10} and β_{20} , the other coefficients do not have any significant pattern over time.
- Their fluctuations over time are more likely to be from sampling noise rather than crude mortality pattern.
- They can be simplified to constant coefficients for mitigating over-parameterization.
- BIC is improved to 13, 140, 030 from 13, 141, 351 after the simplification. The simplified model has formula of:

$$\begin{aligned}\log m_{itx} = & \kappa_t^1 + \kappa_t^2(x - \bar{x}) + (\beta_{10,t} + \beta_{11}(x - \bar{x})) \times \text{incOld} \\ & + (\beta_{20,t} + \beta_{21}(x - \bar{x})) \times \text{pNoqual65} \\ & + (\beta_{30} + \beta_{31}(x - \bar{x})) \times \text{numb} \\ & + (\beta_{40} + \beta_{41}(x - \bar{x})) \times \text{widerb}\end{aligned}$$

1. Shrinkage

For variable selection purpose, ridge regression and LASSO are two commonly used machine learning approaches. They are based on penalized model fitting criteria, i.e. MLE or OLS.

Ridge Regression Penalized Log-likelihood

$$l^R(\underline{\beta}) = l(\underline{\beta}) - \lambda \sum_{j=1}^p \beta_j^2$$

LASSO Penalized Log-likelihood

$$l^L(\underline{\beta}) = l(\underline{\beta}) - \lambda \sum_{j=1}^p |\beta_j|$$

Variable Selection (cont.)

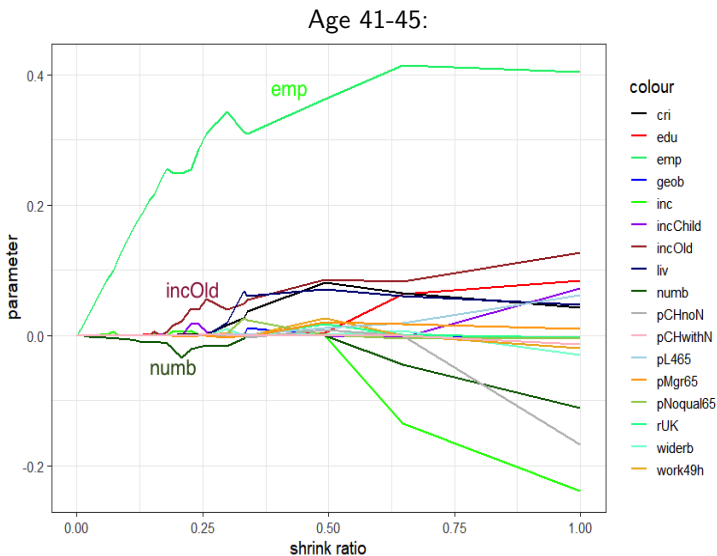
Simplified GLM (no age interaction or time dependency) is used to implement Poisson MLE for LASSO, as the target is to select important variables instead of modelling accurately.

$$\log m_{itx} = \log m_{tx}^b + \beta_0 + \sum_{j=1}^p \beta_j X_{ji}$$

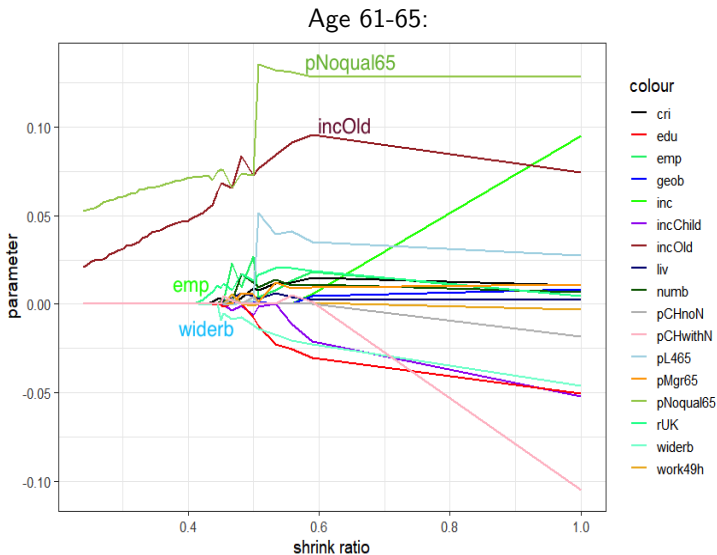
with m_{tx}^b representing the base mortality derived from crude observations taking all LSOAs into account, $m_{tx}^b = \frac{\sum_i D_{itx}}{\sum_i E_{itx}}$.

The simplified model should be fitted over relatively narrow age buckets as there is no age interaction, i.e. 5-year or 10-year.

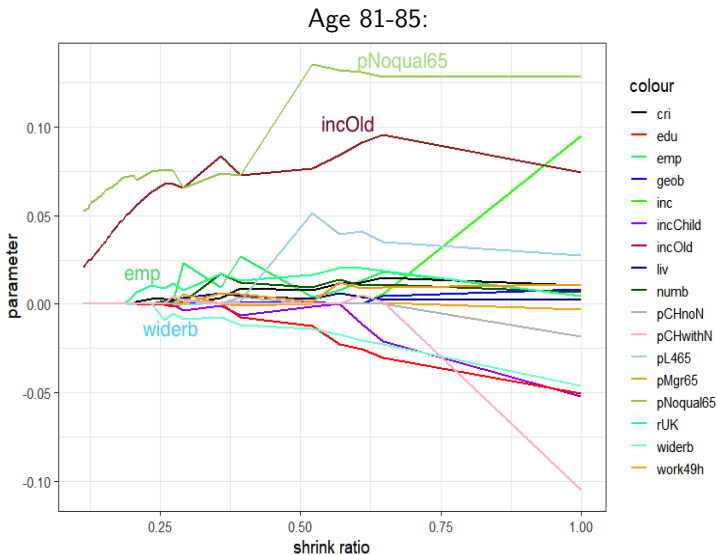
Variable Selection (cont.)



Variable Selection (cont.)



Variable Selection (cont.)



Variable Selection (cont.)

LASSO suggests the important variables over different age buckets to be:

- 41-45: emp, incOld, numb
- 61-65: pNoqual65, incOld, emp, widerb
- 81-85: pNoqual65, incOld, emp, widerb

Variables suggested to be important by LASSO have been broadly consistent with those having the most preferable BIC of the GLM model.

2. Regression Tree and Random Forest (RF)

- Tree-based supervised learning methods (non-parametric methods, no closed-form formula) recognise complex patterns in labelled datasets.
- They can be alternative to BIC, LASSO, etc. for selecting the most important variables.
- They split the training dataset (LSOAs) by values of input variables \mathbf{X} into different groups, all observations within the same group have the same regression output \hat{y} .
- Input response variable y here is set by the A-E ratio over a certain age range to represent mortality level of individual LSOAs.
- All in all, the trees are all about the non-linear regression problem: $\hat{y} = f(X_1, X_2, \dots, X_p)$, while f is a non-linear tree function.

Variable Selection (cont.)

- Each split in the tree is called a 'node', the two distinct groups under each node are called 'leaves'.
- The final groups splitted are called the 'terminal nodes'.
- Within each terminal node (group), the estimated \hat{y} is calculated as the average of all observations in that node, i.e.

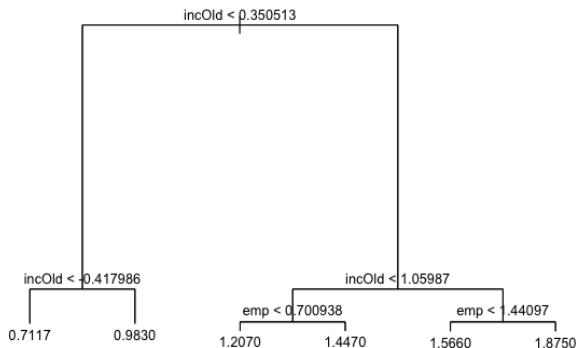
$$\hat{y}_i = \frac{\sum_{i \in R} y_i}{N_R} \quad \forall i \in R$$

- Variable to split at each node is selected such that the overall mean-squared error (MSE) is reduced the most. Splitting stops when certain stopping criterion is achieved.

Variable Selection (cont.)

Single regression tree:

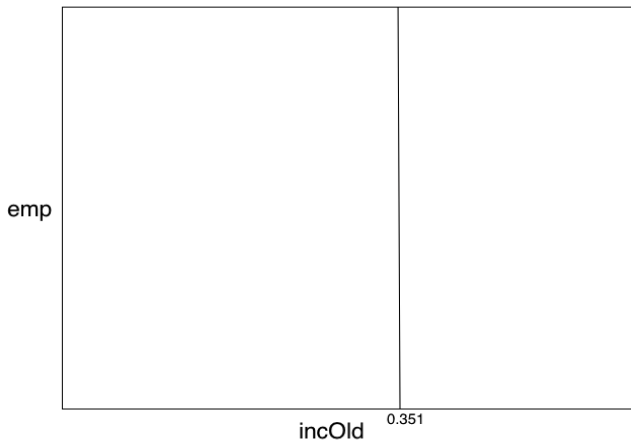
$y = \text{A-E ratio of age 60-69}; \hat{y} = f(X_1 = \text{incOld}, X_2 = \text{emp})$



Variable Selection (cont.)

Single regression tree:

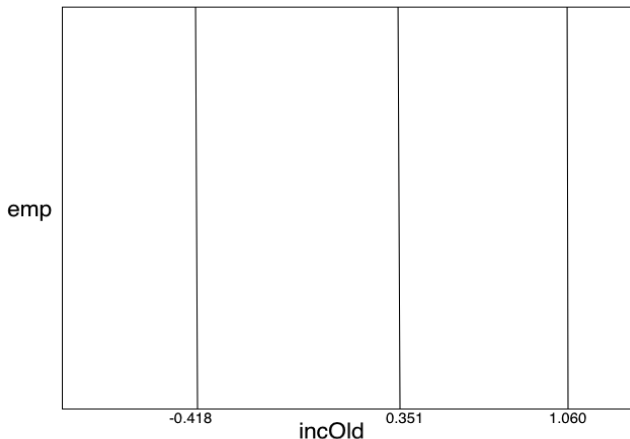
Splitting of nodes shown in the 2D plan of ($X_1 = incOld$, $X_2 = emp$):



Variable Selection (cont.)

Single regression tree:

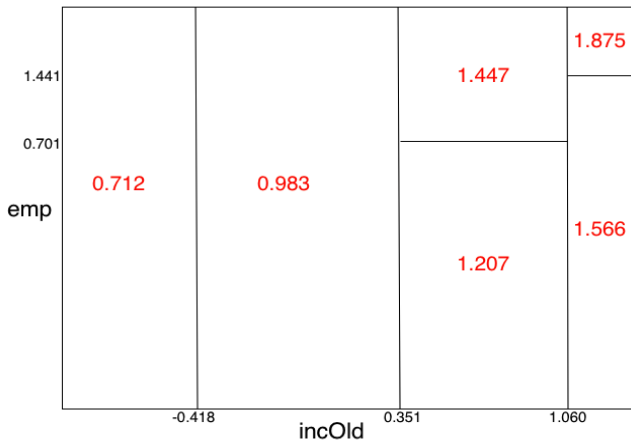
Splitting of nodes shown in the 2D plan of ($X_1 = incOld$, $X_2 = emp$):



Variable Selection (cont.)

Single regression tree:

Splitting of nodes shown in the 2D plan of ($X_1 = incOld$, $X_2 = emp$):



Variable Selection (cont.)

- RF grows a set of trees, with each tree grown from a randomly selected subset of observations and variables out of the training dataset.
- Output of RF is calculated as an average over outputs from all individual trees in the forest, $\hat{y} = \frac{\sum_{t=1}^T \hat{y}^t}{T}$
- RF can be used for regression (when y is numerical) or classification (when y is categorical) of new observations. We apply them here to investigate variable importance.
- Other tree-based regression tools are available and might lead to further improvements in accuracy, i.e. Gradient Boosting Machine (GBM).

Variable Selection (cont.)

$$\begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} & \dots & X_{1,p-1} & X_{1p} \\ X_{21} & X_{22} & X_{23} & X_{24} & \dots & X_{2,p-1} & X_{2p} \\ X_{31} & X_{32} & X_{33} & X_{34} & \dots & X_{3,p-1} & X_{3p} \\ X_{41} & X_{42} & X_{43} & X_{44} & \dots & X_{4,p-1} & X_{4p} \\ X_{51} & X_{52} & X_{53} & X_{54} & \dots & X_{5,p-1} & X_{5p} \\ X_{61} & X_{62} & X_{63} & X_{64} & \dots & X_{6,p-1} & X_{6p} \\ X_{71} & X_{72} & X_{73} & X_{74} & \dots & X_{7,p-1} & X_{7p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \\ X_{N-1,1} & X_{N-1,2} & X_{N-1,3} & X_{N-1,4} & \dots & X_{N-1,p-1} & X_{N-1,p} \\ X_{N1} & X_{N2} & X_{N3} & X_{N4} & \dots & X_{N,p-1} & X_{Np} \end{pmatrix}$$

Red: randomly selected LSOAs with variables used to grow one tree

Black: LSOAs not included in growing of the tree, which are called out-of-bag (OOB) samples.

Variable Selection (cont.)

RF produces two different measures of variable importance under two aspects of model criteria.

- Percentage increase in MSE: increase in out-of-bag MSE after a specific variable is randomly shuffled (after trees are grown), as a percentage of the original out-of-bag MSE. It is calculated for each individual tree and averaged over all trees (out-of-sample).
- Increase in node purity: total reduction of in-sample MSE from all splits in a tree that is driven by a specific variable, averaged over all trees (in-sample).

The first one is more of interest as it is directly related to the out-of-bag performance.

Variable Selection (cont.)

RF with A-E ratio over age 60-69 as the response variable y , over the set of 27 explanatory variables in \mathbf{X} :

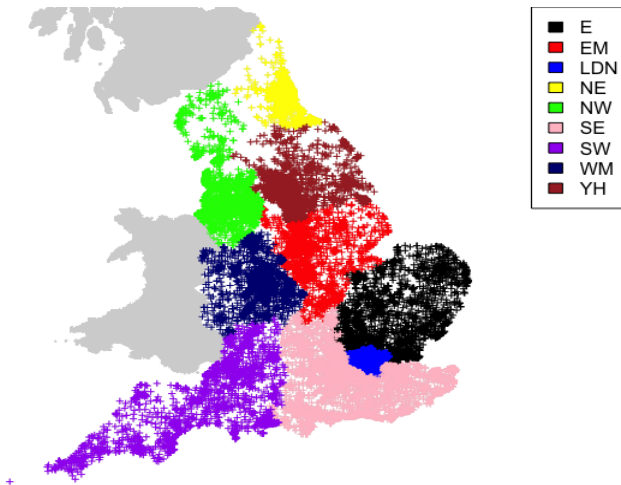
Variable	%IncMSE	Variable	IncNodePurity
pCHwithN	108.89	incOld	652.95
rUK	74.64	inc	456.64
pNoqual	59.48	emp	367.77
numb	53.76	incChild	273.15
work49h	48.23	pNoqual65	268.75
geob	47.52	numb	203.69
incOld	46.54	pMgr	172.00
liv	45.24	edu	145.74
bHS	45.18	pMgr65	135.03
pMgr65	43.57	adultSkill	113.10

Nine Geographical regions:

- East
- East Midlands
- London
- North East
- North West
- South East
- South West
- West Midlands
- Yorkshire and The Humber

Region/Urban-Rural Pattern (cont.)

Distribution of LSOAs over the 9 Regions:



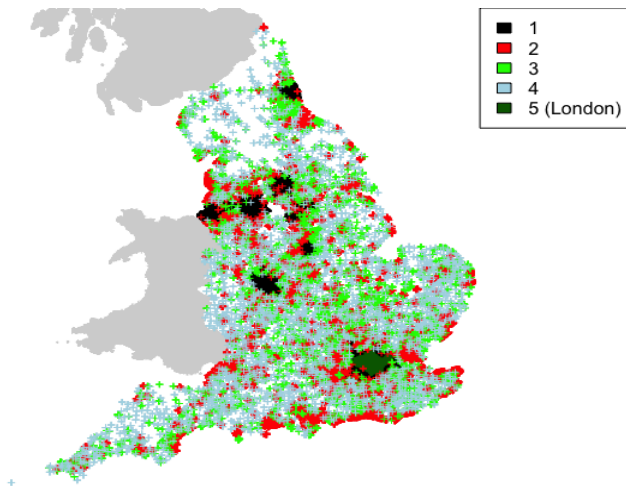
Region/Urban-Rural Pattern (cont.)

Five urban-rural classes:

- Class 1: Urban conurbation (except London)
- Class 2: Urban city and town
- Class 3: Rural town and fringe
- Class 4: Rural village and dispersed
- Class 5: Urban conurbation in London

Region/Urban-Rural Pattern (cont.)

Distribution of LSOAs over the urban-rural classes:



1. Principal Component Analysis (PCA)

PCA is an unsupervised learning approach dealing with unlabelled datasets (while RF requires 'labelled' data) that extracts informative features while abandoning redundant information from a set of factors, $\mathbf{X} = (X_1, X_2, \dots, X_p) (N \times p)$.

- It constructs p principal components (PC) as linear combinations of the input variables \underline{X} .
- PCs can be considered as new, transformed variables from \mathbf{X} with selected information that have maximum variation. They are not correlated with each other ("orthogonal").
- The first PC (PC1) explains the highest proportion of total variation in the data, the second PC (PC2) explains the second highest, etc.

Region/Urban-Rural Pattern (cont.)

Algebraic principles - for LSOA $i = 1, 2, \dots, 32,844$ and with p variables $X_{1i}, X_{2i}, \dots, X_{pi}$:

$$P_i^1 = \phi_{11}X_{1i} + \phi_{21}X_{2i} + \dots + \phi_{p1}X_{pi}$$

$$P_i^2 = \phi_{12}X_{1i} + \phi_{22}X_{2i} + \dots + \phi_{p2}X_{pi}$$

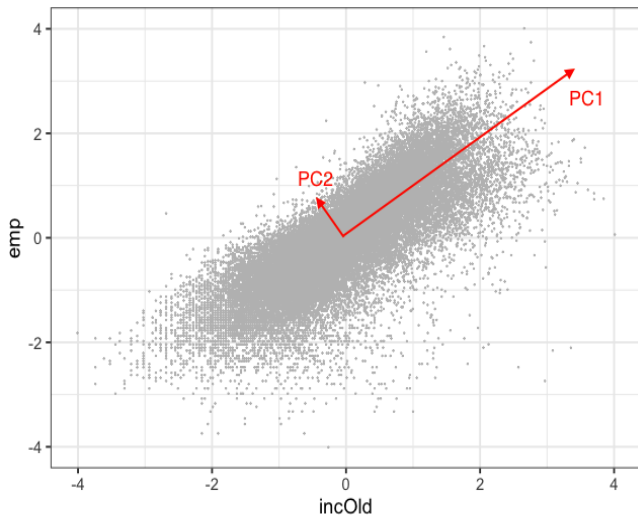
...

$$P_i^p = \phi_{1p}X_{1i} + \phi_{2p}X_{2i} + \dots + \phi_{pp}X_{pi}$$

Vice versa, each of the variables \mathbf{X} can be written as linear combinations of the principle components based on the same set of loading vectors ϕ .

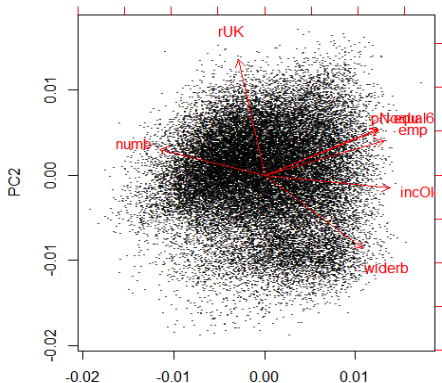
Region/Urban-Rural Pattern (cont.)

2-dimension data matrix $\mathbf{X} = (X_1, X_2)$ as example: $X_1 = \text{incOld}$, $X_2 = \text{emp}$



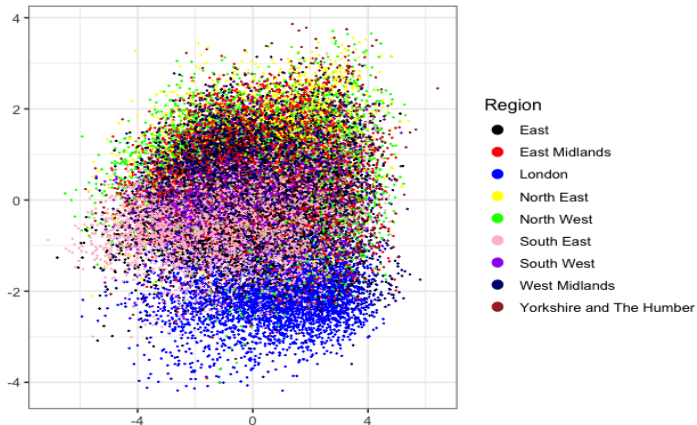
Region/Urban-Rural Pattern (cont.)

More complex example: projection of all LSOAs onto the 2D plane spanned by the first two PCs (almost 90% of variation explained); Red arrows: projection of the 7 predictors consisting of \mathbf{X} decomposed onto the same 2D plain. 7 variables are in \mathbf{X} : incOld, emp, edu, pNoqual65, numb, rUK, widerb



Region/Urban-Rural Pattern (cont.)

Projection of LSOAs onto PC1-PC2 plain with colour scheme specifying the 9 Regions.



Region/Urban-Rural Pattern (cont.)

- There is heterogeneity across different Regions or Urban-Rural classes.
- The model is fitted over LSOAs in separate Regions/Urban-Rural classes separately.
- PCA has revealed significant heterogeneity between London-based LSOAs and the others.
- Urban-Rural class (1 to 5) describes the level of urbanization from large cities to rural villages over class 1 to 4. London-based LSOAs form a separate class 5.
- Variables selected by BIC and Lasso generally shows quite a heavy contribution to the first 2 PCs capturing most of the data variances. This proves their importance to some extent.

2. Region/UR specific GLM model

A distinct model for each individual region or urban-rural class:

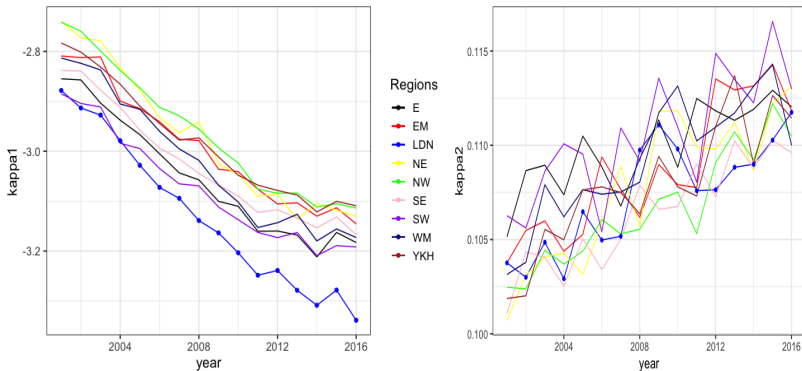
$$\log m_{itx} = \kappa_{t,r(i)}^1 + (x - \bar{x})\kappa_{t,r(i)}^2 + \sum_j \left(\beta_{j0,r(i)} + \beta_{j1,r(i)}(x - \bar{x}) \right) X_{ji,r(i)}$$

$r(i)$ is the corresponding Region or Urban-Rural class of LSOA i . It is equivalent to decomposing the national-level CBD model into a multi-population version with groups of LSOAs.

Region/Urban-Rural Pattern (cont.)

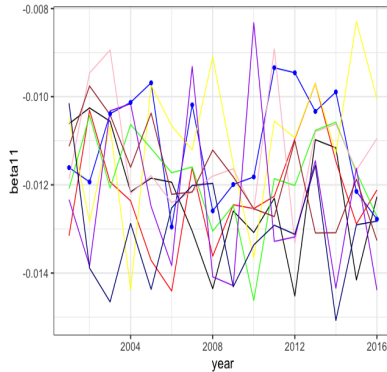
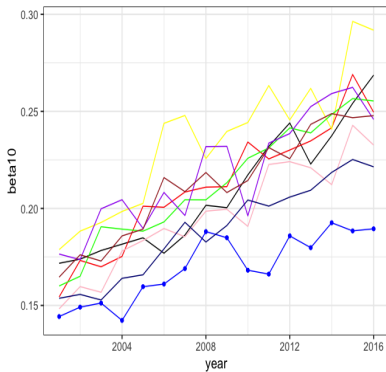
The CBD-GLMx model fitted with *incOld* only - over age 65-89 for LSOAs in different Regions separately.

Period-effect factors κ at Region-level CBD term:



Region/Urban-Rural Pattern (cont.)

Income-related coefficients β_{10} and β_{11} :



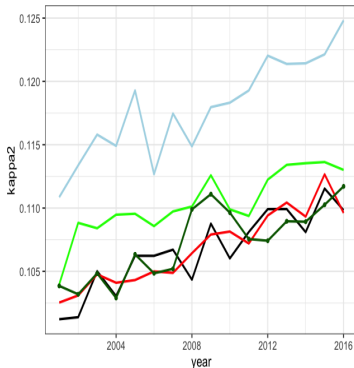
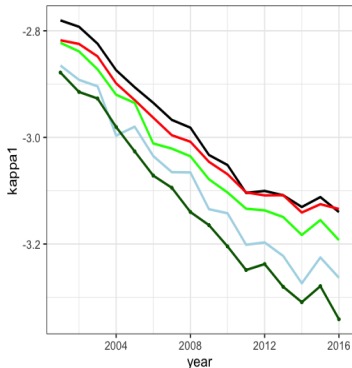
Region/Urban-Rural Pattern (cont.)

- There is significant heterogeneity across separate Regions shown by both κ 's and the β_{10} coefficients with *incOld*.
- London has the lowest baseline time factor κ^1 , and have mortality improvement not affected as much by the 2011 slowdown as other regions - the most significant improvement over 2001-2016 appears in London.
- North West and North East are the two with highest mortalities over 2001-2016.

Region/Urban-Rural Pattern (cont.)

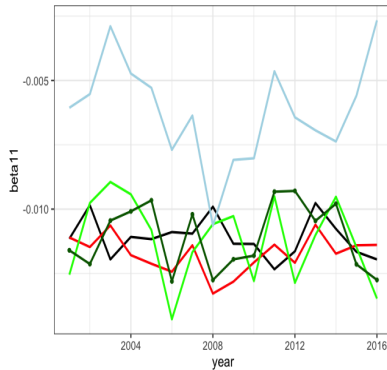
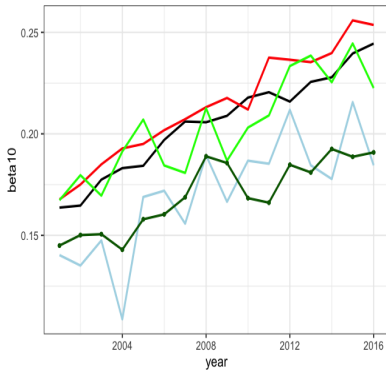
The same model with *incOld* only is also fitted over LSOAs in separate urban-rural classes (1 to 5).

Period-effect factor κ 's of UR-level CBD term:



Region/Urban-Rural Pattern (cont.)

Income-related coefficients β_{10} and β_{11} :



Conclusions of region/urban-rural analysis

- The London-based LSOAs have significantly different behaviour from other LSOAs in class 1 (conurbations). Allowing for London effect in the GLM could improve the model's predictive power.
- The 2011 slowdown of mortality improvement affects LSOAs in urban-rural class 1 to 3 more than 4 and 5.
- Regions with higher mortality have had least improvements over 2001-2016, which widens the gap of mortality over high and low-mortality areas.
- The LSOAs in class 4 (least urbanized) are most sensitive to age, reflected by the two age-interacted parameters κ^2 .

Highlights

- Income, employment, education and housing related factors have relatively strong predictive power to mortality.
- Income is one of the main contributors to the increasing inequality of mortality across different areas.
- There are significant heterogeneity of mortality over different regions and urban-rural classes in England.
- London-based LSOAs have significantly different behaviour from most of the others with respect to socio-economic factors.

Key Takeaways (cont.)

Future Researches

- More research on the selective age interaction and time dependency for GLM model - some variables might not need them.
- Based on the current findings, build up the GLM model with the most appropriate set of variables including socio economic factors and other factors such as spatial information (region/urban-rural) that gives the best BIC.

ANY QUESTIONS?