



Introduction à l'Apprentissage Statistique

Modèles Linéaires

M2 Actuariat – ISFA – 2021/2022

Pierrick Piette
Actuaire à Seyna
pierrick.piette@gmail.com

Seyna.

In God we trust, all others bring data
- William Edwards Deming





Régression linéaire

Objectif de la régression

On observe n réalisations de $(\mathbf{X}, Y) \in \mathbb{R}^{p+1}$

- Y est la variable d'intérêt ou expliquée
- \mathbf{X} sont les variables explicatives

L'idée est très simple (et c'est ce qu'on va faire le reste du cours)

- Décrire Y en fonction de \mathbf{X}

$$Y = f(\mathbf{X})$$

Idée simple, mais problème difficile à résoudre.

Régression linéaire simple

Expliquer Y en fonction de X_1, \dots, X_p

- Comprendre l'impact des variables explicatives
- Prédiction de la variable d'intérêt en supposant une relation linéaire

$$Y = \beta_0 + \sum_{k=1}^p \beta_k X_k + \epsilon$$

Modèle simple mais avec des hypothèses

- Non colinéarité des variables explicatives
- Erreurs indépendantes et normales $\mathcal{N}(0, \sigma^2)$
- Exogénéité et Homoscédasticité

Estimation des moindres carrés

Estimation des coefficients en minimisant les erreurs quadratiques

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{k=1}^p \beta_k x_{i,k} \right)^2$$

Dans le cadre classique on a l'estimateur MCO

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Propriété

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$$

Test d'hypothèses

Rappel sur les tests d'hypothèses en statistique classique

- On a deux hypothèses que l'on veut tester l'une contre l'autre
- Hypothèse *null* H_0
- contre l'hypothèse alternative H_1
- On accepte ou on rejette l'hypothèse *null*

	Accepte H_0	Rejette H_0
H_0 est vraie	✓	Faux positif (type I)
H_1 est vraie	Faux négatif (type II)	✓

Introduction de la notion de fausse alarme

p-value

Accepter ou rejeter n'est pas très informatif

- Si je rejette à un niveau α , alors je vais aussi le rejeter à un niveau $\alpha' > \alpha$

Définition de la *p*-value

- La probabilité d'obtenir sous H_0 une valeur plus extrême que celle observée
- Plus la *p*-value est faible, plus l'évidence contre H_0 est forte

Niveaux de *p*-value

- En économétrie, on note les *p*-value avec des symboles pour les différents niveaux de confiance

Niveau de <i>p</i> -value	Notation
< 0.001	***
0.001 – 0.01	**
0.01 – 0.05	*
0.05 – 0.10	.

La guerre des étoiles

Le mythe des 5%

- Seuils « imposés » par Fisher dans les années 1920
- « we shall not often be astray if we draw a conventional line at 5% »
- Dogme qui est resté mais qui commence à être remis en question

Comprendre la p -value pour ce qu'elle est

- Une valeur continue
- On ne compare pas des étoiles mais des probabilités

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	HIGHLY SIGNIFICANT
0.02	HIGHLY SIGNIFICANT
0.03	HIGHLY SIGNIFICANT
0.04	SIGNIFICANT
0.049	SIGNIFICANT
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	ON THE EDGE OF SIGNIFICANCE
0.07	HIGHLY SUGGESTIVE,
0.08	SIGNIFICANT AT THE P<0.10 LEVEL
0.09	SIGNIFICANT AT THE P<0.10 LEVEL
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	THIS INTERESTING SUBGROUP ANALYSIS



Vision économétrique

Estimation du taux de criminalité

- On teste $H_0 : \beta_k = 0$ versus $H_1 : \beta_k \neq 0$
- basé sur le test de Student

Covariate	$\widehat{\beta}_j$	$\widehat{se}(\widehat{\beta}_j)$	t value	p-value
(Intercept)	-589.39	167.59	-3.51	0.001 **
Age	1.04	0.45	2.33	0.025 *
Southern State	11.29	13.24	0.85	0.399
Education	1.18	0.68	1.7	0.093
Expenditures	0.96	0.25	3.86	0.000 ***
Labor	0.11	0.15	0.69	0.493
Number of Males	0.30	0.22	1.36	0.181
Population	0.09	0.14	0.65	0.518
Unemployment (14–24)	-0.68	0.48	-1.4	0.165
Unemployment (25–39)	2.15	0.95	2.26	0.030 *
Wealth	-0.08	0.09	-0.91	0.367



Sélection de modèle

Choix des variables : la p-value ?

Covariate	$\hat{\beta}_j$	$\hat{s.e}(\hat{\beta}_j)$	t value	p-value
(Intercept)	-589.39	167.59	-3.51	0.001 **
Age	1.04	0.45	2.33	0.025 *
Southern State	11.29	13.24	0.85	0.399
Education	1.18	0.68	1.7	0.093
Expenditures	0.96	0.25	3.86	0.000 ***
Labor	0.11	0.15	0.69	0.493
Number of Males	0.30	0.22	1.36	0.181
Population	0.09	0.14	0.65	0.518
Unemployment (14–24)	-0.68	0.48	-1.4	0.165
Unemployment (25–39)	2.15	0.95	2.26	0.030 *
Wealth	-0.08	0.09	-0.91	0.367

Le problème de la *p*-value

« A key issue with applying small-sample statistical inference to large samples is that even minuscule effects can become statistically significant. The increased power leads to a dangerous pitfall as well as to a huge opportunity. The issue is one that statisticians have long been aware of : the *p*-value problem. [...] The question is not whether differences are significant (they nearly always are in large samples, but whether they are interesting. Forget statistical significance, what is the practical significance of the results ? » M. Lin, H. Lucas, G. Shmueli, 2010.

Un coefficient de corrélation égale à 0,002 est significativement différent de 0 si $n = 10^6$, mais il est totalement inutile.

« Statistical significance plays a minor or no rôle in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, even if they are statistically significant, results in improved prediction accuracy. » M. Lin, H. Lucas, G. Shmueli, 2010.



Qualité de la régression linéaire

En régression linéaire un critère de qualité de modèle est $R^2 = 1 - SSR/SST$

$$\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Implication pour les modèles linéaires

- Terme de gauche : variation totale (*Sum of Squares Total*)
- 1^{er} terme de droite : variation expliquée par les résidus (*Sum of Squared Residuals*)
- 2^e terme de droite : variation expliquée par la régression (*Sum of Squares Explained*)

Si le R^2 est proche de 1, le modèle est « bon »

Erreur d'un modèle

Décomposition des erreurs de prédiction d'un modèle $y = f(x) + \epsilon$

$$E \left[(y - \hat{f}(x))^2 \right] = Bias[\hat{f}(x)]^2 + Var[\hat{f}(x)] + \sigma^2$$

Avec

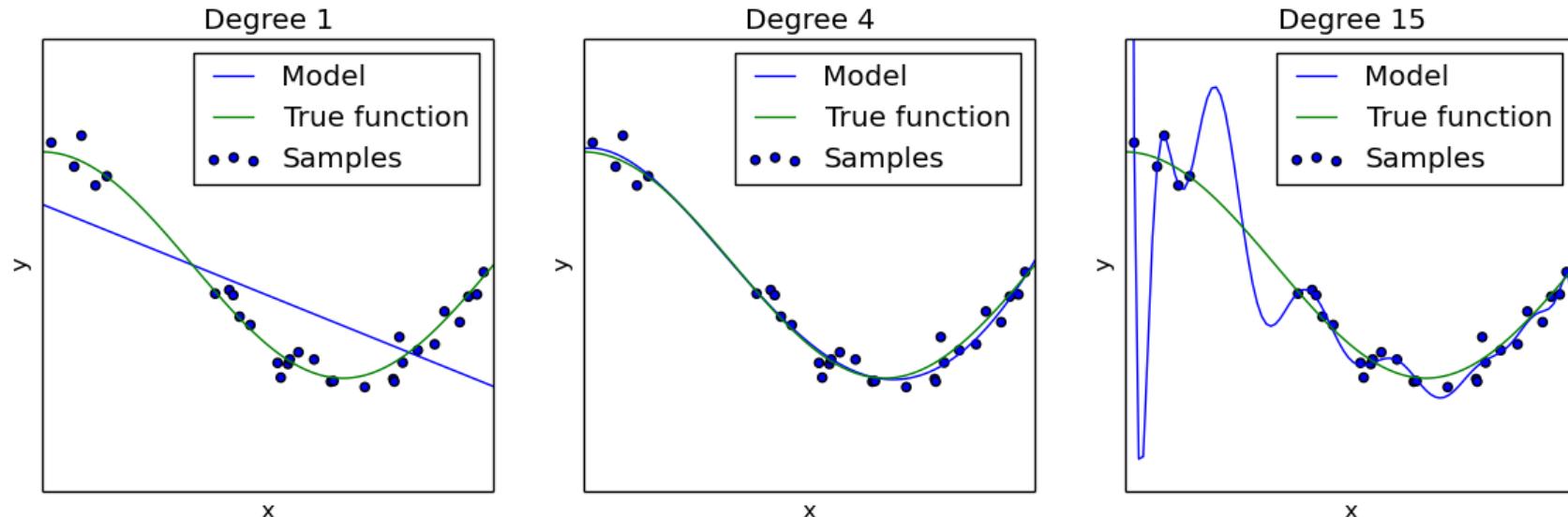
$$Bias[\hat{f}(x)] = E[\hat{f}(x) - f(x)]$$

$$Var[\hat{f}(x)] = E \left[(\hat{f}(x) - E[\hat{f}(x)])^2 \right]$$

Dilemme biais-variance

Implication pour les modèles linéaires

- Pas assez de variables explicatives : on rate des effets → Biais important
- Trop de variables explicatives : on sur-apprend des données → Variance importante
- Trouver un juste milieu : LA grande question de l'apprentissage statistique



A GOOD EXAMPLE OF



Fléau de la dimension

Dimensions du problème de la régression

- n : nombre d'observations ou taille de l'échantillon
- p : nombre de variables observées sur cet échantillon

Si n est grand

- pas de problème a priori, bien au contraire pour la théorie asymptotique !
- problème de puissance de calculs, mais pas un problème mathématique

Si p est grand

- Les estimateurs conservent les propriétés de normalité asymptotique si $p^2/n \rightarrow 0$, lorsque $p, n \rightarrow \infty$
- Données en grande dimension : $p > \sqrt{n}$

Quel problème de convergence ?

Conditionnement de $\mathbf{X}^T \mathbf{X}$

- Non colinéarité des variables explicatives

Si $n \gg p$

- les chances que la corrélation entre les variables explicatives soit très grande diminue
- $\mathbf{X}^T \mathbf{X}$ est symétrique définie positive, donc inversible
- Le problème a une solution $\hat{\boldsymbol{\beta}}$ unique !

Si les variables ont une forte corrélation

- mauvais conditionnement de $\mathbf{X}^T \mathbf{X}$, les coefficients de régression sont très élevés en valeur absolue sur certains facteurs avec des signes peu intuitif (effet de compensation)
- Grande sensibilité de la solution $\hat{\boldsymbol{\beta}}$ à de faibles variations de \mathbf{X} ou \mathbf{Y}

Réduction de la dimension

Pour éviter ces problèmes de convergence, on réduit la dimension

- partir de la dimension initiale p
- pour arriver à une dimension $d < p$

Première solution évidente

- on ne garde que d variables explicatives
- sélection de modèle parmi toutes les possibilités

Seconde solution en statistique « classique »

- création de nouvelles variables regroupant les variables initiales
- Analyse en composantes principales

Pénalisation du R^2

Problème du R^2 classique

- Fonction croissante de d
- Plus la dimension est grande, plus le SSR diminue
- Vision explicative, et non prédictive

Pénalisation

- compenser la baisse naturelle de SSR par une pénalité qui défavorise les modèles de grande dimension
- on veut minimiser $SSR(d) + \text{pen}(d)$

$$R_{adj}^2 = R^2 - (1 - R^2) \times \frac{d - 1}{n - d}$$

Critère AIC

Critère d'Akaike est la vraisemblance du modèle pénalisé par le nombre de paramètres à estimer

$$\text{AIC} = 2k - 2 \ln(L)$$

Formulation pour les échantillons de petite taille.

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

Critère BIC

Critère d'information bayésien qui pénalise suivant le nombre de paramètres mais aussi en fonction de la taille de l'échantillon.

$$\text{BIC} = -2 \ln(L) + \ln(n)k$$

Plus sujet à discussion

- Fondements théoriques plus discutables que l'AIC (suppose que le « vrai modèle » est présent)
- Ne fonctionne pas trop en grande dimension
- Pas asymptotiquement optimal

Stratégies de sélection

En pratique, on ne peut pas parcourir l'ensemble des 2^p modèles possibles.

Sélection ascendante

- On part du modèle nul, puis ajout des variables explicatives 1 à 1.
- La variable ajoutée est celle qui diminue le plus le critère d'AIC
- Arrêt quand l'AIC ne diminue plus

Sélection descendante

- On part du modèle complexe, et on retire les variables 1 à 1
- Choix des variables selon la même stratégie que la solution ascendante.

Mélange des deux

- Sélection ascendante avec possibilité à chaque étape de supprimer une variable déjà ajoutée



Régression pénalisée

Régression pénalisée

Même but que la sélection de modèle, mais avec une approche très différente

Au lieu d'inférer des estimateurs par les données observées, puis de pénaliser le contraste du modèle par la dimension du modèle ...

... on infère des estimateurs par les données observées **en tenant compte de la pénalisation lors de la procédure d'optimisation !**

On laisse le modèle statistique **apprendre** lui-même la forme qu'il doit avoir.

Ridge

Optimisation

$$\hat{\beta} = \operatorname{argmin} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{k=1}^p \beta_k x_{i,k} \right)^2 + \lambda \sum_{k=1}^p \beta_k^2 \right\}$$

Utilisation

- Dans le cas de beaucoup de variables explicatives corrélées : forte variance de l'OLS
- Un coefficient positif peut venir contrebalancer un coefficient négatif
- La contrainte en norme L_2 permet limiter cette variance

Solution du ridge

Solution explicite

$$\hat{\beta}_\lambda^{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbb{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

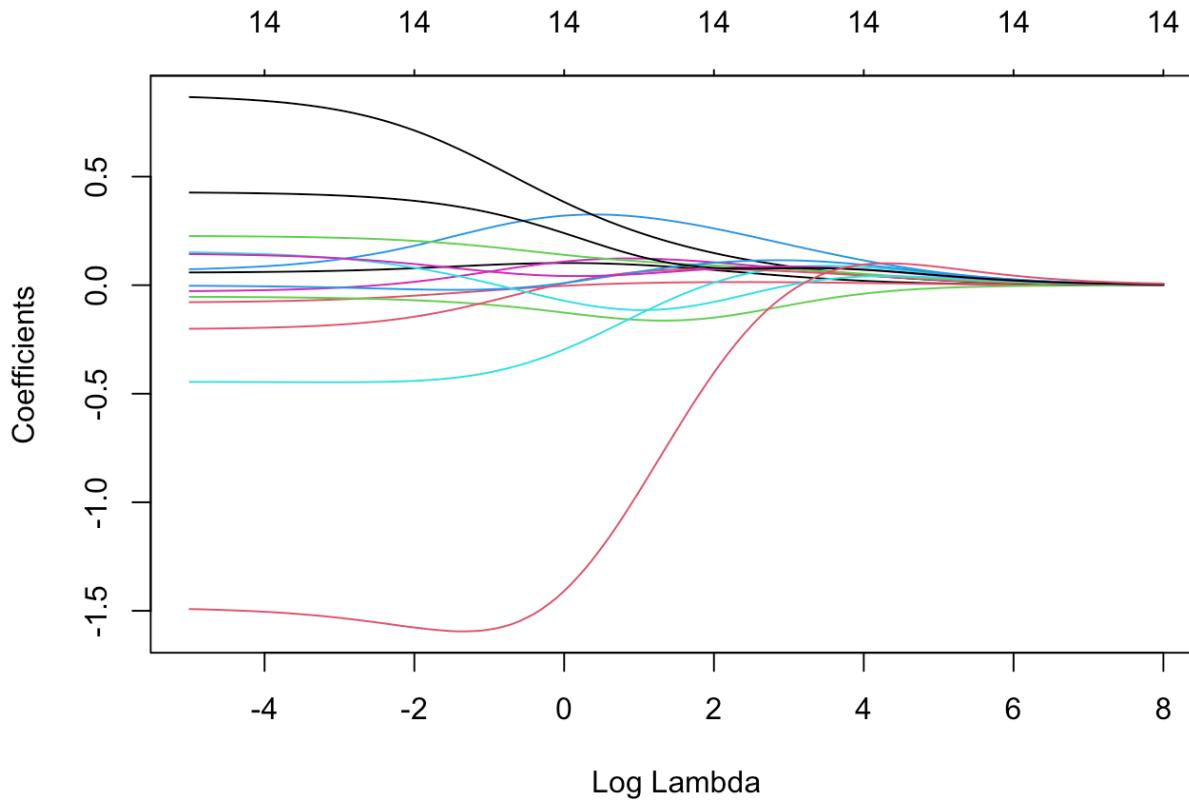
Impact du lambda

- Si $\lambda \rightarrow 0$, alors on retombe sur l'estimateur OLS
- Si $\lambda \rightarrow \infty$, alors l'estimateur ridge tend vers 0

Propriétés

- Variance diminuée par rapport à l'estimateur OLS
- Mais l'estimateur ridge est biaisé

Coefficients vs. Lambda



LASSO

Optimisation du Least Absolute Shrinkage and Selection Operator)

$$\widehat{\beta}_\lambda^{LASSO} = \operatorname{argmin} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{k=1}^p \beta_k x_{i,k} \right)^2 + \lambda \sum_{k=1}^p |\beta_k| \right\}$$

Utilisation

- Sélection de variables quand on a beaucoup de variables
- Problème de très grande dimension $p \gg N$: génétique
- La contrainte en norme L_1 force les coefficients à 0

Solution du LASSO

Contrairement au ridge, pas de solution explicite

- Utilisation d'un algorithme spécifique

Impact du lambda

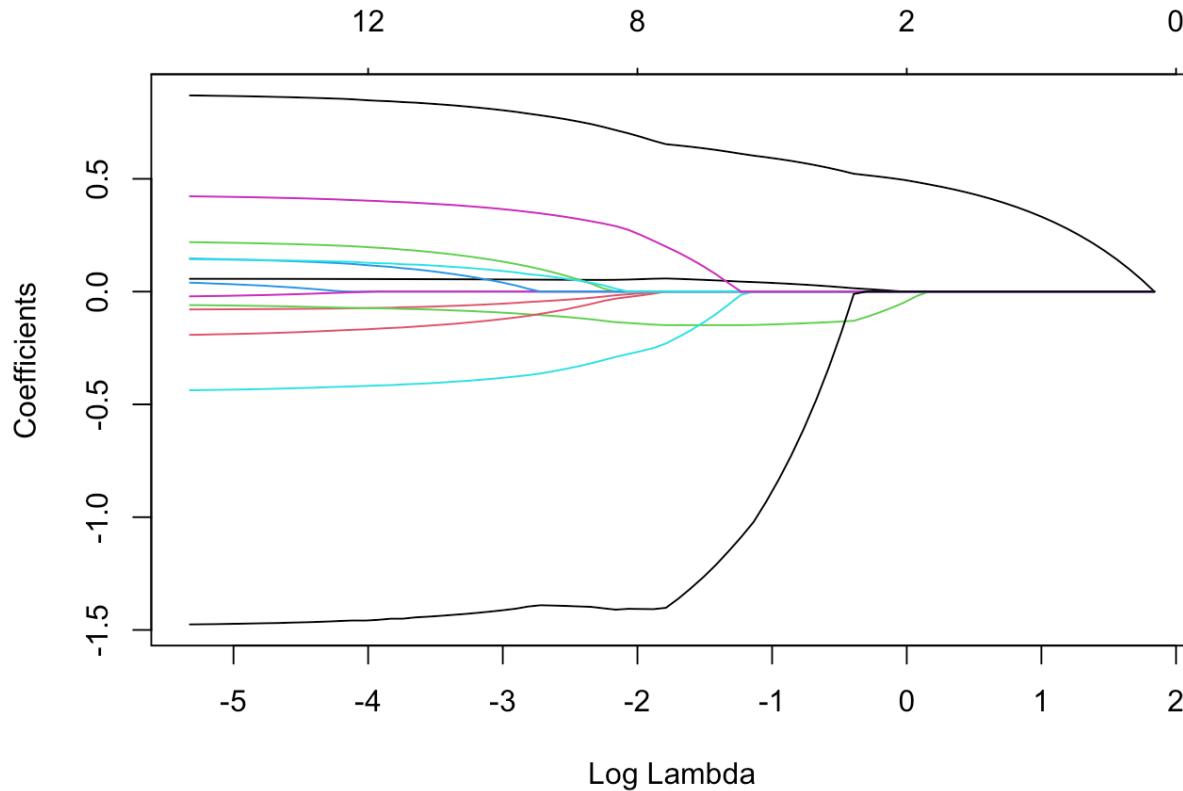
- Si $\lambda \rightarrow 0$, alors on retombe sur l'estimateur OLS
- Si $\lambda \rightarrow \infty$, alors l'estimateur LASSO tend vers 0



Propriétés de sélection

- Ne retient que certains coefficients, les autres sont mis à 0
- Mais ne peut pas retenir plus de n coefficients

Coefficients du LASSO vs. Lambda



Elastic-Net

Optimisation

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{k=1}^p \beta_k x_{i,k} \right)^2 + \lambda_1 \sum_{k=1}^p |\beta_k| + \lambda_2 \sum_{k=1}^p \beta_k^2 \right\}$$

Utilisation

- En cas de corrélation trop importante entre les variables, le LASSO a tendance à n'en choisir qu'une parmi le groupe
- En mixant les deux contraintes, on va garder (ou non) l'ensemble du groupe

Fused LASSO

Optimisation

$$\hat{\beta} = \operatorname{argmin} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{k=1}^p \beta_k x_{i,k} \right)^2 + \lambda_1 \sum_{k=1}^p |\beta_k| + \lambda_2 \sum_{k=2}^p |\beta_{k-1} - \beta_k| \right\}$$

Utilisation

- Comme le LASSO, une estimation parcimonieuse
- Forcer les coefficients qui se suivent à avoir les mêmes valeurs

Cross-validation

Comment choisir la force de pénalisation λ ?

Objectif : augmenter le pouvoir prédictif

- Base d'apprentissage : on estime les coefficients pour un λ donné
- Base de validation : on calcule les erreurs sur de l'*outsample* à partir du modèle
- On répète le processus pour plusieurs λ et on garde celui avec la meilleure mesure de validation

Cross-validation K -folds

- Séparation de la base de données en K sous-échantillons
- Estimation du modèle sur $K - 1$ sous-échantillons et validation sur le dernier sous-échantillon
- On répète le procédé K fois en changeant l'échantillon de validation
- Chaque point aura servi à l'apprentissage et à la validation



GLM Pénalisé

Generalized Linear Model (GLM)

Limites du modèle linéaire simple

- C'est bien quand la variable Y varie dans toutes les directions indéfiniment
- Si je veux estimer une probabilité de sinistre : $Y \in (0,1)$
- Si je veux estimer une charge sinistre : $Y \geq 0$

On suppose que la variable Y est générée par une distribution

- Introduction d'une **fonction lien g** qui définit la loi que suit Y

$$g(Y) = \beta_0 + \sum_{k=1}^p \beta_k X_k + \epsilon$$

Régression logistique

Régression d'une variable binaire $Y \in \{0,1\}$

- Classification binaire d'un événement de probabilité $P(Y = 1) = p_i$
- Marketing quantitatif : churn, transformation de l'assurance

Modèle logit

$$\ln \frac{p_i}{1 - p_i} = \beta_0 + \sum_{k=1}^p \beta_k X_k + \epsilon$$

Estimation par maximum de vraisemblance

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^N \left(\frac{e^{x_i^T \boldsymbol{\beta} + \boldsymbol{\beta}_0}}{1 + e^{x_i^T \boldsymbol{\beta} + \boldsymbol{\beta}_0}} \right)^{y_i} \left(\frac{1}{1 + e^{x_i^T \boldsymbol{\beta} + \boldsymbol{\beta}_0}} \right)^{1-y_i}$$

Extension de la pénalisation

Les régressions pénalisées s'étendent facilement aux GLM

$$\hat{\beta}_\lambda^{LASSO} = \operatorname{argmin} \left\{ -\log(\mathcal{L}(\boldsymbol{\beta})) + \lambda \sum_{k=1}^p |\beta_k| \right\}$$

Et toutes les autres pénalisations...



Application à la mortalité

Taux de mortalité

$$m_{x,t} = \frac{D_{x,t}}{E_{x,t}}$$

Dimension en âge x :

- classe 0-100 ans pour toute la courbe,
- ou 45-95 ans pour les contrats d'assurance vie / retraite

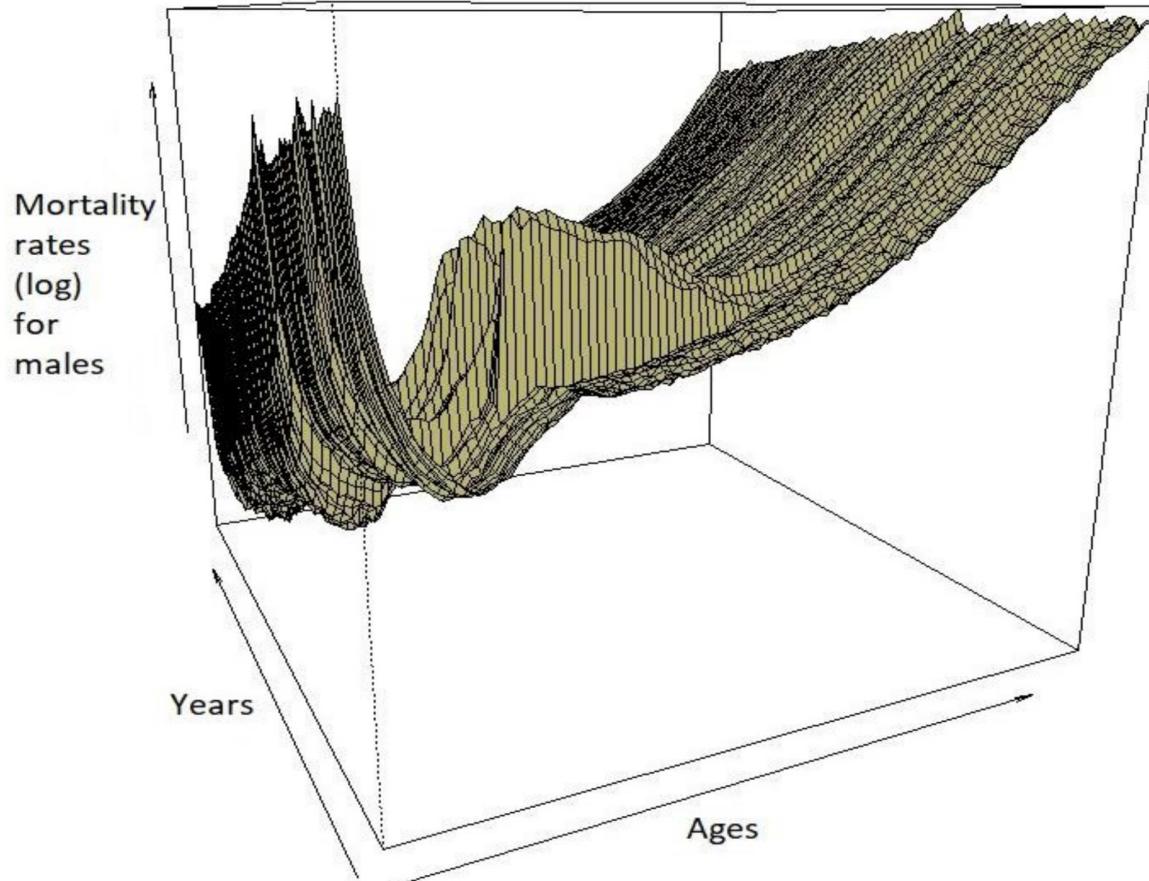
Dimension temporelle t

- Estimation depuis 1950 – 1980 ;
- Projection sur 50 ans pour les calculs actuariels.

Données :

- Human Mortality Database (www.mortality.org)
- Âge et période par année.

Surface de mortalité



Modèles de mortalité à facteurs

Lee-Carter (1992) et Renshaw-Haberman (2006)

$$\ln(m_{x,t}) = \alpha_x + \beta_x^{(1)} \kappa_t + \beta_x^{(2)} \gamma_{t-x}$$

CBD (Cairns et al., 2006) et M7 (Cairns et al., 2009)

$$\ln\left(\frac{q_{x,t}}{1-q_{x,t}}\right) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \kappa_t^{(3)}((x - \bar{x})^2 - \hat{\sigma}^2) + \gamma_{t-x}$$

Vecteur Autorégressif (VAR)

Modélisation des taux d'amélioration de la mortalité en tant que séries temporelles

$$\Delta y_{x,t} = y_{x,t} - y_{x,t-1} = \ln(m_{x,t}) - \ln(m_{x,t-1})$$

Projection de la courbe par une approche vectorielle

$$\Delta Y_t = (\Delta y_{x_{\min},t}, \Delta y_{x_{\min}+1,t}, \dots, \Delta y_{x_{\max}-1,t}, \Delta y_{x_{\max},t})^T$$

Liberté significative dans la structure de dépendance spatio-temporelle grâce au VAR

$$\Delta Y_t = C + \sum_{k=1}^p A_k \Delta Y_{t-k} + E_t$$

Grande dimension

L'estimation du modèle VAR soulève la problématique de grande dimension

- Le nombre d de séries temporelles peut être important
- Les p matrices autorégressives sont de dimension d^2

Particulièrement vrai dans l'analyse de la mortalité

- Classe d'âge 0-100 ans
- Estimation sur 70 ans d'historique
- Seulement 7,070 points d'observation
- VAR(1) implique 10,302 coefficients
- $N \ll p$

Besoin d'une méthodologie d'estimation spécifique pour éviter le sur-apprentissage

VAR - ENET

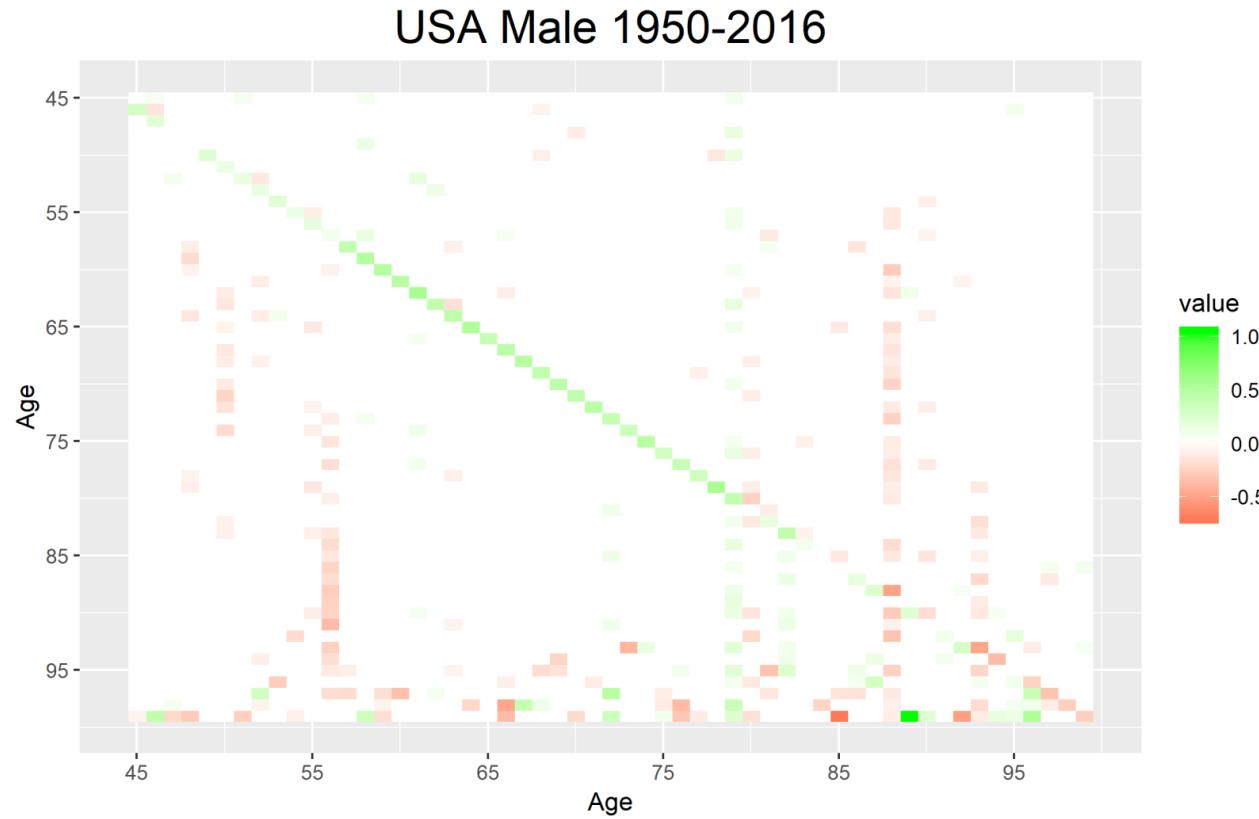
Estimation pénalisée des coefficients autorégressifs

$$\frac{1}{T-p} \sum_{t=p}^T \| \Delta Y_t - C - \sum_{k=1}^p A_k \Delta Y_{t-k} \|_2^2 + \alpha \lambda \sum_{k=1}^p \| A_k \|_1 + \frac{(1-\alpha)\lambda}{2} \sum_{k=1}^p \| A_k \|_2^2$$

Elastic-Net est une double pénalisation

- LASSO : sélection de variable pour des matrices A_k parcimonieuses
- Ridge : effet de groupement car taux d'amélioration de mortalité fortement corrélés entre eux
- Application récente aux modèles VAR, mais surtout en économie et finance.

Matrice auto-régressive parcimonieuse



Projection à court-moyen terme

L'adaptabilité du modèle grâce à l'Elastic-net se retrouve dans les erreurs de projections

Mesure de validation : MSE

Procédure

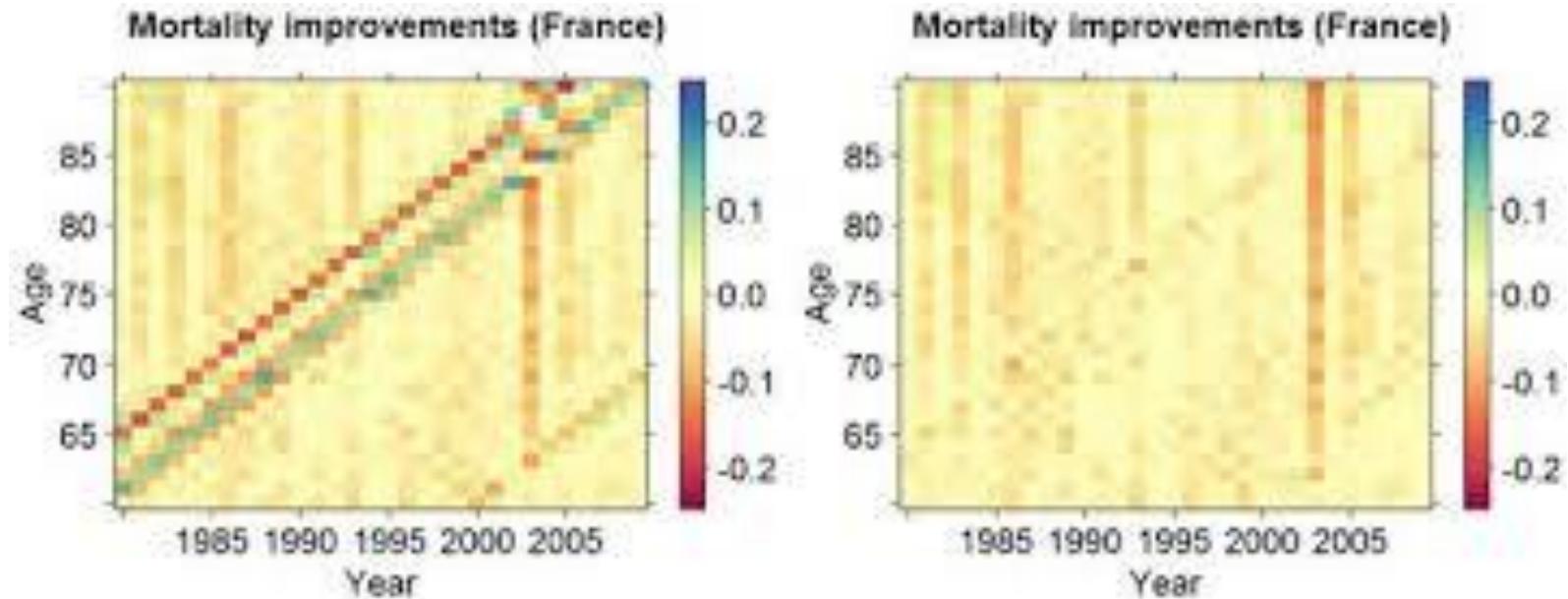
- Classe d'âges 45-99 ans
- Estimation des modèles sur 1950-2000
- Projection sur 2001-2016

	VAR	LC	M7	HU	RESPECT	STAR
FR Female	0,088	0,111	0,676	0,082	0,083	0,098
FR Male	0,110	0,113	0,193	0,112	0,091	0,127
FR Total	0,078	0,067	0,257	0,067	0,071	0,417
UK Female	0,095	0,142	0,228	0,109	0,281	0,083
UK Male	0,087	0,141	0,099	0,138	0,230	0,115
UK Total	0,080	0,138	0,145	0,122	0,296	0,156
US Female	0,078	0,085	0,237	0,061	0,110	0,071
US Male	0,116	0,122	0,144	0,141	0,081	0,050
US Total	0,078	0,087	0,135	0,085	0,075	0,049
Mean	0,090	0,112	0,235	0,102	0,146	0,130
St. Dev.	0,014	0,027	0,174	0,029	0,094	0,113

Détection de l'effet cohorte

Contrairement à d'autres modèles, pas d'effet cohort imposé *a priori*

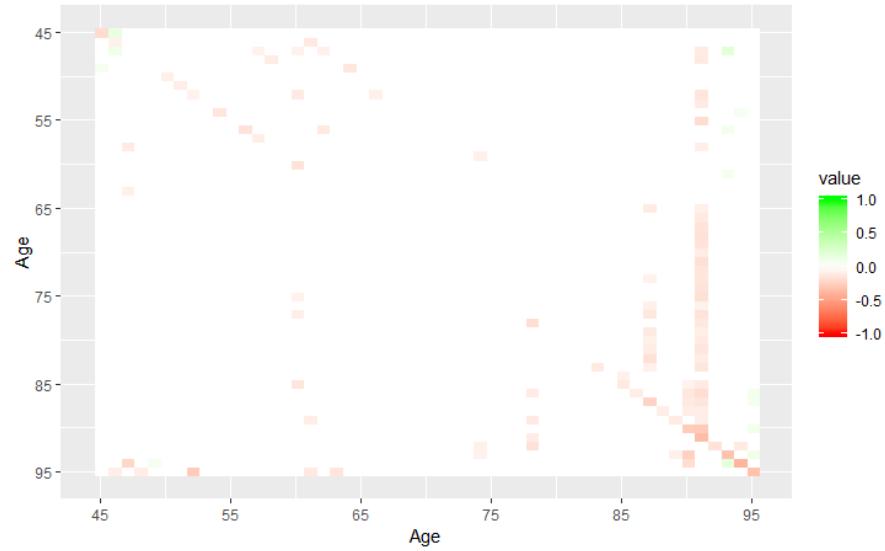
- Cairns et al. (2016) et Boumezoued (2016) notent des faux effets cohorte dans les données.
- Correction de la HMD, notamment pour la France (ici téléchargement en Octobre 2017 et Janvier 2019).
- Estimation d'un VAR(1) sur la population masculine 1950-2012.



Détection de l'effet cohorte

Contrairement à d'autres modèles, pas d'effet cohort imposé *a priori*

- Cairns et al. (2016) et Boumezoued (2016) notent des faux effets cohorte dans les données.
- Correction de la HMD, notamment pour la France (ici téléchargement en Octobre 2017 et Janvier 2019).
- Estimation d'un VAR(1) sur la population masculine 1950-2012.



Détection de lissage - Wyoming

