Mémoire présenté devant le Conservatoire National des Arts et Métiers

pour l'obtention du diplôme du Master Droit Economie Gestion mention Actuariat

et l'admission à l'Institut des Actuaires

le 14 Décembre 2021

Par : Marc-Antoine DEFRANSURE

Titre: Analyse des déterminants de l'intensité d'entrée en dépendance à l'aide de méthodes de régression pénalisée

Confidentialité : ☑ NON ☐ OUI (Durée : ☐ 1 an ☐ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Signatures :

Présidente du jury

Sandrine LEMERY

Membres présents du jury de l'Institut des Actuaires

Michel NANG

Jonathan BASTIEN

Membres présents du jury du Conservatoire National des Arts et Métiers

Olivier DESMETTRE

David FAURE

François WEISS

Entreprise :

Nom : Prim'Act

Directeur de mémoire en entreprise :

Nom : Frédéric PLANCHET

Signature :

Secrétariat

Signature du candidat

Bibliothèque :

« On commence à vieillir quand on finit d'apprendre.»

Proverbe japonais

Remerciements

Merci au docteur **Michael Schwarzinger** pour m'avoir offert la très belle opportunité de travailler sur les données du PMSI.

Merci à mon tuteur **Frédéric Planchet** pour m'avoir permis de traiter ce sujet aussi passionnant. Merci sincèrement pour la confiance que tu m'as accordé, pour ta disponibilité ainsi que tes conseils avisés.

Merci également **Quentin Guibert** et **David Faure** pour les nombreuses pistes proposées ainsi que pour les longues relectures qui s'en sont suivies.

Merci au **cabinet Prim'Act** et à tous ses associés pour m'avoir ouvert les portes du monde du conseil et de l'actuariat, dans un environnement de camaraderie et toujours porté vers l'excellence.

Merci à **mes parents** et à **mon frère** pour leur soutien sans faille et ceci bien avant que le mot actuariat n'ait de sens pour moi !

Merci enfin à **Michel Fromenteau** pour m'avoir ouvert les portes du CNAM et m'avoir transmis, ainsi qu'à une génération entière d'étudiants, la passion pour cette belle discipline qu'est l'actuariat.

Table des matières

Introduction Générale

Le 28 Mars 2019, le rapport Libault [1] sur la Concertation Grand âge et autonomie a été remis au Premier Ministre. Ce rapport annonce le début d'un grand chantier pour le gouvernement pour répondre à la question cruciale du financement de la dépendance. Des mesures chocs comme le report de l'âge de départ à la retraite à 63 ans ou l'allongement de la durée de cotisation sont sur la table. En effet, l'affaire n'est pas mince, il s'agit pour l'Etat de trouver 9,2 milliards d'euros d'ici l'horizon 2030 pour financer la prise en charge de la dépendance. Ce grand défi a pour but de faire face au choc démographique du vieillissement de la population. En effet, 1,5 million de français ont plus de 85 ans aujourd'hui, on en comptera plus de 5 million en 2050 [2].

Si la dépendance est l'un des chantiers majeurs du gouvernement pour les prochaines décennies, un autre sujet brûlant est sur sa table, celui de l'intelligence artificielle. En effet, un an jour pour jour avant le rapport Libault (le 28 mars 2018), le célèbre mathématicien et député Cédric Villani, a remis au gouvernement français son rapport sur l'intelligence artificielle. Parmi les sujets proposés par l'ancien médaillé Fields, pour que la France se positionne à l'avant-garde de l'IA, figure l'identification de quatre secteurs prioritaires où la France doit particulièrement concentrer son effort de développement : la santé, les transports, l'environnement et la défense. Ces recommandations mettent en lumière l'importance grandissante apportée à deux domaines connexes de l'actuaire, le domaine des statistiques à travers le concept d'intelligence artificielle et celui de la santé. Dans un contexte où le risque dépendance et l'intelligence artificielle sont sur le devant de la scène, l'actuaire dispose des outils et connaissances pour tenter de répondre à ces deux défis majeurs qui s'annoncent. A la croisée de ces deux thématiques, l'ambition de ce mémoire est d'utiliser une volumineuse base de données médicales pour tenter d'expliquer le phénomène d'entrée en dépendance par le biais de méthodes statistiques d'apprentissage.

Dans un univers où les grands volumes de données et le phénomène du Big Data n'existaient pas, la démarche statistique « classique » consistait à construire un estimateur sur un jeu de données unique et utiliser une théorie asymptotique pour permettre de juger de sa qualité et de construire des intervalles de confiance. Dans la logique de la théorie de l'apprentissage, l'abondance de données permet de séparer celles-ci en deux sous-échantillons, l'un dit d'apprentissage et l'autre de validation ; la qualité des estimateurs n'est alors plus jugée par l'intermédiaire de critères asymptotiques, mais en fonction de l'adéquation à l'échantillon de validation par l'intermédiaire d'une mesure de l'erreur de prédiction.

Une première partie s'attachera à expliquer le phénomène d'entrée en dépendance au moyen d'études médicales reconnues de manière à connaitre a priori les principaux facteurs et pathologies influençant l'entrée dans l'état de dépendance. En effet, les variables disponibles pour cette étude sont issues du Programme de Médicalisation des Systèmes d'Information (PMSI), celles-ci correspondent à des pathologies médicales dont ni l'apparition, ni l'évolution, ni les conséquences ni même les appellations ne sont connues ou familières de l'actuaire. La connaissance des causes médicales connues a priori est en effet indispensable pour valider les modèles obtenus a posteriori.

Par la suite, après avoir présenté de façon théorique le modèle de Cox, une étude statistique va être menée en opérant la phase de sélection de variables avec des méthodes "classiques" pas à pas basées sur le critère AIC. Le modèle de Cox sera ensuite paramétré pour analyser les déterminants de l'intensité d'entrée dans l'état de dépendance (et démence) sur les populations disponibles pour l'étude. Le caractère chronophage de cette première méthode de sélection sera en particulier soulevé.

1. Dominique Libault est le président du Haut Conseil du financement de la protection sociale
2. Source, Le Parisien et solidarites-sante.gouv.fr

Dans une troisième partie, un jeu de données sera simulé de manière à tester dans un cadre analogue de données censurées différentes méthodes de régressions pénalisées (LASSO, Ridge, Enet, Adaptative LASSO). Le but de cette partie est d'étudier la qualité de sélection et la justesse de prédiction des estimateurs obtenus grâce à ces méthodes par rapport à la procédure classique exposée dans la partie précédente. Après avoir défini de façon théorique la régression pénalisée, la théorie de l'apprentissage et les différentes mesures de prédiction qui en découlent, les concepts seront mis en pratique sur ces bases de données simulées. L'influence de la taille des bases, des corrélations entre variables, les techniques d'échantillonnage utilisées, le choix des mesures de prédiction et le taux de censures seront notamment mesurés. Une fois définies et appréhendées, ces techniques seront alors appliquées sur les bases de données médicales du PMSI. Les résultats seront comparés avec ceux obtenus dans la partie utilisant les méthodes classiques et la qualité de sélection sera confrontée aux attendus théoriques préalablement établis dans la première partie du mémoire à l'aide des études reconnues par la profession médicale.

Résumé

Définition de la dépendance et des données disponibles

La dépendance est définie et mesurée différemment selon que l'on se place dans le périmètre assurantiel ou médical. Dans un contexte assurantiel, l'état de dépendance est défini comme la quantité de ressources extérieures nécessaires à une personne pour effectuer les actes basiques de la vie quotidienne ce qui permet aux pouvoirs publics et aux organismes d'assurance français de distinguer in fine deux niveaux de risque dépendance pour jalonner leurs aides ou couvertures : la dépendance partielle et la dépendance totale. Ce travail s'adresse à un public du secteur de l'assurance mais les données disponibles provenant du milieu médical, une première phase de définition et de comparaison [3] a permis de réconcilier ces deux définitions. Les notions de dépendance physique et démence étudiées dans ce mémoire correspondent à un niveau de dépendance totale.

Les variables disponibles pour cette étude sont issues du Programme de Médicalisation des Systèmes d'Information (PMSI), celles-ci correspondent à des pathologies médicales dont ni l'apparition, ni l'évolution, ni les conséquences ni même les dénominations ne sont connues ou familières de l'actuaire. La connaissance des causes médicales connues a priori est en effet indispensable pour valider les modèles obtenus a posteriori. A la lumière d'articles scientifiques, il a été établi dans cette phase exploratoire que les causes médicales identifiées par les spécialistes du domaine médicale pour l'apparition de la dépendance cognitive ou démence sont :

- Parmi les facteurs de risques comportementaux :
  - ☐ L'alcool
- Parmi les pathologies neurologiques :
  - ☐ L'épilepsie ;
  - ☐ La maladie de Parkinson ;
    (La maladie se déclarant aux alentours de 70 ans, et la démence Parkinsonienne 10 à 15 ans après, celle-ci touche les personnes avec des âges très avancés.)
  - ☐ La sclérose en plaque
  - ☐ L'hydrocéphalie à pression normale
  - ☐ L'encéphalite
- Les accidents vasculaires cérébraux
- Les pathologies cardiaques :
  - ☐ L'hypertension artérielle
- Le diabète

L'objet de cette liste de causes médicales de démence, non exhaustive, est de servir d'étalon pour arbitrer quant à la pertinence de la sélection de variable opérée par les modèles paramétrés par le biais des méthodes d'apprentissage.

3. Issu de SCHWARZINGER [2018]

Application d'un modèle multiplicatif à hasard proportionnel.

Dans ces travaux, les sorties sont modélisées au moyen d'un modèle multiplicatif à hasard proportionnel de Cox. Ce modèle, qui sera décrit plus en détail par la suite, a pour objectif de mesurer les écarts entre différentes sous-populations, sous réserve de respecter certaines hypothèses dont l'indépendance des covariables avec le temps et l'hypothèse de proportionnalité des risques. L'avantage de ce modèle est qu'il respecte le critère informatif de la censure. Cependant, dans un souci de simplification, la censure aléatoire induite par le décès a été traitée comme une censure simple ce qui induit un biais de modélisation et affecte l'estimation des écarts relatifs entre les différentes sous-populations étudiées. L'étude et l'introduction de concurrences entre les différentes causes de sorties du périmètre d'étude permet de lever ce biais, on parle de modèles à risques concurrents. Ce mémoire, dont l'accent est mis sur la comparaison avec des méthodes d'apprentissage, se limite à l'utilisation de modèles non concurrents.

Afin de vérifier les hypothèses d'application d'un modèle à hasard proportionnel et indépendant du temps, la comparaison des taux bruts avec les modélisations de Cox a été opérée sur 4 sous-populations suffisamment hétérogènes en termes de risque de dépendance (pour capter des écarts) et suffisamment volumineuses (pour limiter les problèmes dus à une trop grande volatilité). Cette comparaison pour les (picards, parisiennes) x (alcoolique, non alcoolique) pour les deux causes de sorties ont permis de valider la pertinence de l'utilisation de ce type de modèle et valider les conditions nécessaires d'indépendance temporelle et de risques proportionnels.

D'autres méthodes, comme l'analyse des résidus de Schönefeld ou des tests d'adéquation du χ^2, ont été menées mais ne permettaient pas de valider ou invalider clairement les hypothèses de Cox.

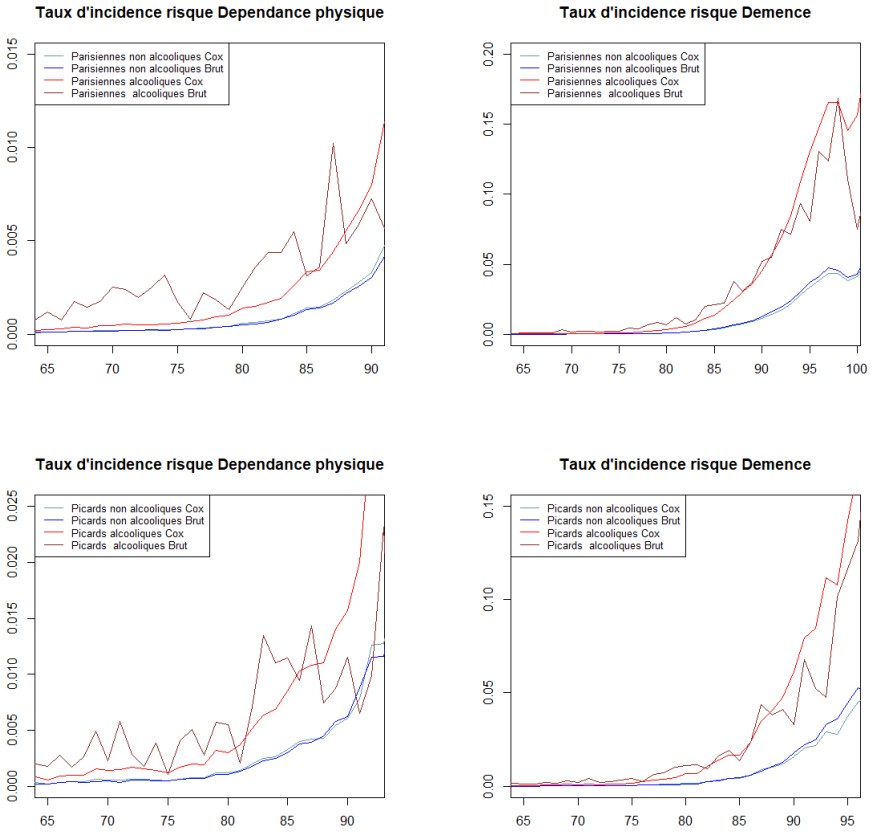


FIGURE 1 – Taux bruts et taux prédits par Cox pour différentes sous-populations

On remarque que la courbe des prédictions approxime bien la courbe des observations brutes. De plus, les deux sous-populations ont des tendances similaires ce qui permet d'accepter l'hypothèse de proportionnalité des risques. Le modèle de Cox étant bien adapté, une méthode de sélection de variable à l'aide du critère AIC et de méthode pas à pas a permis d'opérer une première sélection de variables.
Les pathologies préalablement identifiées comme à risque ressortent relativement bien.

	Parisienne Demence	Picard Demence	Parisienne Dependance	Picard Dependance		Parisienne Demence	Picard Demence	Parisienne Dependance	Picard Dependance
Covariable	Stepwise	Stepwise	Stepwise	Stepwise	Covariable	Stepwise	Stepwise	Stepwise	Stepwise
fdr_aud_all	**4,14**	**4,13**	3,77	4,02	DIG0_VB_METABO	0,87	0,97	0,89	0,98
fdr_smoker	2,94	2,50	3,27	2,60	CARDIO0_IHD_3noTTT	0,96	0,97	0,97	0,96
CANCER1_SMOKER	2,13	1,75	2,15	1,77	RESP0_LRI_ATCD	0,98	0,96	0,98	0,97
CANCER1_PC_POOR	2,03	1,60	2,09	1,63	RHEUM0_AUD	0,94	0,96	0,93	0,95
fdr_obesity_all	1,77	1,26	2,01	1,36	CARDIO1_INSUF_CHRO	0,96	0,96	0,95	1,00
CANCER1_AUD_SMOKER	1,66	1,41	1,68	1,42	TRAUMA0_CHUTE	0,88	0,96	0,93	0,92
CANCER1_PC_GOOD	1,60	1,22	1,63	1,24	CARDIO1_IHD_1MI	0,84	0,96	0,88	1,01
CANCER1_BREAST	1,44	-	1,51	-	TRAUMA0_3FRACTURE	0,94	0,95	0,93	0,93
NEURO1_DEM_FDR	**1,36**	**1,42**	1,34	1,32	RHEUM0_ARTHROSE_OTHE	0,91	0,95	0,91	0,92
NEURO1_PARKINSON	**1,29**	**1,11**	1,22	-	CARDIO0_HBP	0,92	0,95	0,92	0,92
DIG1_LIVER_1CirrD	1,24	1,21	1,25	1,22	DIG0_OTHER_noFDR	0,93	0,95	0,93	0,94
CANCER1_COLORECTAL	1,24	1,23	1,30	1,23	CARDIO0_RYTHME_3noTT	0,93	0,94	0,93	0,92
CANCER1_HEMATO	1,24	1,24	1,31	1,31	RESP0_APNEE_SOM	0,89	0,93	0,92	0,91
STROKE1_1HEMO	**1,22**	-	1,22	-	CARDIO0_RYTHME_2TTT	0,91	0,93	0,90	0,94
NEURO1_EPILEPSIE	**1,21**	**1,21**	1,13	1,21	SENSE0_CATARACTE	0,90	0,93	0,90	0,93
INFECT1_SEPSIS	1,20	1,16	1,19	1,18	KIDNEY1_1INSUF_CHRO	0,96	0,92	0,95	0,93
DIG1_STOMIE	1,16	1,32	1,21	1,30	DIG1_HEMORRAGIE	0,97	0,93	0,95	0,91
BLOOD0_2OTHER	1,13	1,05	1,15	-	RHEUM0_ARTHROSE_HIP	0,92	0,91	0,92	0,90
cp_dipl0	1,13	1,05	1,16	1,06	RHEUM0_ARTHROSE_KNEE	0,87	0,90	0,87	0,89
TRAUMA0_SUICIDE	1,10	-	1,08	-	TRAUMA1_1CRANE	0,92	0,89	0,94	0,85
DIG0_LIVER_ETIO_ANY	1,09	1,05	1,10	1,05	DIG1_PERITONITE	0,99	0,89	0,97	0,89
RESP1_2INSUF_AIGUE	1,09	-	1,11	-	CANCER0_SKIN	0,92	0,88	0,90	0,89
BLOOD0_2ANEMIA_LYSE	1,08	1,21	1,10	1,22	DIG0_PANCREAS_AUD	0,91	0,85	0,89	0,84
BLOOD1_1TRANSFUSION	1,08	1,10	1,11	1,10	NEURO1_OTHER	1,11	0,72	1,14	0,68
RHEUM0_SYSTEME	1,08	-	1,12	-	RESP0_2INTERSTI	1,29	1,04	1,26	1,08
ENDOC0_DIABETE	1,08	1,03	1,07	1,04	DIG0_LIVER_2CirrC	0,98	1,02	0,99	1,05
ENDOC1_METABO	1,06	-	1,06	-	CV1_MTE	1,08	1,01	1,07	1,04
ENDOC1_GLD_OTHER	1,06	1,16	1,06	1,12	KIDNEY0_2GN	0,99	1,01	1,03	1,04
RESP0_2OTHER	1,06	-	1,09	-	CARDIO0_IHD_2TTT	0,97	1,01	0,97	1,03
CARDIO0_VALVE_ANY	1,05	1,03	1,07	1,04	cp_immi	0,99	1,00	1,00	0,99
NEURO0_SNP_METABO	1,05	1,10	1,06	1,08	BLOOD0_2ANEMIA_IRON	0,95	0,99	0,94	0,98
DIG1_OCCLUSION	1,05	-	1,03	-	KIDNEY0_CYSTITE_ATCD	0,91	0,99	0,89	0,98
RESP1_1INSUF_CHRO	1,05	0,92	1,08	0,94	TRAUMA0_OSTEOPOROSE	1,08	1,01	1,11	0,98
KIDNEY1_2INSUF_AIGUE	1,02	1,12	1,03	1,11	CANCER0_ATCD_FAM	0,94	0,99	0,95	0,99
cp_dep	1,01	-	1,01	-	KIDNEY0_2OTHER	0,92	0,99	0,93	1,00
ENDOC0_DYSLIPIDEMIA	0,99	0,97	0,97	0,96	KIDNEY0_2PYELONEPHRI	0,89	1,01	0,86	0,99
ENDOC0_THYROIDE	0,99	-	0,97	0,94	CARDIO0_OTHER	1,00	1,01	1,01	0,99
RESP0_2COPD	0,98	-	-	-	KIDNEY0_2UROLITHIASE	0,95	0,97	0,97	0,99
STROKE0_AIT_noSEQ	0,98	-	0,97	-	RESP0_2ASTHMA	1,01	1,00	0,99	1,01
CANCER0_BENIN	0,97	0,90	0,98	0,90	RESP0_BRONCHITE_ATCD	1,06	1,00	1,02	1,02
CARDIO1_RYTHME_1ACFA	0,97	0,87	-	0,87	STROKE0_2TTT	1,06	1,04	1,03	1,03
CANCER0_INSITU	0,97	-	0,98	-	CANCER1_PROSTATE	1,04	0,00	1,03	-
ENDOC0_CARENCE	0,97	0,89	0,91	0,83	CV1_PVD	1,01	0,98	1,01	1,01
STROKE0_3noTTT	0,97	-	0,96	-	**STROKE1_1ISCHEMIC**	0,95	**1,02**	**0,97**	1,00
					TRAUMA1_2SEVERE	0,94	1,03	0,97	1,03

TABLE 1 – Selection AIC. En gras les covariables pré-identifiées

Cependant, plusieurs problèmes sont mis en lumière. Tout d'abord la procédure pas à pas est extrêmement chronophage. Celle-ci a en effet mis plus de 24h à être exécutée malgré la grande puissance de calcul disponible. Par ailleurs, l'étude de corrélation normalement indispensable en amont du processus de sélection n'a pu être mené au regard de la taille excessive de la base (89 variables) et des inter-corrélations très importantes entre toutes les pathologies.

Utilisation de techniques de régression pénalisée

L'idée de la régression pénalisée est d'introduire l'étape de pénalisation directement dans le problème d'optimisation de la vraisemblance en ajoutant une contrainte sur la norme des coefficients de régression de la forme :

$$\sum_{j=1}^{p} |\beta_j|^\delta \leq C$$

La région des valeurs possibles pour les coefficients de régression dépendent alors de la valeur du paramètre δ. Par exemple, en dimension 2, les régions autorisées des coefficients β_1 et β_2 en fonction du paramètre δ ont les

formes suivantes :

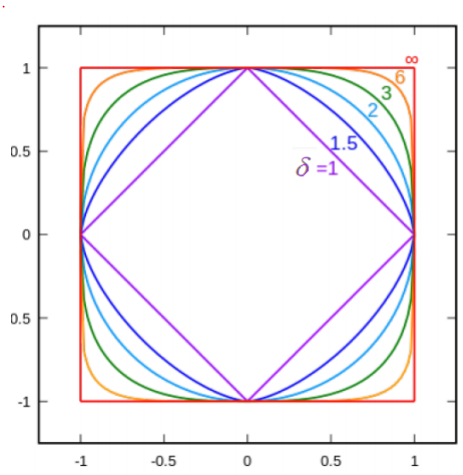


FIGURE 2 – Régions des contraintes en fonction de δ

Les travaux de VERWEIJ [1996] permettent de transposer la méthode de pénalisation couramment rencontrée pour la méthode d'optimisation des moindres carrés ordinaires au calcul d'optimisation de la vraisemblance. Le fait d'imposer une contrainte sur la norme des coefficients revient alors à exprimer directement la vraisemblance partielle pénalisée de la façon suivante :

$$L_{partielle}(\beta) = \prod_{i}^{D} \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - p(\beta, \lambda)$$

Où $p(\beta, \lambda)$ est la fonction de pénalisation. λ est le paramètre d'intensité ou paramètre de régularisation. Ce paramètre modère l'intensité de la pénalisation de façon linéaire.

En faisant varier le valeur de la constante de régularisation, on obtient alors l'ensemble de solutions suivant :

$$\{\hat{\beta}(\lambda) \in [0, +\infty[\text{ où } \hat{\beta}(\lambda) = argmin_\beta \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - p(\beta, \lambda)\}$$

Un certain nombre de simulations a été mené de manière à tester les différentes pénalisations afin de sélectionner celle qui serait la plus adéquate à appliquer sur le jeu de données réel disponible pour cette étude.

Pour l'ensemble des simulations, des jeux de données censurées, dont la variable d'intérêt vérifie la relation de Cox, ont été simulés :

$$\alpha(t|\mathbf{Z}_i(t)) = \alpha_0(t) exp^{\mathbf{Z}_i^T \beta}$$

où :

- La fonction de hasard de base a été prise égale à $\alpha_0(t) = 2.t$ afin d'avoir une expression simple et dépendante du temps (le facteur 2 a été choisi pour qu'il se simplifie lors de l'intégration pour le calcul de la loi de survie). Ce choix revient à considérer qu'en l'absence de l'effet des autres covariables, le taux d'incidence croit linéairement au cours du temps ;
- Pour le vecteur des covariables, il a été pris pour chaque individu i une valeur aléatoire suivant une loi uniforme sur l'intervalle [-1 ;1] : $\forall\ i \in [\![1; n]\!]\ \mathbf{Z}_i \sim \mathcal{U}[-1; 1]$
  Pour certaines simulations, des corrélations plus ou moins fortes entre certaines variables ont été appliquées afin d'en mesurer les effets ;
- Le fait de fixer certains coefficients à zero a enfin permis de mesurer la qualité de sélection des estimateurs puisque cela signifie d'imposer que les variables associées n'ont aucune influence sur la variable de sortie. Un bon estimateur devant les exclure.

La loi de survie en fonction du temps pour un individu i étant :

$$S(t, \mathbf{Z}_i) = exp[-\int_0^t \alpha(s, \mathbf{Z}_i)ds]$$

L'intensité de la loi de survie ayant été choisie de la forme $\alpha_0(t) = 2.t$, on obtient alors :

$$S(t, \mathbf{Z}_i) = exp[-\int_0^t 2s.exp^{\mathbf{Z}_i^T \beta}ds]$$

$$S(t, \mathbf{Z}_i) = exp(-t^2.e^{\mathbf{Z}_i^T \beta})$$

Pour simuler les temps T_i de survenance pour chaque individu i il a été utilisé une méthode de Monte-Carlo en inversant la fonction de survie conditionnelle et en simulant des variables aléatoires uniformes $Y_i \in \mathcal{U}[0; 1]$

On obtient alors pour chaque individu i :

$$T_i = \sqrt{-exp^{-\mathbf{Z}_i^T \beta} log(Y_i)}$$

La méthode de Monte Carlo consiste en effet à simuler la variable aléatoire Y_i (très simple avec le logiciel R) pour déduire les temps T_i distribués selon une loi de Cox.

En fixant à l'avance les paramètres intrinsèques de la distribution (lesβ), il a alors été possible d'appliquer les différentes techniques de régression pénalisée LASSO, Ridge, Elastic-net et adaptative Lasso pour les comparer à l'attendu et ainsi déduire quelle pénalité était la plus adaptée au regard de jeux de données incluant différents niveaux de corrélations, de taux de censure ou encore pour mesurer l'influence de la taille des échantillons.

Il ressort de ces simulations que la régression Lasso permet d'effectuer de la sélection de variable puisqu'elle autorise les coefficients à être nuls. La régression Ridge permet quant à elle de gérer les problèmes liés à la colinéarité. La régression Elastic-net enfin permet de combiner les effets des deux pénalisations en effectuant une sélection et en gérant les effets non désirés de la colinéarité. Si ces conclusions sont relativement classiques dans le cadre de la pénalisation d'une régression par méthode des moindres carrés, celles-ci se retrouvent bien en ce qui concerne la pénalisation de la vraisemblance partielle de Cox.

La pénalisation retenue au regard de la grande dimension de la base d'étude, de sa forte colinéarité et de son caractère censuré a ainsi été la régression Elastic-net.

Le problème d'optimisation de la vraisemblance partielle de Cox pénalisée par une pénalité Elastic-Net est ainsi le suivant :

$$L_{partielle}(\beta) = \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - \lambda(\frac{1}{2}\sum_{j=1}^p \beta_j^2 + \frac{1}{2}\sum_{j=1}^p |\beta_j|)$$

En faisant varier le valeur de la constante de régularisation, on obtient alors l'ensemble de solutions suivant :

$$\{\hat{\beta}(\lambda) \in [0, +\infty[\text{ où } \hat{\beta}(\lambda) = argmin_\beta \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - \lambda(\frac{1}{2}\sum_{j=1}^p \beta_j^2 + \frac{1}{2}\sum_{j=1}^p |\beta_j|)\}$$

Le chemin de régularisation ci-dessous est la représentation de l'ensemble des valeurs prises par chacun des coefficients de régression en fonction de la valeur du paramètre de régularisation λ.

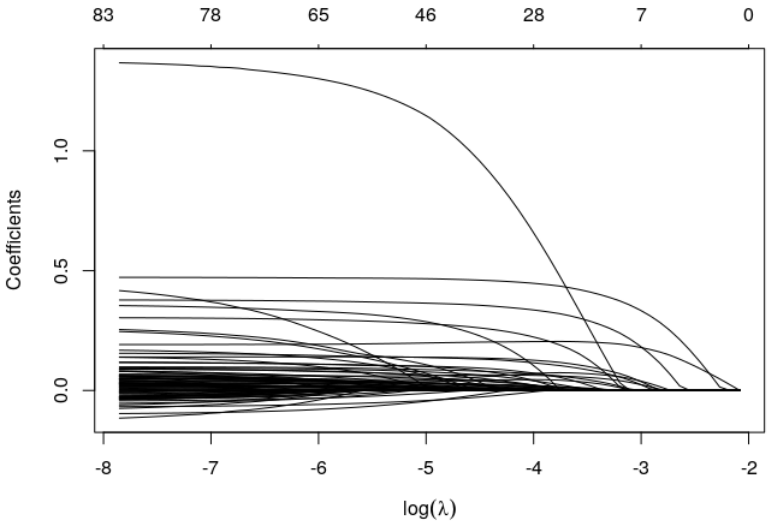


FIGURE 3 – démence des parisiennes

Ce travail a été effectué sur les populations parisiennes, picardes pour les risques de dépendance physique et démence.

Le paramètre de régularisation optimal qui minimise la vraisemblance est une combinaison entre le paramètre α qui donne plus ou moins de poids aux deux composantes de l'elastic net relatives au norme 1 et au norme 2 et le paramètre β qui module l'intensité de la pénalisation.

En utilisant une 10-folds validation croisée [4] avec la déviance comme mesure de prédiction on détermine le paramètre de régularisation optimal :

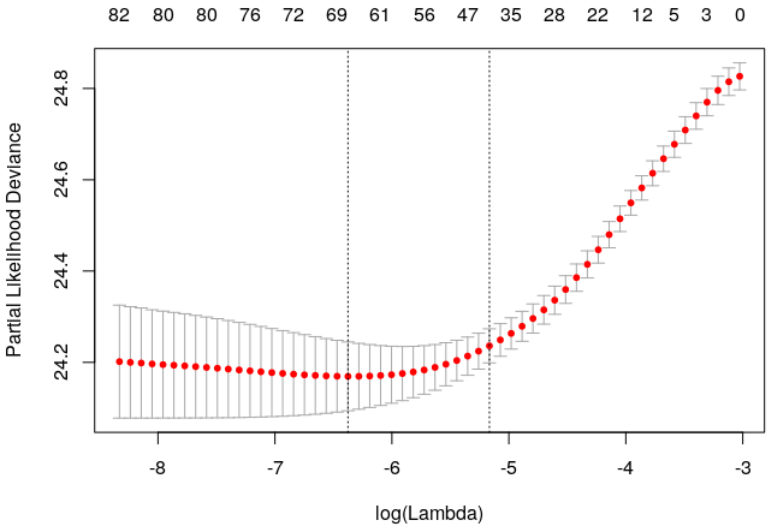


FIGURE 4 – Détermination du paramètre λ de l'Elastic-net pour la dépendance des parisiennes

Les paramètres optimaux et le paramétrage du modèle ont été obtenus pour un temps de calcul de quelques minutes avec la même puissance de calcul que lors de la procédure de sélection AIC qui avait mis plus de 24h à

4. L'étude de sensibilité menée sur le nombre de sous-échantillon de la validation croisée nous a invité à choisir un découpage optimal en 10.

s'exécuter.

Un tableau comparatif donne les coefficients obtenus selon les deux méthodes :

	Parisienne Demence		Picard Demence	
Covariable	Enet	AIC	Enet	AIC
fdr_aud_all	**2,46**	**4,14**	**2,21**	**4,13**
NEURO1_PARKINSON	**1,57**	**1,29**	**-**	**1,11**
NEURO1_EPILEPSIE	**1,43**	**1,21**	**1,18**	**1,21**
NEURO1_DEM_FDR	**1,33**	**1,36**	**2,28**	**1,42**
STROKE1_1HEMO	**1,33**	**1,22**	**1,67**	**-**
ENDOC0_CARENCE	1,20	0,97	0,92	0,89
STROKE1_1ISCHEMIC	**1,15**	**-**	**1,37**	**-**
DIG1_OCCLUSION	1,15	1,05	1,05	-
DIG1_STOMIE	1,14	1,16	1,32	1,32
ENDOC1_GLD_OTHER	1,12	1,06	1,04	1,16
NEURO1_OTHER	1,11	0,72	1,45	1,14
INFECT1_SEPSIS	1,11	1,20	1,44	1,16
fdr_smoker	1,10	2,94	1,65	2,50
KIDNEY0_2PYELONEPHRI	1,09	-	-	0,85
fdr_obesity_all	1,09	1,77	1,23	1,26
NEURO0_SNP_METABO	1,08	1,05	1,07	1,10
CV1_MTE	1,08	-	1,09	1,07
ENDOC0_DIABETE	1,07	1,08	-	1,03
KIDNEY0_CYSTITE_ATCD	1,07	-	1,12	0,90
TRAUMA1_1CRANE	1,06	0,89	-	-
RESP1_2INSUF_AIGUE	1,05	1,09	1,11	-
TRAUMA0_SUICIDE	1,05	1,10	-	-
KIDNEY0_2GN	1,04	-	1,03	-
KIDNEY1_2INSUF_AIGUE	1,04	1,02	-	1,12
KIDNEY1_1INSUF_CHRO	1,04	0,93	1,00	-
DIG1_LIVER_1CirrD	1,04	1,24	1,14	1,21
BLOOD0_2ANEMIA_IRON	1,04	-	0,98	0,93
RHEUM0_AUD	1,04	0,96	-	0,93
DIG0_LIVER_2CirrC	1,04	-	1,01	-
TRAUMA0_3FRACTURE	1,03	0,95	1,02	0,93
DIG0_LIVER_ETIO_ANY	1,03	1,09	1,00	1,05
CANCER1_HEMATO	1,03	1,24	1,17	1,24
CARDIO0_OTHER	1,03	-	0,99	-
ENDOC1_METABO	1,03	1,06	1,04	-
TRAUMA0_CHUTE	1,03	0,96	0,98	0,93
ENDOC0_THYROIDE	1,02	0,99	0,97	-
cp_dipl0	1,02	1,13	1,02	1,05
CANCER1_BREAST	1,02	1,44	1,31	-
CARDIO1_RYTHME_1ACFA	1,02	0,97	0,97	0,87
ENDOC0_DYSLIPIDEMIA	1,02	0,99	0,98	0,97
RESP0_LRI_ATCD	1,02	0,96	1,01	-
CARDIO0_HBP	1,02	0,95	0,96	0,92
TRAUMA0_OSTEOPOROSE	1,02	-	-	1,11
cp_dep	1,01	1,01	0,88	-
cp_immi	1,01	-	1,02	-

	Parisienne Demence		Picard Demence	
Covariable	Enet	AIC	Enet	AIC
STROKE0_3noTTT	1,01	0,97	0,92	-
TRAUMA1_2SEVERE	1,00	-	1,25	-
BLOOD1_1TRANSFUSION	1,00	1,08	1,17	1,10
CANCER1_SMOKER	-	2,13	1,55	1,75
CANCER1_PC_POOR	-	2,03	1,68	1,60
CANCER1_AUD_SMOKER	-	1,66	1,22	1,41
CANCER1_PC_GOOD	-	1,60	1,20	1,22
CANCER1_COLORECTAL	-	1,24	1,08	1,23
BLOOD0_2OTHER	-	1,13	1,04	1,05
RHEUM0_SYSTEME	-	1,08	1,05	-
BLOOD0_2ANEMIA_LYSE	-	1,08	1,18	1,21
RESP0_2OTHER	-	1,06	1,01	-
RESP1_1INSUF_CHRO	-	1,05	1,10	0,92
RESP0_2INTERSTI	-	-	1,01	1,27
CARDIO0_VALVE_ANY	-	1,05	-	1,03
CARDIO0_IHD_2TTT	-	-	0,99	0,97
CANCER0_BENIN	-	0,97	0,98	0,90
DIG0_VB_METABO	-	0,97	0,94	0,89
CANCER0_INSITU	-	0,97	1,03	-
STROKE0_AIT_noSEQ	-	0,98	1,03	-
DIG0_OTHER_noFDR	-	0,95	0,98	0,93
RHEUM0_ARTHROSE_OTHE	-	0,95	1,01	0,91
CARDIO0_RYTHME_3noTT	-	0,94	0,98	0,93
DIG1_HEMORRAGIE	-	0,93	-	-
RESP0_APNEE_SOM	-	0,93	0,99	0,92
RHEUM0_ARTHROSE_KNEE	-	0,90	-	0,87
DIG1_PERITONITE	-	0,89	0,98	-
DIG0_PANCREAS_AUD	-	0,85	-	0,89
CARDIO1_INSUF_CHRO	-	0,96	1,00	0,94
CANCER0_ATCD_FAM	-	-	0,99	0,95
KIDNEY0_2OTHER	-	-	-	0,93
KIDNEY0_2UROLITHIASE	-	-	-	-
RESP0_2ASTHMA	-	-	-	-
RESP0_2COPD	-	0,98	1,03	-
RESP0_BRONCHITE_ATCD	-	-	1,03	-
STROKE0_2TTT	-	-	-	-
CANCER1_PROSTATE	-	-	1,02	-
CV1_PVD	-	-	1,01	-
CARDIO1_IHD_1MI	1,00	0,96	0,91	0,88
CARDIO0_IHD_3noTTT	0,99	0,96	0,96	0,97
RHEUM0_ARTHROSE_HIP	0,99	0,91	-	0,92
SENSE0_CATARACTE	0,98	0,93	0,97	0,90
CARDIO0_RYTHME_2TTT	0,95	0,93	0,94	0,90
CANCER0_SKIN	0,95	0,88	0,97	0,90

TABLE 2 – Comparaison des résultats obtenus entre les méthodes AIC et Elastic-net pour la démence

Les covariables identifiées comme causes d'apparition de la démence dans la première partie de ce travail sur la base d'études reconnues du milieu médical apparaissent en gras.
On remarque que toutes les causes identifiées a priori ressortent en tête du classement.
Ceci confirme que la sélection automatique avec une pénalisation elastic-net est performante et ce malgré les fortes corrélations qui existent dans la base d'étude.

On constate par ailleurs que la sélection est meilleure avec les méthodes d'apprentissage qu'avec la méthode classique. On remarque de plus que le poids du facteur de risque "Alcool" est beaucoup plus important avec la méthode AIC qu'avec l'Elastic net. En effet, par analogie avec l'analyse effectuée sur les jeux de données simulées, cette covariable capte l'information de l'ensemble des variables qui lui sont corrélées, chose que l'Elastic net gère grâce à la partie de la pénalisation de type Ridge qui pénalise la norme l2 des coefficients de régression.

Les taux d'incidence pour les 4 sous-populations (Picards,Parisiennes) x (alcool, non alcool) ont été tracés pour les deux sorties démence et dépendance et comparé aux travaux effectués par le groupe de travail Qalydays opérés sur la France entière et dont les résultats ont fait l'objet de plusieurs publications dans des revues scientifiques mondialement reconnues. Les populations alcooliques ressortent plus risquées que la population française de référence Qalydays et les populations non alcooliques ressortent moins à risque ce qui semble cohérent. De plus, comme attendu, les Picards ont des incidences plus importantes que les Parisiennes du fait de leur différence de sexe, de niveau de vie et d'accès aux soins.

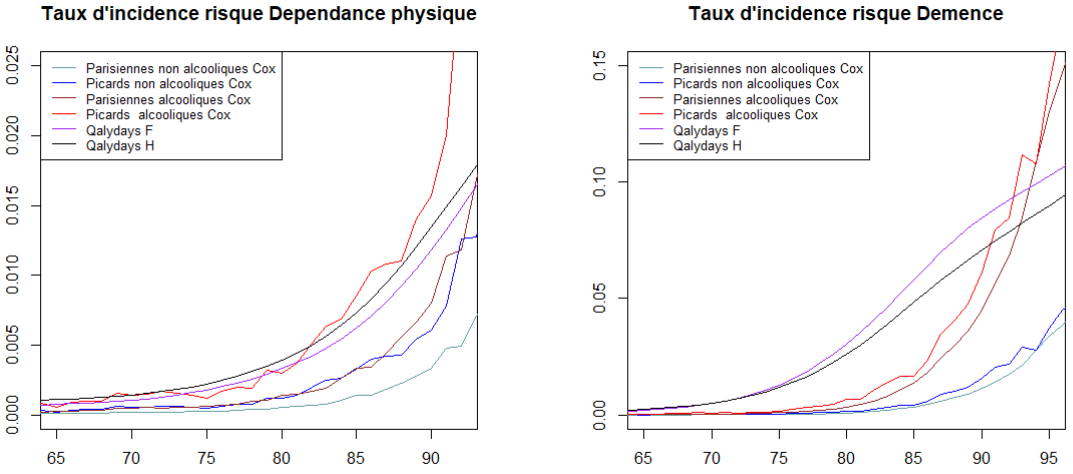


FIGURE 5 – Comparaison des modèles prédictifs de Cox pour la dépendance physique et la démence

Si la qualité d'ajustement n'est pas parfaite, l'avantage du paramétrage des lois par l'intermédiaire de l'Elastic-net est la grande souplesse et la grande rapidité d'exécution et de sélection tout en gérant les problèmes de colinéarité particulièrement complexes lorsque le nombre de variables est aussi important. Cette méthode permet d'industrialiser la modélisation à un grand nombre de sous-populations avec une grande flexibilité et avec des résultats satisfaisants.

Abstract

Definition of dependency and description of available data

Dependency is defined and measured differently, depending on whether you are in the insurance or the medical perimeter. In an insurance context, dependency is defined as the quantity of external resources necessary for a person to carry out the basic acts of daily life, which allows the French public authorities and insurance organisations to distinguish in fine two levels of dependency risk to mark out their aid or coverage : partial dependency or total dependency.
This work is intended for an audience in the insurance sector, but the available data comes from the medical community, a first phase of definition and comparison[5] has reconciled these two definitions. The notions of physical dependence and dementia studied in this thesis correspond to a level of total dependence.

The variables available for this study are from the Information Systems Medicalization Program (IMSP), which correspond to medical pathologies whose appearance, evolution, consequences or even denominations are not known or familiar to the actuary. Knowledge of the medical causes known a priori is indeed essential to validate the models obtained a posteriori. In the light of scientific articles, it was established in this exploratory phase that the medical causes identified by specialists in the medical field for the development of cognitive dependence or dementia are :

- Behavioural risk factors are :
  - ☐ Alcohol
- Among the neurological pathologies :
  - ☐ Epilepsy ;
  - ☐ Parkinson's disease ;
    (The disease occurs around the age of 70, and Parkinson's dementia 10 to 15 years later, it affects people with very advanced ages.)
  - ☐ Multiple sclerosis
  - ☐ Hydrocephalus at normal pressure
  - ☐ Encephalitis
- Strokes
- Cardiac pathologies :
  - ☐ High blood pressure
- Diabetes

The purpose of this non-exhaustive list of medical causes of dementia is to serve as a benchmark to arbitrate the relevance of the variable selection made by the models parameterized by learning methods.

Application of a hasard proportional multiplicative model

In this work, the outputs are modelled using a Cox proportional hasard multiplicative model. This model, which will be described in more details later, aims to measure the differences between different sub-populations, subject to certain assumptions, including the independence of the covariates over time and the assumption of risk

5. From SCHWARZINGER [2018]

proportionality. The advantage of this model is that it meets the informative criterion of censorship. However, for the sake of simplicity, random censorship induced by death has been treated as a simple censorship, which leads to modelling bias and affects the estimation of relative differences between the different sub-populations studied. The study and the introduction of competition between the differents causes of exit from the study scope makes it possible to remove this bias, we are considering models with competing risks. This thesis, which focuses on comparison with learning methods, is limited to the use of non-competing models.

In order to verify the hypotheses of application of a proportional and time-independent random model, the comparison of gross rates with Cox models was carried out on 4 sub-populations that were sufficiently heterogeneous in terms of dependency risk (to capture differences) and sufficiently large (to limit problems due to too high volatility). This comparison for the (Picards, Parisian) x (alcoholic, non-alcoholic) for the two causes of exits validated the relevance of using this type of model and validated the necessary conditions of temporal independence and proportional risk.

Other methods such as Schönefeld residues analysis or χ^2 adequacy tests were conducted but did not clearly validate or invalidate Cox's assumptions.

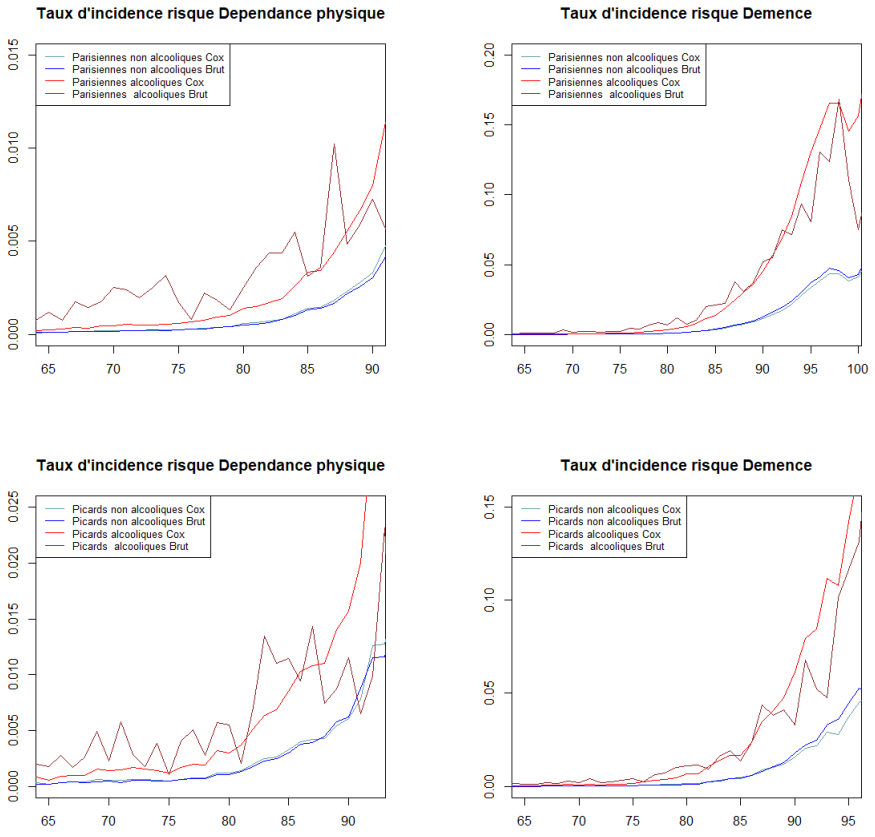


FIGURE 6 – Gross rates and predicted rates by Cox for different sub-populations

It can be seen that the prediction curve approximates the raw observation curve. In addition, the two sub-populations have similar trends, which makes it possible to accept the assumption of proportionality of risks. The Cox model being well adapted, a variable selection method using the AIC criterion and a step-by-step method made it possible to make a first variable selection. The pathologies previously identified as at risk stand out relatively well.

	Parisienne Demence	Picard Demence	Parisienne Dependance	Picard Dependance		Parisienne Demence	Picard Demence	Parisienne Dependance	Picard Dependance
Covariable	Stepwise	Stepwise	Stepwise	Stepwise	Covariable	Stepwise	Stepwise	Stepwise	Stepwise
fdr_aud_all	**4,14**	**4,13**	3,77	4,02	DIG0_VB_METABO	0,87	0,97	0,89	0,98
fdr_smoker	2,94	2,50	3,27	2,60	CARDIO0_IHD_3noTTT	0,96	0,97	0,97	0,96
CANCER1_SMOKER	2,13	1,75	2,15	1,77	RESP0_LRI_ATCD	0,98	0,96	0,98	0,97
CANCER1_PC_POOR	2,03	1,60	2,09	1,63	RHEUM0_AUD	0,94	0,96	0,93	0,95
fdr_obesity_all	1,77	1,26	2,01	1,36	CARDIO1_INSUF_CHRO	0,96	0,96	0,95	1,00
CANCER1_AUD_SMOKER	1,66	1,41	1,68	1,42	TRAUMA0_CHUTE	0,88	0,96	0,93	0,92
CANCER1_PC_GOOD	1,60	1,22	1,63	1,24	CARDIO1_IHD_1MI	0,84	0,96	0,88	1,01
CANCER1_BREAST	1,44	-	1,51	-	TRAUMA0_3FRACTURE	0,94	0,95	0,93	0,93
NEURO1_DEM_FDR	**1,36**	**1,42**	1,34	1,32	RHEUM0_ARTHROSE_OTHE	0,91	0,95	0,91	0,92
NEURO1_PARKINSON	**1,29**	**1,11**	1,22	-	CARDIO0_HBP	0,92	0,95	0,92	0,92
DIG1_LIVER_1CirrD	1,24	1,21	1,25	1,22	DIG0_OTHER_noFDR	0,93	0,95	0,93	0,94
CANCER1_COLORECTAL	1,24	1,23	1,30	1,23	CARDIO0_RYTHME_3noTT	0,93	0,94	0,93	0,92
CANCER1_HEMATO	1,24	1,24	1,31	1,31	RESP0_APNEE_SOM	0,89	0,93	0,92	0,91
STROKE1_1HEMO	**1,22**	-	1,22	-	CARDIO0_RYTHME_2TTT	0,91	0,93	0,90	0,94
NEURO1_EPILEPSIE	**1,21**	**1,21**	1,13	1,21	SENSE0_CATARACTE	0,90	0,93	0,90	0,93
INFECT1_SEPSIS	1,20	1,16	1,19	1,18	KIDNEY1_1INSUF_CHRO	0,96	0,92	0,95	0,93
DIG1_STOMIE	1,16	1,32	1,21	1,30	DIG1_HEMORRAGIE	0,97	0,93	0,95	0,91
BLOOD0_2OTHER	1,13	1,05	1,15	-	RHEUM0_ARTHROSE_HIP	0,92	0,91	0,92	0,90
cp_dipl0	1,13	1,05	1,16	1,06	RHEUM0_ARTHROSE_KNEE	0,87	0,90	0,87	0,89
TRAUMA0_SUICIDE	1,10	-	1,08	-	TRAUMA1_1CRANE	0,92	0,89	0,94	0,85
DIG0_LIVER_ETIO_ANY	1,09	1,05	1,10	1,05	DIG1_PERITONITE	0,99	0,89	0,97	0,89
RESP1_2INSUF_AIGUE	1,09	-	1,11	-	CANCER0_SKIN	0,92	0,88	0,90	0,89
BLOOD0_2ANEMIA_LYSE	1,08	1,21	1,10	1,22	DIG0_PANCREAS_AUD	0,91	0,85	0,89	0,84
BLOOD1_1TRANSFUSION	1,08	1,10	1,11	1,10	NEURO1_OTHER	1,11	0,72	1,14	0,68
RHEUM0_SYSTEME	1,08	-	1,12	-	RESP0_2INTERSTI	1,29	1,04	1,26	1,08
ENDOC0_DIABETE	1,08	1,03	1,07	1,04	DIG0_LIVER_2CirrC	0,98	1,02	0,99	1,05
ENDOC1_METABO	1,06	-	1,06	-	CV1_MTE	1,08	1,01	1,07	1,04
ENDOC1_GLD_OTHER	1,06	1,16	1,06	1,12	KIDNEY0_2GN	0,99	1,01	1,03	1,04
RESP0_2OTHER	1,06	-	1,09	-	CARDIO0_IHD_2TTT	0,97	1,01	0,97	1,03
CARDIO0_VALVE_ANY	1,05	1,03	1,07	1,04	cp_immi	0,99	1,00	1,00	0,99
NEURO0_SNP_METABO	1,05	1,10	1,06	1,08	BLOOD0_2ANEMIA_IRON	0,95	0,99	0,94	0,98
DIG1_OCCLUSION	1,05	-	1,03	-	KIDNEY0_CYSTITE_ATCD	0,91	0,99	0,89	0,98
RESP1_1INSUF_CHRO	1,05	0,92	1,08	0,94	TRAUMA0_OSTEOPOROSE	1,08	1,01	1,11	0,98
KIDNEY1_2INSUF_AIGUE	1,02	1,12	1,03	1,11	CANCER0_ATCD_FAM	0,94	0,99	0,95	0,99
cp_dep	1,01	-	1,01	-	KIDNEY0_2OTHER	0,92	0,99	0,93	1,00
ENDOC0_DYSLIPIDEMIA	0,99	0,97	0,97	0,96	KIDNEY0_2PYELONEPHRI	0,89	1,01	0,86	0,99
ENDOC0_THYROIDE	0,99	-	0,97	0,94	CARDIO0_OTHER	1,00	1,01	1,01	0,99
RESP0_2COPD	0,98	-	-	-	KIDNEY0_2UROLITHIASE	0,95	0,97	0,97	0,99
STROKE0_AIT_noSEQ	0,98	-	0,97	-	RESP0_2ASTHMA	1,01	1,00	0,99	1,01
CANCER0_BENIN	0,97	0,90	0,98	0,90	RESP0_BRONCHITE_ATCD	1,06	1,00	1,02	1,02
CARDIO1_RYTHME_1ACFA	0,97	0,87	-	0,87	STROKE0_2TTT	1,06	1,04	1,03	1,03
CANCER0_INSITU	0,97	-	0,98	-	CANCER1_PROSTATE	1,04	0,00	1,03	-
ENDOC0_CARENCE	0,97	0,89	0,91	0,83	CV1_PVD	1,01	0,98	1,01	1,01
STROKE0_3noTTT	0,97	-	0,96	-	**STROKE1_1ISCHEMIC**	0,95	**1,02**	**0,97**	1,00
					TRAUMA1_2SEVERE	0,94	1,03	0,97	1,03

TABLE 3 – AIC fitting and selection. Pre-identify covariate are in bold

However, several problems are highlighted. First of all, the step-by-step procedure is extremely time-consuming. It took more than 24 hours to execute despite the high computing power available. In addition, the correlation study normally required prior to the selection process could not be carried out given the excessive size of the database (89 variables) and the very important inter-correlationships between all pathologies.

Use of penalized regression

The idea of penalized regression is to introduce the step of penalization directly into the problem of optimizing likelihood by adding a constraint on the standard of regression coefficients of the form :

$$\sum_{j=1}^{p} |\beta_j|^{\delta} \leq C$$

The region of possible values for the regression coefficients then depends on the value of the parameter δ. For instance, in dimension 2, possible values for the regression coefficients β_1 and β_2 regarding parameter δ have the following forms :

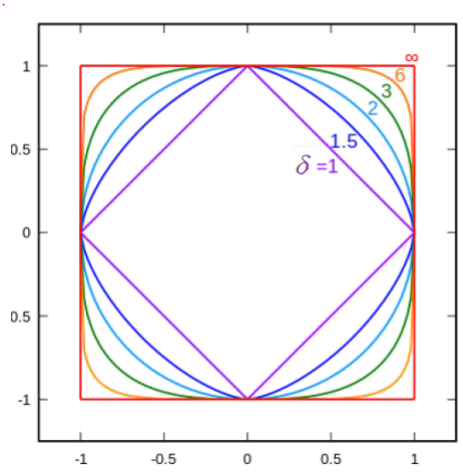


FIGURE 7 – Regions of constraints according to δ

The paper of VERWEIJ[1996] makes it possible to transpose the penalty method commonly used for the ordinary least squares optimization method to the likelihood optimization calculation.
Imposing a constraint on the coefficient standard then amounts to directly expressing the penalized partial likelihood in the following way :

$$L_{partielle}(\beta) = \prod_{i}^{D} \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - p(\beta, \lambda)$$

where $p(\beta, \lambda)$ is the penalized function. λ is the intensity parameter or regulation parameter. This parameter moderates the intensity of the penalty in a linear way.
By varying the value of the regularization constant, we then obtain the following set of solutions :

$$\{\hat{\beta}(\lambda) \in [0, +\infty[\text{ où } \hat{\beta}(\lambda) = argmin_\beta \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - p(\beta, \lambda)\}$$

A number of simulations were conducted to test the different penalties in order to select the most appropriate one to apply to the actual data set available for this study.

For all simulations, censored datasets whose variable of interest verifies the Cox relationship were simulated :

$$\alpha(t|\mathbf{Z}_i(t)) = \alpha_0(t) exp^{\mathbf{Z}_i^T \beta}$$

where :

- The basic hazard function was taken equal to $\alpha_0(t) = 2.t$ to have a simple and time-dependent expression (factor 2 was chosen to simplify it when integrating for the calculation of the survival law). This choice amounts to considering that in the absence of the effect of the other variables, the incidence rate increases linearly over time ;
- For the vector of covariates, it was taken for each individual i a value random according to a uniform law on the interval [-1 ;1] : $\forall\ i \in [\![1; n]\!]\ \mathbf{Z}_i \sim \mathcal{U}[-1; 1]$
  For some simulations, more or less strong correlations between certain variables were applied to measure their effects ;
- The fact of setting some coefficients to zero has finally made it possible to measure quality selection of estimators since this means imposing that the associated variables have no influence on the output variable. A good estimator to exclude them. ;

The law of survival as a function of time for an individual i being :

$$S(t, \mathbf{Z}_i) = exp[-\int_0^t \alpha(s, \mathbf{Z}_i) ds]$$

The intensity of the survival law having been chosen from the form $\alpha_0(t) = 2.t$, we obtain :

$$S(t, \mathbf{Z}_i) = exp[-\int_0^t 2s.exp^{\mathbf{Z}_i^T \beta} ds]$$

$$S(t, \mathbf{Z}_i) = exp(-t^2.e^{\mathbf{Z}_i^T \beta})$$

To simulate the occurrence times T_i for each individual i a Monte Carlo method was used by reversing the conditional survival function and simulating uniform random variables $Y_i \in \mathcal{U}[0; 1]$

We then obtain for each individual i :

$$T_i = \sqrt{-exp^{-\mathbf{Z}_i^T \beta} log(Y_i)}$$

The Monte Carlo method consists in simulating the random variable Y_i (very simple with the R software) to deduct the T_i times distributed according to a Cox distribution. .

By setting the intrinsic β parameters of the distribution in advance, it was then possible to apply the different regression techniques penalized by LASSO, Ridge, Elastic-net and Adaptive-Lasso to compare them to the expected and thus deduce which penalty was the most appropriate for data sets including different levels of correlations, censorship rates or to measure the influence of sample size.
These simulations show that Lasso regression allows variable selection since it allows coefficients to be zero. The Ridge regression is used to manage problems related to collinearity. Finally, the Elastic-net regression allows the effects of the two penalties to be combined by making a selection and managing the unwanted effects of collinearity. While these conclusions are relatively classic in the context of the penalisation of a regression by the least squares method, they are well reflected in the penaliszation of Cox's partial likelihood.

The penalty chosen in view of the large size of the study base, its high collinearity and its censored nature was thus the Elastic-net regression.

The problem of optimizing Cox's partial likelihood penalized by an Elastic-Net penalty is as follows :

$$L_{partielle}(\beta) = \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - \lambda(\frac{1}{2}\sum_{j=1}^p \beta_j^2 + \frac{1}{2}\sum_{j=1}^p |\beta_j|)$$

By varying the value of the regularization constant, we then obtain the following set of solutions :

$$\{\hat{\beta}(\lambda) \in [0, +\infty[\text{ où } \hat{\beta}(\lambda) = argmin_\beta \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - \lambda(\frac{1}{2}\sum_{j=1}^p \beta_j^2 + \frac{1}{2}\sum_{j=1}^p |\beta_j|)\}$$

The regularization path below is the representation of all the values taken by each of the regression coefficients as a function of the value of the regularization parameter λ.

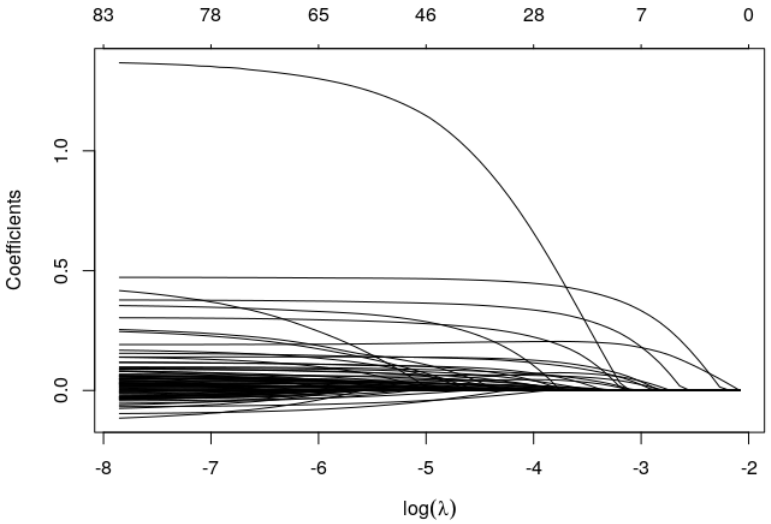


FIGURE 8 – Demencia of Parisian women

This work was carried out on Parisian women population and Picard men population for the risks of physical dependency and dementia.

The optimal regulation parameter that minimizes likelihood is a combination of the α parameter that gives more or less weight to the two components of the net elastic relative to norme l1 and norme l2 and the β parameter that modulates the intensity of the penalty.

Using a 10-folds cross-validation[6] was with deviance as a predictive measure we determine the optimal regulation parameter :

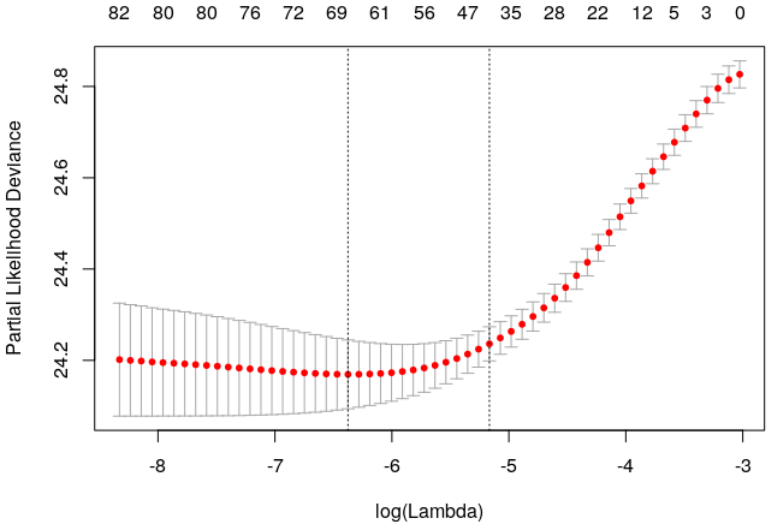


FIGURE 9 – Determination of the Elastic-net parameter λ for the demencia of the Parisian women

The optimal parameters and model settings were obtained for a computation time of few minutes with the same computation power as during the AIC selection procedure, which took more than 24 hours to execute.

6. The sensitivity study conducted on the number of sub-samples of the cross-validation invited us to choose an optimal breakdown into 10.

A comparative table gives the coefficients obtained according to the two methods :

	Parisienne Demence		Picard Demence	
Covariable	Enet	AIC	Enet	AIC
fdr_aud_all	**2,46**	**4,14**	**2,21**	**4,13**
NEURO1_PARKINSON	**1,57**	**1,29**	**-**	**1,11**
NEURO1_EPILEPSIE	**1,43**	**1,21**	**1,18**	**1,21**
NEURO1_DEM_FDR	**1,33**	**1,36**	**2,28**	**1,42**
STROKE1_1HEMO	**1,33**	**1,22**	**1,67**	**-**
ENDOC0_CARENCE	1,20	0,97	0,92	0,89
STROKE1_1ISCHEMIC	**1,15**	**-**	**1,37**	**-**
DIG1_OCCLUSION	1,15	1,05	1,05	-
DIG1_STOMIE	1,14	1,16	1,32	1,32
ENDOC1_GLD_OTHER	1,12	1,06	1,04	1,16
NEURO1_OTHER	1,11	0,72	1,45	1,14
INFECT1_SEPSIS	1,11	1,20	1,44	1,16
fdr_smoker	1,10	2,94	1,65	2,50
KIDNEY0_2PYELONEPHRI	1,09	-	-	0,85
fdr_obesity_all	1,09	1,77	1,23	1,26
NEURO0_SNP_METABO	1,08	1,05	1,07	1,10
CV1_MTE	1,08	-	1,09	1,07
ENDOC0_DIABETE	1,07	1,08	-	1,03
KIDNEY0_CYSTITE_ATCD	1,07	-	1,12	0,90
TRAUMA1_1CRANE	1,06	0,89	-	-
RESP1_2INSUF_AIGUE	1,05	1,09	1,11	-
TRAUMA0_SUICIDE	1,05	1,10	-	-
KIDNEY0_2GN	1,04	-	1,03	-
KIDNEY1_2INSUF_AIGUE	1,04	1,02	-	1,12
KIDNEY1_1INSUF_CHRO	1,04	0,93	1,00	-
DIG1_LIVER_1CirrD	1,04	1,24	1,14	1,21
BLOOD0_2ANEMIA_IRON	1,04	-	0,98	0,93
RHEUM0_AUD	1,04	0,96	-	0,93
DIG0_LIVER_2CirrC	1,04	-	1,01	-
TRAUMA0_3FRACTURE	1,03	0,95	1,02	0,93
DIG0_LIVER_ETIO_ANY	1,03	1,09	1,00	1,05
CANCER1_HEMATO	1,03	1,24	1,17	1,24
CARDIO0_OTHER	1,03	-	0,99	-
ENDOC1_METABO	1,03	1,06	1,04	-
TRAUMA0_CHUTE	1,03	0,96	0,98	0,93
ENDOC0_THYROIDE	1,02	0,99	0,97	-
cp_dipl0	1,02	1,13	1,02	1,05
CANCER1_BREAST	1,02	1,44	1,31	-
CARDIO1_RYTHME_1ACFA	1,02	0,97	0,97	0,87
ENDOC0_DYSLIPIDEMIA	1,02	0,99	0,98	0,97
RESP0_LRI_ATCD	1,02	0,96	1,01	-
CARDIO0_HBP	1,02	0,95	0,96	0,92
TRAUMA0_OSTEOPOROSE	1,02	-	-	1,11
cp_dep	1,01	1,01	0,88	-
cp_immi	1,01	-	1,02	-

	Parisienne Demence		Picard Demence	
Covariable	Enet	AIC	Enet	AIC
STROKE0_3noTTT	1,01	0,97	0,92	-
TRAUMA1_2SEVERE	1,00	-	1,25	-
BLOOD1_1TRANSFUSION	1,00	1,08	1,17	1,10
CANCER1_SMOKER	-	2,13	1,55	1,75
CANCER1_PC_POOR	-	2,03	1,68	1,60
CANCER1_AUD_SMOKER	-	1,66	1,22	1,41
CANCER1_PC_GOOD	-	1,60	1,20	1,22
CANCER1_COLORECTAL	-	1,24	1,08	1,23
BLOOD0_2OTHER	-	1,13	1,04	1,05
RHEUM0_SYSTEME	-	1,08	1,05	-
BLOOD0_2ANEMIA_LYSE	-	1,08	1,18	1,21
RESP0_2OTHER	-	1,06	1,01	-
RESP1_1INSUF_CHRO	-	1,05	1,10	0,92
RESP0_2INTERSTI	-	-	1,01	1,27
CARDIO0_VALVE_ANY	-	1,05	-	1,03
CARDIO0_IHD_2TTT	-	-	0,99	0,97
CANCER0_BENIN	-	0,97	0,98	0,90
DIG0_VB_METABO	-	0,97	0,94	0,89
CANCER0_INSITU	-	0,97	1,03	-
STROKE0_AIT_noSEQ	-	0,98	1,03	-
DIG0_OTHER_noFDR	-	0,95	0,98	0,93
RHEUM0_ARTHROSE_OTHE	-	0,95	1,01	0,91
CARDIO0_RYTHME_3noTT	-	0,94	0,98	0,93
DIG1_HEMORRAGIE	-	0,93	-	-
RESP0_APNEE_SOM	-	0,93	0,99	0,92
RHEUM0_ARTHROSE_KNEE	-	0,90	-	0,87
DIG1_PERITONITE	-	0,89	0,98	-
DIG0_PANCREAS_AUD	-	0,85	-	0,89
CARDIO1_INSUF_CHRO	-	0,96	1,00	0,94
CANCER0_ATCD_FAM	-	-	0,99	0,95
KIDNEY0_2OTHER	-	-	-	0,93
KIDNEY0_2UROLITHIASE	-	-	-	-
RESP0_2ASTHMA	-	-	-	-
RESP0_2COPD	-	0,98	1,03	-
RESP0_BRONCHITE_ATCD	-	-	1,03	-
STROKE0_2TTT	-	-	-	-
CANCER1_PROSTATE	-	-	1,02	-
CV1_PVD	-	-	1,01	-
CARDIO1_IHD_1MI	1,00	0,96	0,91	0,88
CARDIO0_IHD_3noTTT	0,99	0,96	0,96	0,97
RHEUM0_ARTHROSE_HIP	0,99	0,91	-	0,92
SENSE0_CATARACTE	0,98	0,93	0,97	0,90
CARDIO0_RYTHME_2TTT	0,95	0,93	0,94	0,90
CANCER0_SKIN	0,95	0,88	0,97	0,90

TABLE 4 – Comparison of results obtained between AIC and Elastic-net methods for demencia

The covariates identified as causes of the onset of dementia in the first part of this work based on recognized studies from the medical community appear in bold. It should be noted that all the causes identified a priori are at the top of the ranking. This confirms that automatic selection with elastic-net penalisation is effective despite the strong correlations that exist in the study database.

It can also be seen that the selection is better with learning methods than with the traditional method. It should also be noted that the weight of the risk factor "Alcohol" is much higher with the AIC method than with the Elastic Net. Indeed, by analogy with the analysis carried out on the simulated data sets, this covariate captures the information of all the variables correlated to it, something that the Net Elastic manages thanks to the part of the Ridge type penalty that penalizes the l2 norm of regression coefficients.

The incidence rates for the 4 sub-populations (Picard, Paris) x (alcohol, non-alcohol) were plotted for the two dementia and dependencies outings and compared to the work carried out by the Qalydays working group operating throughout France, the results of which have been published in several internationally recognised scientific journals. Alcoholic populations appear to be more risky than the French reference population Qalydays and non-alcoholic populations appear to be less at risk, which seems consistent. Moreover, as expected, Picards have a greater impact than Parisian women because of their difference in gender, standard of living and access to healthcare.

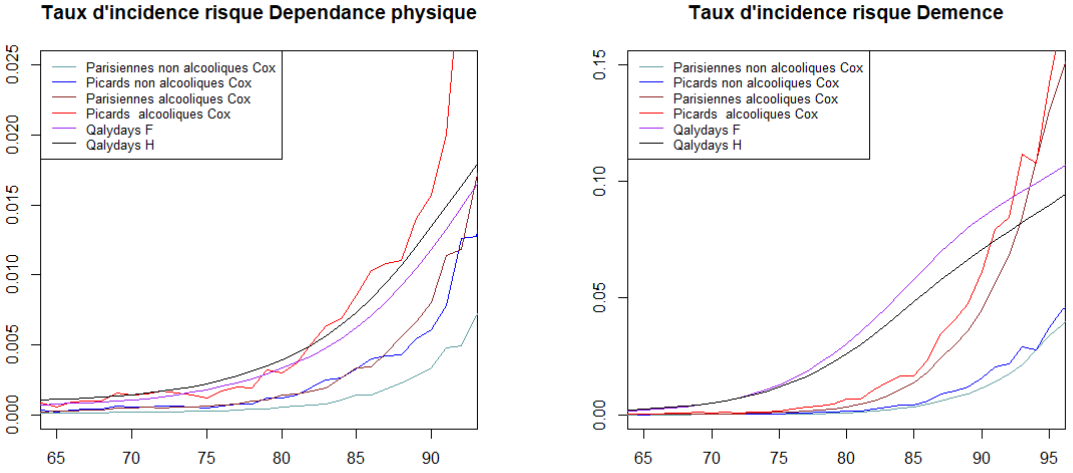


FIGURE 10 – Comparison of Cox predictive model for demencia

Although the adjustment quality is not perfect, the advantage of law parameter setting via the Elastic-net is the great flexibility and speed of execution and selection while managing particularly complex collinearity problems when the number of variables is so large. This method makes it possible to industrialize modelling to a large number of sub-populations with great flexibility and satisfactory results.

Chapitre 1

Cadre de l'étude

1.1 Les risques dépendance et démence

1.1.1 Définitions et mesure de la « Dépendance », la « Démence » et la « Perte d'autonomie»

Point de vue des assureurs

L'avancée en âge s'accompagne d'un accroissement des risques de perte d'autonomie, avec pour corollaire l'augmentation des situations de dépendance. Si les notions de perte d'autonomie et de dépendance sont très proches, le terme "perte d'autonomie" est souvent préféré puisque moins stigmatisant.
On parle d'état de dépendance lorsque une personne a besoin de faire appel à un tiers pour effectuer un ou plusieurs Actes de la Vie Quotidienne (AVQ). L'origine de cet état peut être physique ou psychique. D'où la nécessité d'appréhender ces deux composantes dans la définition du risque couvert.
Pour mesurer ce risque, les pouvoirs publics français ont mis en place une grille qui permet de quantifier le degré de dépendance.
Ce degré est évalué sur la base de dix critères, mesurant l'autonomie physique et psychique.
En fonction de leur degré de difficulté pour réaliser ces actions, les personnes sont classées en six groupes, dits « iso ressources » (GIR), de la dépendance la plus légère, le GIR 6, à la plus élevée, le GIR 1. Ces six groupes sont censés regrouper des personnes qui peuvent avoir des profils d'incapacités différents, mais ont besoin d'une même quantité d'heures de soins. Ces GIR permettent alors de distinguer deux catégories : la Dépendance Totale et la Dépendance Partielle qui modulent le niveau de garantie présent dans les contrats d'assurance dépendance en France.

La perte d'autonomie totale a fait l'objet d'une standardisation récente, le label GAD 2016. Ce label comprend l'évaluation de 5 AVQ et du MMSE pour le déficit cognitif.
La perte d'autonomie totale est ainsi identifiée selon son type :

- Physique : ≥ 4 AVQ4 (en dépendance totale) ;
- Cognitive : déficit cognitif sévère (MMSE ≤ 10) et ≥ 2 AVQ3 (en dépendance partielle voire totale) ;
- Mixte : déficit cognitif modéré (MMSE ≤ 15) et ≥ 3 AVQ4 (en dépendance totale).

Dans un contexte assurantiel, l'état de dépendance est donc défini comme la quantité de ressources extérieures nécessaires à une personne pour effectuer les actes basiques de la vie quotidienne. Les termes « Perte d'Autonomie » et « Dépendance » sont synonymes et regroupent les notions de « dépendance physique » et « dépendance cognitive » (ou « démence ») sans distinction ; seules leurs conséquences en terme de niveau d'aide par un tiers importent. Ce niveau d'aide est quantifiable (via la grille AGGIR ou le nombre d'AVQ) ce qui permet aux pouvoirs publiques et aux organismes d'assurance français de distinguer in fine deux niveaux de risque dépendance pour jalonner leurs aides ou couvertures. La dépendance partielle et la dépendance totale.

Point de vue de la profession Médicale

SCHWARZINGER [2018] introduit une correspondance entre la définition médicale de la perte d'autonomie, c'est-à-dire celle respectant le codage médical préconisé par l'Organisation Mondiale de la Santé, le code CIM-10, et la définition assurantielle qui différencie la dépendance totale et la dépendance partielle par l'intermédiaire des notions d'AVQ et de Groupe Iso Ressources.
Pour ce faire, le docteur Schwarzinger s'intéresse à une population de patients qui ont l'avantage d'avoir été d'une part admis à l'hôpital, donc bénéficiant du codage médical CIM-10 et d'autre part ayant été admis dans des secteurs hospitaliers impliquant également l'évaluation systématique des 6 AVQ.

Identification de la perte d'autonomie à l'hôpital via les standards de l'OMS

Le paragraphe suivant issu de SCHWARZINGER [2018] définit l'identification de la perte d'autonomie à l'hôpital :
«Le codage médical à l'hôpital est standardisé selon la CIM-10. Il permet d'identifier deux types de perte d'autonomie à l'hôpital :

- *Le codage médical d'une « démence » au sens large (maladies d'Alzheimer et apparentées) permet d'identifier toute perte d'autonomie dite « cognitive » (SCHWARZINGER et al. [2018]). De plus, le codage médical permet de préciser le niveau de déficit cognitif, notamment l'existence d'un déficit cognitif sévère (CIM-10 : F00.xx2 ; F01.xx2 ; F02.xx2 ; F03.xx2), i.e., une perte d'autonomie cognitive « totale ». ;*
- *Le codage médical d'un état « grabataire » permet d'identifier une perte d'autonomie « physique » totale. En effet, un état grabataire (CIM-10 : R26.30) est défini par l'« état d'une personne confinée au lit ou au fauteuil par sa maladie, incapable de subvenir seule sans aide et en toute sécurité à ses besoins alimentaires, d'hygiène personnelle, d'élimination et d'exonération, de transfert et de déplacement.*

Dans cette étude, nous avons considéré deux types de perte d'autonomie (cognitive ou physique) au caractère mutuellement exclusif. En effet, dans les rares cas (<5 %) où une démence et un état grabataire sont enregistrés dans le suivi du patient, l'enregistrement de la démence précède le plus souvent celui de l'état grabataire que nous avons alors identifié comme une « perte d'autonomie cognitive totale ».
Le codage médical d'une « démence » sans enregistrement d'un déficit cognitif sévère interroge sur le caractère partiel ou total de la perte d'autonomie cognitive. En particulier, l'enregistrement du niveau de déficit cognitif suppose une connaissance fine du codage médical. En effet, la facturation des séjours impose un codage CIM-10 de base au 4ème rang alphanumérique, alors que l'enregistrement du niveau de déficit cognitif nécessite de coder jusqu'au 6ème rang alphanumérique (F00.xx2 ; F01.xx2 ; F02.xx2 ; F03.xx2). Dès lors, il est fort probable que seuls les médecins spécialistes (i.e., neurologues) aient recours à cette possibilité sans qu'il soit possible d'inférer le niveau de déficit cognitif pour la grande majorité des cas enregistrés en dehors des services spécialisés. Aussi, nous avons cherché à documenter le niveau de perte d'autonomie pour ces patients à partir des AVQ enregistrées pour la partie des patients hospitalisés en soins de suite (SSR, HAD, PSY).»

Caractère « total » de la perte d'autonomie identifiée à l'hôpital

L'étude de correspondance entre les codes CIM-10 de l'OMS et la vision assurantielle d'AVQ permet de classer la perte d'autonomie constatée à l'hôpital comme une dépendance totale dans l'échelle assurantielle (SCHWARZINGER [2018]) :
« Dans trois secteurs hospitaliers (SSR, HAD, PSY), les soignants évaluent systématiquement 6 AVQ à l'admission puis toutes les semaines pendant la durée d'hospitalisation. Les 6 AVQ mesurés à l'hôpital présentent logiquement de fortes similarités avec les 5 AVQ du label GAD, notamment les 4 AVQ évaluant la dimension « physique » de la perte d'autonomie.
De plus, le niveau d'autonomie est évalué pour chaque AVQ selon la même échelle ordinale :

- indépendance (sans l'intervention d'un tiers) ;
- supervision ou arrangement (présence d'un tiers sans contact physique) ;
- dépendance partielle (aide par un tiers) ;
- dépendance totale (réalisation par un tiers). »

Une approche psychométrique basée sur la théorie de la réponse à l'item a été menée par l'auteur en considérant

qu'une échelle latente du niveau d'autonomie sous-tend l'ensemble des évaluations des 6 AVQ (De Ayala [2009]). L'analyse de ces résultats et la comparaison avec les données des patients admis en soins de suite a permis à l'auteur d'établir que le codage médical de toute perte d'autonomie à l'hôpital (démence ou état grabataire) correspondait globalement à une perte d'autonomie « totale » dans l'échelle de mesure assurantielle du risque Dépendance. Pour prendre connaissance de la démarche complète, le lecteur est invité à se référer à l'article original
(http ://www.ressources-actuarielles.net/C1256CFC001E6549/0/3F7A221B51D221
A2C125839500242D68).

Identification des causes médicales de l'apparition de la démence.

La correspondance entre la définition assurantielle et la définition médicale de la notion de dépendance a été décrite précédemment de manière à pouvoir établir un lien entre la variable de sortie étudiée dans la base du PMSI et les standards assurantiels.
Cependant, si la base de données disponible pour l'étude contient certaines variables sociologiques ou des facteurs de risque identifiables et compréhensibles par l'actuaire, la base contient également des variables dont la compréhension et l'interprétation sont difficiles voire impossibles pour toutes personnes non-initiées au milieu médical.
Dans la suite de ce mémoire, des méthodes dites d'apprentissage vont être utilisées sur ces données médicales. Ces méthodes sont connues pour prétendre pouvoir réussir à s'affranchir de « l'avis d'expert » normalement indispensable à toute étude statistique de tous domaines confondus. Pour pouvoir vérifier la véracité de cette assertion, il nous faut pouvoir établir un lien a priori entre la variable de sortie et les variables explicatives pathologiques disponibles afin de pouvoir vérifier in fine si les modèles d'apprentissage obtenus sont satisfaisants et ont par conséquent réussis à réellement s'affranchir de « l'avis d'expert médical».
Un travail de recherche documentaire a été mené de manière à établir au moyen d'études médicales reconnues, les facteurs pathologiques susceptibles d'induire le passage à l'état de dépendance (ici cognitive).
Les liens établis ici seront alors confrontés aux résultats obtenus avec les méthodes d'apprentissage, notamment en termes de sélection des variables pertinentes nécessaires à la construction des modèles de prédiction de la dépendance.
Le lecteur est invité à se référer à l'annexe 1 pour y trouver la description et la mise en évidence des causes et facteurs médicaux de l'apparition de la démence. Les notions de prévalence et d'incidence y sont également rappelées. En effet, pour quantifier la présence (ou l'apparition) d'un phénomène, le médecin utilise souvent la notion de prévalence tandis que l'actuaire s'appuie lui sur celle de l'incidence.

Synthèse des causes médicales d'apparition de la démence.

A la lumière de ces articles, nous pouvons établir que les causes médicales identifiées par les spécialistes du domaine médicale pour l'apparition de la dépendance cognitive ou démence sont :

- Parmi les facteurs de risques comportementaux :
  - ☐ L'alcool
- Parmi les pathologies neurologiques :
  - ☐ L'épilepsie ;
  - ☐ La maladie de Parkinson ;
    (La maladie se déclarant aux alentours de 70 ans, et la démence Parkinsonienne 10 à 15 ans après, celle-ci touche les personnes avec des âges très avancés.)
  - ☐ La sclérose en plaque
  - ☐ L'hydrocéphalie à pression normale
  - ☐ L'encéphalite
- Les accidents vasculaires cérébraux
- Les pathologies cardiaques :
  - ☐ L'hypertension artérielle
- Le diabète

Cette liste n'est pas exhaustive. Cependant le nombre relativement important de causes médicales de démence identifiées ici va nous servir d'étalon pour arbitrer quant à la pertinence de la sélection de variables opérée par les modèles paramétrés par le biais des méthodes d'apprentissage. Le travail de comparaison entre les modèles prédictifs de la dépendance physique avec ses causes médicales connues n'a pas été traité mais le niveau de pertinence de la sélection constatée sur la variable de sortie "démence" permettra, par analogie, de juger la pertinence de sélection des modèles de prédiction de la dépendance physique.

1.1.2 Description de contrats Dépendance

Les assurés qui se couvrent contre le risque dépendance ont généralement le choix en ce qui concerne la forme du versement des prestations une fois la dépendance survenue : sous forme de rente ou sous forme de capital. Dans les deux cas, l'assuré devra payer une prime d'assurance (généralement annuelle) calculé en fonction du montant des prestations souhaitées et de l'âge de l'assuré à partir du moment de la souscription jusqu'à la survenance. La bonne évaluation du montant de cotisation à fixer par l'assureur est donc fonction d'une part des taux d'incidence en dépendance (traités dans ce mémoire) mais également des taux de maintien dans cet état (autrement dit de l'espérance de vie des dépendants, puisque cet état est irréversible).

Les caractéristiques du contrat de Dépendance moyen réalisé sur 52 contrats dépendance par Profidéo pour les dossiers de l'épargne donne les éléments suivants :

- L'âge moyen minimum d'adhésion est 33 ans pour la dépendance totale contre 30 ans pour la dépendance partielle ;
- L'âge moyen d'adhésion est de 60 ans pour la dépendance totale ainsi que pour la dépendance partielle ;
- L'âge moyen maximum d'adhésion est de 74 ans pour la dépendance totale ainsi que pour la dépendance partielle ;
- Les rentes dépendance moyennes s'étalent de 360 euros à 2500 euros par mois ;
- Les capitaux dépendance moyen vont de 6 000 euros à 100 000 euros ;
- Le délai de carence est de 9 mois pour les maladies physiques et de 33 mois en moyenne pour les maladies psychiques (ceci pour lutter contre le phénomène d'anti-selection) ;
- La sélection médicale est variable d'un contrat à l'autre ;
- La plupart des contrats prévoient une franchise, de l'ordre de 90 jours ;

Les produits peuvent par ailleurs proposer des garanties supplémentaires comme un capital fracture, un capital décès, un capital Alzheimer ou encore différents types d'assistances particulières (télé assistance, aide ménagère, formation de l'aidant, répit de l'aidant...).
Le tableau suivant donne un exemple des tarifs de contrats dépendance que l'on peut trouver sur la place pour 4 assureurs différents :

Assureur A	Dépendance totale + Assistance		Dépendance totale + partielle + Assistance	
Montant de la rente (mensuelle)	500 €	1 500 €	500 €	1 500 €
Cotisation assuré de 60 ans (annuelle)	240 €	680 €	305 €	885 €
Cotisation assuré de 70 ans (annuelle)	395 €	1 150 €	510 €	1 500 €
Assureur B	**Dépendance totale + Assistance + Capital equipement (3 200€)**		**Dépendance totale + partielle + Assistance + Capital equipement (3 200€)**	
Montant de la rente (mensuelle)	500 €	1 500 €	500 €	1 500 €
Cotisation assuré de 60 ans (annuelle)	285 €	750 €	510 €	1 360 €
Cotisation assuré de 70 ans (annuelle)	465 €	1 225 €	900 €	2 400 €
Assureur C	**Dépendance totale + Assistance + Capital décès (3 000€)**		**Dépendance totale + partielle + Assistance + Capital décès (3 000€) + Capital equipement (7 500€)**	
Montant de la rente (mensuelle)	500 €	1 500 €	500 €	1 500 €
Cotisation assuré de 60 ans (annuelle)	510 €	1 150 €	860 €	885 €
Cotisation assuré de 70 ans (annuelle)	750 €	1 700 €	1 300 €	2 610 €
Assureur D	**Dépendance totale + Assistance + Capital fracture (500€)**		**Dépendance totale + partielle + Assistance + Capital fracture (500€)**	
Montant du Capital (unique)	10 000 €		100 000 €	
Cotisation assuré de 60 ans (annuelle)	380 €		400 €	
Cotisation assuré de 70 ans (annuelle)	565 €		585 €	

TABLE 1.1 – Comparatif de produits dépendance privés en France

On remarque que les niveaux des tarifs varient sensiblement suivant le niveau de garanties souhaitées ainsi que de l'âge à l'adhésion.

1.2 Les données

Les bases de données disponibles pour ce mémoire proviennent du Programme de médicalisation des systèmes d'information. La richesse et la forte volumétrie de ces données ont permis d'apporter une approche inédite pour modéliser les phénomènes d'entrée en démence et en dépendance en prenant en compte l'influence de nombreux facteurs de risque et d'informations médicales. Ces données ont été manipulées en accord avec le caractère confidentiel que celles-ci revêtent et en conformité avec les dispositions de la CNIL et du RGPD. Dans la partie précédente, nous avons vu que le diagnostic médical ne permettait de différencier que les états "démence", "dépendance" et "sain" sans atteindre la granularité "dépendance totale" et "dépendance partielle" et encore moins les 6 niveaux de contraste de la grille AGGIR. Toute dépendance identifiée dans la base du PMSI est assimilée à une dépendance totale.

1.2.1 Le PMSI

Le Programme de médicalisation des systèmes d'information (PMSI) est un dispositif qui permet de disposer d'informations médicales quantifiées et standardisées. Mis en place en 1982 par Jean de Kervasdoué, responsable de la Direction des Hôpitaux, son objectif est de définir l'activité des établissements et de calculer l'allocation budgétaire qui en découle. Pour chaque séjour d'un patient hospitalisé il est réalisé un résumé de sortie standardisé (RSS). Ce RSS est réalisé le plus tôt possible après la sortie du patient.

Il contient obligatoirement un diagnostic principal, qui est le problème de santé qui a motivé l'admission du patient dans l'unité médicale (UM), déterminé à la sortie de l'UM.

Le RSS peut aussi contenir un « diagnostic relié » (au motif du séjour), et des « diagnostics associés » (« significatifs » s'ils ont consommé des ressources, « documentaires » dans le cas contraire). Son rôle est d'améliorer la précision documentaire du codage en indiquant la pathologie à l'origine du motif de prise en charge.

La base disponible pour l'étude contient ainsi l'ensemble des hospitalisations des individus hospitalisés au moins une fois de 2008 à 2014. Au sein de celle-ci, on retrouve la base exhaustive de tous les cas nécessitant une hospitalisation (diagnostic principal), les enregistrements de multiples informations médicales au cours des hospitalisations quel qu'en soit le motif (diagnostics associés) ainsi que le suivi exhaustif de la 1^{re} hospitalisation à la dernière hospitalisation en 2008-2013 (dont le décès à l'hôpital).

Les diagnostics sont codés d'après la CIM-10 (Classification internationale des maladies et recours aux services de santé n^o10) éditée par l'OMS et faisant l'objet d'extensions régulières par le ministère de la santé français.
La liste contient 14 400 codes différents et permet de coder de nombreux diagnostics et situations cliniques ou sociales. Utilisant des sous-classifications facultatives, le nombre de codes peut s'étendre jusqu'à 16 000.

Les données du PMSI restent anonymes. En effet, un algorithme de hachage transforme les chiffres permettant l'identification du patient en une chaîne de caractères, chiffres et lettres, par une fonction de calcul complexe et irréversible.
La fonction de transcription est par ailleurs bijective ce qui implique que pour un patient donné, il existe une et une seule image par cette fonction. L'algorithme attribue alors la même clé d'identification pour deux observations d'un même patient ce qui permet d'effectuer des analyses statistiques sur la population anonymisée. Cet algorithme de hachage a été validé par la CNIL en 2001.

Pour cette étude et au vu de la très forte volumétrie des données disponibles sur la France entière, nous nous sommes limités aux périmètres des hommes picards et des femmes parisiennes. Ces deux populations ont été choisies du fait de leur forte disparité. Les femmes vivant plus longtemps que les hommes et le niveau de vie et la facilité de l'accès aux soins étant sensiblement différents entre Paris et la Picardie.

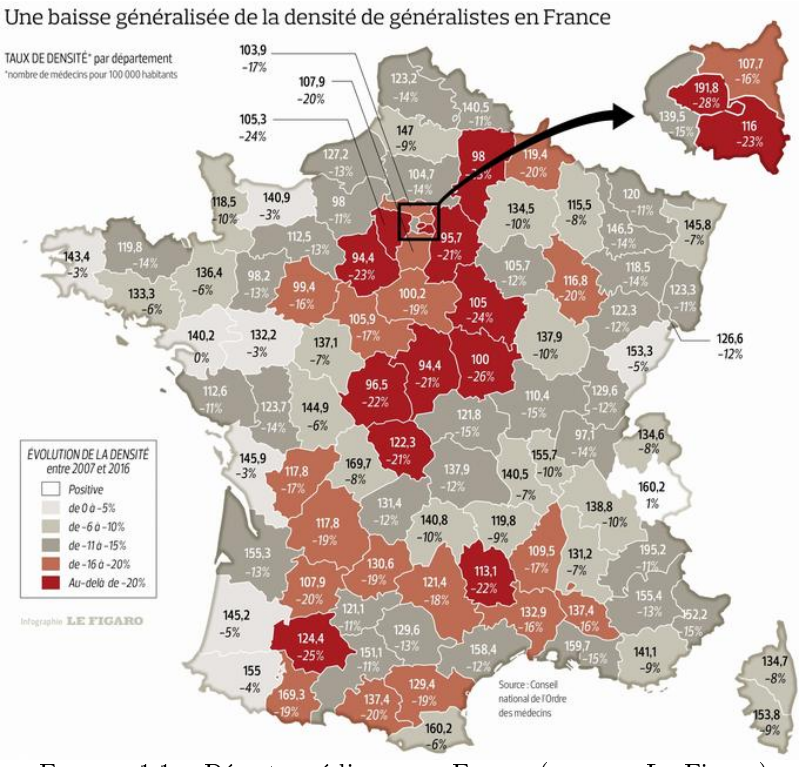


FIGURE 1.1 – Déserts médicaux en France (source : Le Figaro)

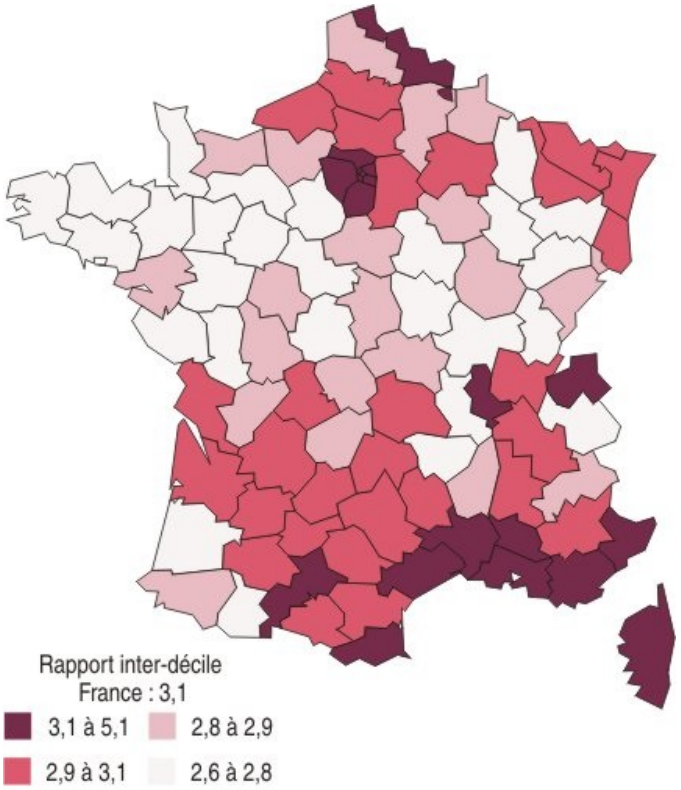


FIGURE 1.2 – Niveaux de vie en France (source : INSEE)

1.2.2 La période d'observation : censure et troncature

Censure et troncature

Les données du PMSI utilisées pour cette étude concernent la période 2008-2013. Nous disposons ainsi des Résumés des Sorties Standardisés de tous les hommes hospitalisés en Picardie et de toutes les femmes hospitalisées à Paris pendant ces 4 années.

Parmi eux, certains sont entrés dans l'état de démence ou de dépendance avant le 01/01/2008. On parle de données censurées à gauche. Celles-ci ne nous intéresse pas pour mesurer l'incidence des phénomène de perte d'autonomie. Nous avons donc exclu ces patients du champ d'étude. Plus largement, tous les patients ayant contracté une maladie grave avant le 01/01/2008 ont été retirés du champ d'étude. On part donc du principe qu'au début de la période d'observation le 01/01/2008, l'ensemble des patients sont considérés comme sains.

Par ailleurs, certains patients sont entrés à l'hôpital après 2008 et y sont restés jusqu'à fin 2013 sans entrer pour autant dans l'état de dépendance ou de démence. Ces données sont des données censurées à droite. Celles-ci sont gardées dans le champ de l'étude car la non survenance de l'événement considéré apporte des informations au modèle.

Enfin il convient de souligner que les patients entrés à l'hôpital après 2008 et qui sont décédés avant la fin 2013 (à l'hôpital ou en dehors) sans être passés par l'état de dépendance physique ou cognitive sont également des données censurées, on parle cette fois de censure droite aléatoire. En effet on ne sait pas si la dépendance serait survenue et le cas échéant, quand elle serait survenue, si le patient n'était pas décédé. On sait tout de même que, avant le décès, la survenance de l'événement à modéliser (dépendance ou démence) n'avait pas encore été constatée.

Dans ces travaux, les sorties sont modélisées au moyen d'un modèle de Cox. Ce modèle, qui sera décrit plus en détail par la suite, a pour objectif de mesurer les écarts entre différentes sous-populations, sous réserve de res-

pecter certaines hypothèses dont l'indépendance des covariables avec le temps et l'hypothèse de proportionnalité des risques. L'avantage de ce modèle est qu'il respecte le critère informatif de la censure. Cependant, dans un soucis de simplification, la censure aléatoire induite par le décès sera traitée comme une censure simple ce qui induit un biais de modélisation et affecte l'estimation des écarts relatifs entre les différentes sous-populations étudiées. L'étude et l'introduction de concurrences entre les différentes causes de sortie du périmètre d'étude permet de lever ce biais, on parle de modèles à risques concurrents. Ce mémoire, dont l'accent est mis sur la comparaison avec des méthodes d'apprentissage, se limite à l'utilisation de modèles non concurrents.

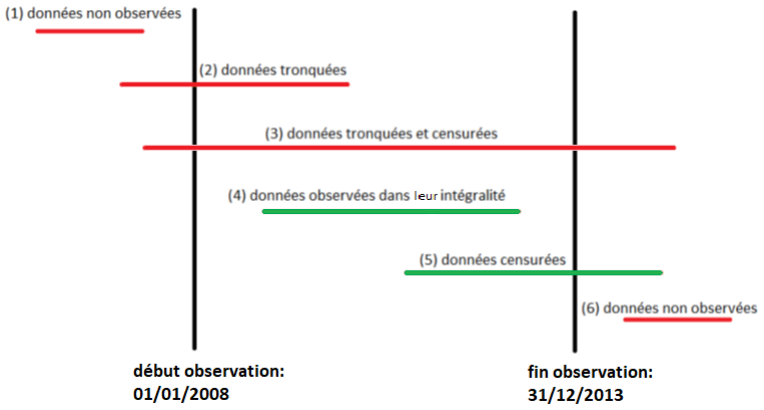


FIGURE 1.3 – Données tronquées et censurées

1.2.3 Les données de l'étude

Étant donné le nombre très important de codes disponibles dans les données du PMSI, nous nous sommes appuyés pour cette étude sur les travaux de regroupements opérés dans les travaux SCHWARZINGER [2018].

Regroupement de pathologie

Les codes du PMSI ont été regroupés en 14 groupes de pathologies :

- BLOOD : maladies sanguines ;
- CANCER : cancers ;
- STROKE : AVC ;
- CARDIO : maladies cardiaques ;
- RESP : maladies respiratoires ;
- KiDNEY : maladies rénales ;
- DIG : maladies digestives ;
- ENDOC : maladies endocriniennes ;
- GYNECO : maladies gynécologiques ;
- INFECT : septicémie ;
- NEURO : maladies neurologiques ;
- RHEUM : maladies rhumatologique ;
- TRAUMA : traumatismes ;
- SENSE : cataracte.

Pour chaque groupe, les pathologies sont divisés en "PATHO0" et en "PATHO1" :

Les "PATHO0" sont des pathologies avérées (que l'on a groupées avec d'autres), qui ont été effectivement diagnostiquées.

Les PATHO1 correspondent quant à elles à des observations d'un ou plusieurs signes cliniques que l'on peut rattacher à un groupe de pathologies et qui témoignent d'un état de santé dégradé. A noter qu'une "PATHO1" est nécessairement précédée d'une "PATHO0"

Les covariables utilisées

Le tableau suivant donne la liste exhaustive des différentes pathologies présentes dans la base (hors facteurs de risque), leur codage dans la base et des statistiques élémentaires. Les pathologies a priori responsables de la démence recensées au chapitre 1 apparaissent sur fond orangé.

Pathologies	Codage dans la base	Nombre	(% du total)	Nombre	(% du total)	Minimum	Moyen	Maximum	Minimum	Moyen	Maximum
Toutes causes		**171 694**	**100%**	**1 177 120**	**100%**	**44,0**	**68,0**	**110,0**	**44,0**	**64,4**	**110,3**
Anémie de type carence martiale	BLOOD0_2ANEMIA_IRON	5 736	3,3%	34 382	2,92%	44,0	73,4	109,0	45,7	73,4	101,7
Anémie de type hémolyse	BLOOD0_2ANEMIA_LYSE	5 518	3,2%	40 402	3,43%	44,5	74,7	108,9	44,7	74,7	103,2
Trouble hémato autre qu'anémie	BLOOD0_2OTHER	1 709	1,0%	19 951	1,69%	44,3	70,9	106,0	45,0	70,9	99,1
Anémie: Transfusion sanguine	BLOOD1_1TRANSFUSION	5 041	2,9%	13 232	1,12%	46,1	75,7	109,0	46,2	75,7	95,1
Cancer, antécédant familial	CANCER0_ATCD_FAM	9 197	5,4%	44 104	3,75%	44,0	56,6	107,9	44,0	56,6	94,0
Tumeur bénigne	CANCER0_BENIN	22 000	12,8%	157 172	13,35%	44,1	61,1	109,0	44,2	61,1	103,3
Cancer in situ (dépistage)	CANCER0_INSITU	4 815	2,8%	38 496	3,27%	44,1	65,5	108,0	44,7	65,5	101,1
Cancer cutané (hors mélanome)	CANCER0_SKIN	2 685	1,6%	17 585	1,49%	44,6	72,9	109,0	44,8	72,9	101,1
Cancer ORL œsophage	CANCER1_AUD_SMOKER	393	0,2%	576	0,05%	46,4	68,1	103,0	46,9	68,1	96,9
Cancer sein	CANCER1_BREAST	5 200	3,0%	15 385	1,31%	46,1	65,9	103,0	46,6	65,9	99,0
Cancer colorectal	CANCER1_COLORECTAL	1 603	0,9%	12 471	1,06%	46,3	72,3	104,9	46,1	72,3	100,3
Hémopathie (lymphome)	CANCER1_HEMATO	1 528	0,9%	7 096	0,60%	46,4	74,7	109,0	46,9	74,7	96,2
Cancer de bon pronostic	CANCER1_PC_GOOD	1 465	0,9%	15 988	1,36%	46,2	68,5	106,4	46,5	68,5	100,3
Cancer de mauvais pronostic	CANCER1_PC_POOR	1 963	1,1%	38 558	3,28%	46,3	70,8	104,0	47,2	70,8	101,0
Cancer poumon	CANCER1_SMOKER	1 338	0,8%	30 989	2,63%	46,1	69,8	107,8	46,4	69,8	99,4
Hypertension artérielle	CARDIO0_HBP	43 717	25,5%	391 260	33,24%	44,1	72,1	109,0	44,1	72,1	110,0
Ischemic Heart Disease: angine de poitrine traitée	CARDIO0_IHD_2TTT	1 365	0,8%	34 321	2,92%	44,0	70,5	98,4	44,0	70,5	90,5
Ischemic Heart Disease: angine de poitrine non traitée	CARDIO0_IHD_3noTTT	4 880	2,8%	63 938	5,43%	44,2	71,5	105,7	44,1	71,5	100,0
Maladie cardio vasculaire autre	CARDIO0_OTHER	10 209	5,9%	76 272	6,48%	44,0	64,2	107,0	44,1	64,2	102,3
Tr. Rythme: autre arythmie non traitée	CARDIO0_RYTHME_2TTT	1 680	1,0%	16 955	1,44%	44,1	77,1	104,0	44,3	77,1	102,0
Tr. Rythme: autre arythmie traitée	CARDIO0_RYTHME_3noTT	3 174	1,8%	34 430	2,92%	44,2	72,5	107,1	45,0	72,5	101,0
Valvulopathie (dont endocardite)	CARDIO0_VALVE_ANY	2 485	1,4%	17 291	1,47%	44,1	73,7	107,1	44,7	73,7	100,0
Ischemic Heart Disease: Infarctus	CARDIO1_IHD_1MI	1 001	0,6%	19 889	1,69%	46,1	75,4	104,2	46,7	75,4	110,0
Insuffisance cardiaque (dont arrêt cardiaque)	CARDIO1_INSUF_CHRO	9 607	5,6%	99 095	8,42%	46,2	79,5	109,0	46,7	79,5	103,2
Tr. Rythme: aryhmie complète par fibrillation auriculaire	CARDIO1_RYTHME_1ACFA	6 810	4,0%	74 455	6,33%	46,4	80,1	107,0	46,3	80,1	104,9
Maladie thrombo-embolique	CV1_MTE	3 306	1,9%	24 510	2,08%	46,1	75,0	106,0	46,1	75,0	103,0
Maladie artérielle périphérique (aorte, digestif, rein, amputation)	CV1_PVD	4 530	2,6%	66 557	5,65%	46,3	74,1	107,0	46,3	74,1	110,3
Cirrhose compensée	DIG0_LIVER_2CirrC	1 396	0,8%	24 147	2,05%	44,5	67,5	104,0	44,5	67,5	98,8
Cirrhose étiologie identifiée autre qu'alcool	DIG0_LIVER_ETIO_ANY	1 729	1,0%	12 217	1,04%	44,0	59,9	101,9	44,0	59,9	93,7
Maladie digestive (autre que pancréatite, cholécystite)	DIG0_OTHER_noFDR	53 194	31,0%	380 092	32,29%	44,0	63,5	110,0	44,1	63,5	109,6
Pancréatite	DIG0_PANCREAS_AUD	998	0,6%	12 759	1,08%	44,4	67,3	105,0	44,5	67,3	104,0
Cholécystite (voies billaires)	DIG0_VB_METABO	6 302	3,7%	36 899	3,13%	44,0	65,6	107,0	44,1	65,6	104,0
Complication digestive: hémorragie (toute cause)	DIG1_HEMORRAGIE	4 231	2,5%	38 750	3,29%	46,3	67,5	105,4	46,7	67,5	110,1
Cirrhose décompensée	DIG1_LIVER_1CirrD	792	0,5%	14 514	1,23%	46,3	69,4	101,0	46,5	69,4	98,8
Complication digestive: occlusion (toute cause)	DIG1_OCCLUSION	2 877	1,7%	22 350	1,90%	46,3	76,9	110,0	46,7	76,9	104,9
Complication digestive: péritonite (toute cause)	DIG1_PERITONITE	1 235	0,7%	10 565	0,90%	46,0	69,1	103,1	46,4	69,1	100,1
Complication digestive: stomie (toute cause)	DIG1_STOMIE	768	0,4%	11 305	0,96%	46,3	71,5	109,0	46,7	71,5	97,7
Carence nutritionnelle (autre que dénutrition)	ENDOC0_CARENCE	6 016	3,5%	31 153	2,65%	44,0	77,5	108,9	45,2	77,5	109,6
Diabète	ENDOC0_DIABETE	14 108	8,2%	153 744	13,06%	44,0	68,1	109,0	44,1	68,1	100,0
Dyslipidémue (hypercholétérolémie)	ENDOC0_DYSLIPIDEMIA	17 395	10,1%	196 436	16,69%	44,1	69,6	105,3	44,1	69,6	110,0
Maladie thyroïdienne	ENDOC0_THYROIDE	11 449	6,7%	24 860	2,11%	44,2	68,3	109,0	45,0	68,3	104,1
Maladie endocrinienne (autre que thyroïde)	ENDOC1_GLD_OTHER	1 810	1,1%	10 661	0,91%	46,2	70,1	105,0	46,8	70,1	100,0
Maladie métabolique (autre diabète, dyslipidémie)	ENDOC1_METABO	1 984	1,2%	17 188	1,46%	46,5	70,8	108,2	46,8	70,8	100,1
Maladie gynécologique	GYNECO0_ALL	10 118	5,9%	13 307	1,13%	44,0	58,9	103,0	44,2	58,9	100,0
Sépticémie (toute cause)	INFECT1_SEPSIS	2 880	1,7%	32 860	2,79%	46,4	75,1	108,2	46,7	75,1	101,0
Maladie rénale chronique : Glomérulonéphrite	KIDNEY0_2GN	1 597	0,9%	53 744	4,57%	44,7	74,1	104,8	44,1	74,1	102,0
Maladie rénale chronique: Autre	KIDNEY0_2OTHER	909	0,5%	13 109	1,11%	44,6	68,2	105,1	45,0	68,2	101,0
Maladie rénale chronique: Pyélonéphrite et tubulopathie	KIDNEY0_2PYELONEPHRI	3 166	1,8%	35 644	3,03%	44,0	73,9	106,0	44,7	73,9	98,8
Maladie rénale chronique: Lythiase urinaire	KIDNEY0_2UROLITHIASE	1 712	1,0%	22 960	1,95%	44,1	61,6	103,3	45,0	61,6	105,1
Cystite avant maladie rénale chronique	KIDNEY0_CYSTITE_ATCD	5 612	3,3%	17 435	1,48%	44,2	77,4	109,0	45,0	77,4	99,1
Insuffisance rénale chronique	KIDNEY1_1INSUF_CHRO	4 286	2,5%	36 761	3,12%	46,1	79,8	109,0	46,8	79,8	103,2
Insuffisance rénale aiguë	KIDNEY1_2INSUF_AIGUE	2 424	1,4%	28 356	2,41%	46,3	78,8	108,9	46,7	78,8	102,3
Atteinte du système nerveux périphérique	NEURO0_SNP_METABO	1 814	1,1%	19 951	1,69%	44,3	67,8	102,0	44,4	67,8	98,8
Maladies rares à risque de démence (sclérose en plaque, hydrocéphalie à pression normale, encéphalité)	NEURO1_DEM_FDR	628	0,4%	3 764	0,32%	46,2	65,7	102,0	46,6	65,7	95,1
Epilepsie (et autre convulsion)	NEURO1_EPILEPSIE	1 516	0,9%	16 098	1,37%	46,3	72,8	108,2	46,4	72,8	100,0
Maladie neurologique autre	NEURO1_OTHER	679	0,4%	6 287	0,53%	46,3	68,8	107,0	46,6	68,8	96,0
Maladie de Parkinson (et autre sd. Extrapyramidal)	NEURO1_PARKINSON	1 295	0,8%	13 136	1,12%	46,4	77,4	102,9	46,7	77,4	100,0
Maladie resp. chronique: Asthme	RESP0_2ASTHMA	4 954	2,9%	71 637	6,09%	44,0	66,2	104,0	44,2	66,2	104,0
Maladie resp. chronique: Broncho-pneumonie chronique obstructive	RESP0_2COPD	3 317	1,9%	5 752	0,49%	44,7	70,4	109,0	46,0	70,4	98,2
Maladie resp. chronique: Maladie interstitielle	RESP0_2INTERSTI	433	0,3%	38 448	3,27%	45,0	73,3	105,7	44,7	73,3	102,0
Maladie resp. chronique: Autre	RESP0_2OTHER	3 684	2,1%	46 997	3,99%	44,0	66,5	107,0	44,2	66,5	94,4
Apnée du sommeil	RESP0_APNEE_SOM	3 085	1,8%	8 707	0,74%	44,1	62,9	109,0	44,3	62,9	102,1
Bronchite avant maladie resp. chronique	RESP0_BRONCHITE_ATCD	654	0,4%	48 986	4,16%	44,1	70,6	107,0	44,2	70,6	103,3
Infection respiratoire basse avant maladie resp. chronique	RESP0_LRI_ATCD	5 028	2,9%	38 955	3,31%	44,2	77,9	107,0	44,6	77,9	98,9
Insuffisance résporatoire chronique (dont arrêt respiratoire)	RESP1_1INSUF_CHRO	1 790	1,0%	20 129	1,71%	46,1	74,5	107,0	47,0	74,5	100,2
Insuffisance résporatoire aiguë	RESP1_2INSUF_AIGUE	2 836	1,7%	29 803	2,53%	46,2	76,4	107,1	46,4	76,4	103,0
Arthrose hanche (prothèse)	RHEUM0_ARTHROSE_HIP	5 055	2,9%	29 426	2,50%	44,5	72,0	106,9	44,7	72,0	97,9
Arthrose genou (prothèse)	RHEUM0_ARTHROSE_KNEE	5 717	3,3%	30 612	2,60%	44,0	70,5	104,3	44,2	70,5	100,0
Arthrose autre (vertébrale, multiple…)	RHEUM0_ARTHROSE_OTHE	4 973	2,9%	59 770	5,08%	44,0	72,5	107,0	44,2	72,5	101,0
Maladie thumatologique autre (alcool)	RHEUM0_AUD	5 496	3,2%	12 238	1,04%	44,1	71,6	109,0	44,1	71,6	98,9
Maladie de système (polyarthrite rhumatoïde, lupus…)	RHEUM0_SYSTEME	3 331	1,9%	125 587	10,67%	44,0	65,9	104,0	44,4	65,9	101,1
Cataracte	SENSE0_CATARACTE	2 574	1,5%	5 183	0,44%	44,2	73,7	109,0	46,0	73,7	94,2
AVC: maladie cérébrovasculaire traitée	STROKE0_2TTT	308	0,2%	10 805	0,92%	44,2	70,1	99,9	45,0	70,1	101,0
AVC: maladie cérébrovasculaire non traitée	STROKE0_3noTTT	1 253	0,7%	15 186	1,29%	44,1	72,3	104,2	44,0	72,3	101,0
Antécédent d'accident ischémique transitoire ou AVC sans séquelle	STROKE0_AIT_noSEQ	1 689	1,0%	53 716	4,56%	44,0	72,6	103,2	44,3	72,6	109,0
AVC: hémorragique (grave)	STROKE1_1HEMO	936	0,5%	8 021	0,68%	46,3	75,6	105,0	47,4	75,6	96,7
AVC: ischémique (moins grave)	STROKE1_1ISCHEMIC	2 447	1,4%	23 146	1,97%	46,5	78,3	107,0	46,6	78,3	100,2
Trauma: Fracture (hors crane, sévère)	TRAUMA0_3FRACTURE	14 346	8,4%	30 577	2,60%	44,0	73,9	109,0	44,5	73,9	102,8
Chute	TRAUMA0_CHUTE	2 654	1,5%	5 101	0,43%	45,0	79,2	109,0	46,0	79,2	97,9
Ostéoporose	TRAUMA0_OSTEOPOROSE	4 420	2,6%	10 462	0,89%	44,3	78,1	106,2	44,1	78,1	99,1
Suicide	TRAUMA0_SUICIDE	1 301	0,8%	47 230	4,01%	44,1	58,2	99,7	46,1	58,2	103,2
Trauma: crâne	TRAUMA1_1CRANE	1 431	0,8%	14 864	1,26%	46,2	77,9	105,1	46,9	77,9	101,0
Trauma: sévère (hors crâne)	TRAUMA1_2SEVERE	613	0,4%	7 562	0,64%	46,2	69,3	107,8	46,9	69,3	101,0

TABLE 1.2 – Statistiques descriptives des différentes pathologies

L'objet de ces statistiques descriptives élémentaires est de décrire la base de données. Les taux exhibés ici pour chaque pathologie sont calculés comme le simple rapport entre le nombre de cas recensés par le nombre

d'observations totales. Le caractère censuré des données n'est pas pris en compte. Un patient avec une exposition d'un jour est décompté au même titre qu'un autre avec une exposition de plusieurs années. Ces taux n'ont absolument pas à être considérés comme des incidences.
Parmi les facteurs de risque, les taux d'alcooliques, de fumeurs et d'obèses sont les suivants :

Code couleur:	Cause de démence identifiée au chapitre 1	Nombre d'observations			
		Parisiennes		*Picards*	
Facteur de risque	Codage dans la base	Nombre	(% du total)	Nombre	(% du total)
Total		**1 177 120**	**100%**	**171 694**	**100%**
Alcoolique	fdr_aud_all	15 845	1,3%	10 791	6,3%
Fumeur	fdr_smoker	29 773	2,5%	16 896	9,8%
Obèse	fdr_obesity_all	79 934	6,8%	15 214	8,9%

TABLE 1.3 – Statistiques descriptives des différents facteurs de risque

Dans la base initiale, nous avons une ligne pour chaque événement médical. Un même patient a donc autant de lignes que d'événements médicaux constatés à l'hôpital durant la période d'observation (pas nécessairement durant le même séjour).
Puisque l'on s'intéresse à la dépendance (respectivement la démence), on a créé une base avec une ligne unique pour un patient et on a crée une indicatrice de censure qui vaut 1 lorsque le patient entre en dépendance (respectivement en démence) et 0 sinon (le patient peut sortir de l'hôpital et être perdu de vue, il peut décéder à l'hôpital ou il peut être à l'hôpital au 31/12/2013).

Nous avons créé par ailleurs une variable *time* qui donne l'âge du patient constaté au moment de l'événement médical.

Par ailleurs, pour chaque covariable pathologique, la valeur affectée à la variable correspond à la durée passée dans la pathologie, ceci pour effectuer une pondération des effets de chaque pathologie en fonction de l'importance de la durée passée dans l'état.

Chaque facteur de risque (alcool,tabac,obésité) est représenté par une indicatrice (1 : présence ; 0 : absence).

Enfin, nous avons ajouté des proxys croisés avec des données INSEE à partir de la variable département, sexe et âge pour avoir une mesure de l'influence des deux facteurs sociaux suivants :

- *cp_immi* : variable informant sur le taux d'immigration (0=1^{er} quartile,1=2^{me} quartile,2=3^{me} quartile,3=4^{me} quartile) ;
- *cp_dipl*0 : variable informant sur le taux d'individus diplômés de niveau inférieur (BEPC max) diplômés du BAC (0=1^{er} quartile,1=2^{me} quartile,2=3^{me} quartile,3=4^{me} quartile) ;

La covariable CANCER1_PROSTATE n'apparait pas pour la population parisienne car c'est une pathologie exclusivement masculine et que la pathologie GYNECO0_ALL n'apparait pas pour la population picarde car c'est une pathologie uniquement féminine.

Quelques résultats statistiques

Dans un but descriptif, on trace ci-dessous quelques fonctions de survie relatives aux pathologies dont l'influence sur la dépendance a été identifiée comme importante (au regard de l'étude préliminaire fondée sur des travaux médicaux. Celles-ci sont estimées par l'estimateur de survie de Harrington-Flemming suivant :

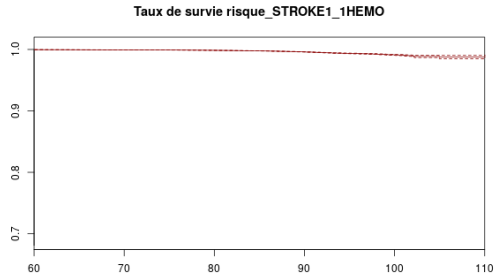
$$\hat{S}(t_i) = \exp\Big(-\sum_{j=1}^{i} \frac{d_j}{Y_j}\Big)$$

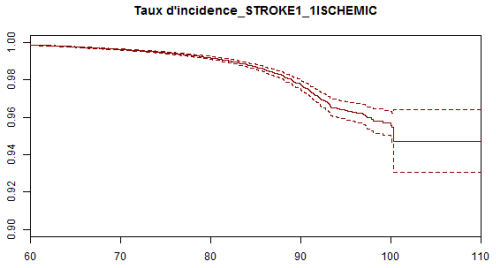
où $t_1 < t_2 < \cdots < t_D$ représentent les dates distinctes d'événements.
Pour chaque $i = 1, \ldots, D$, Y_i représente le nombre de patients non encore atteints juste avant t_i et d_i le nombre d'entrées dans la pathologie à t_i.

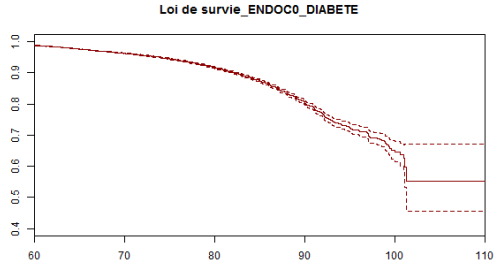
L'avantage de cet estimateur est qu'il est compatible avec la censure à droite, en effet le nombre d'individu à risque Y_i évolue entre chaque date d'événements, ce qui permet d'estimer la fonction de survie sur une somme de petits intervalles de temps complets (au même titre que la méthode de Kaplan-Meier).

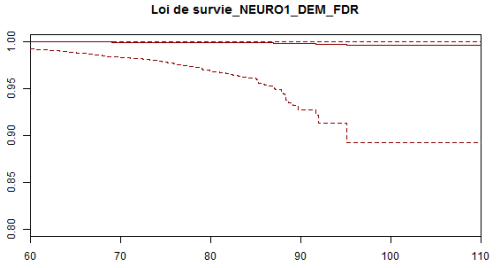
Nous avons choisi de présenter les statistiques de chaque pathologie à l'aide de cet estimateur car il sera réutilisé par la suite pour estimer la fonction de hasard de base dans le modèle de Cox pour estimer les risques démence et dépendance. En effet cet estimateur et sa forme exponentielle sont compatibles avec le modèle de Cox qui est multiplicatif et s'exprime également sous forme exponentielle.

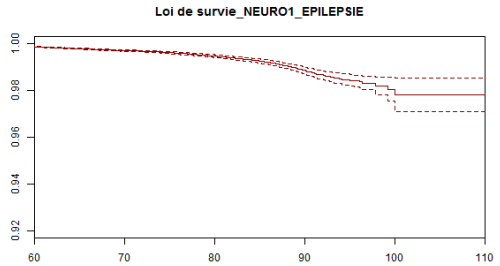
Dans les graphes suivants, tous les individus sont sains au départ, la probabilité de ne pas avoir contracté la maladie est alors égale à 1. Elle décroit ensuite au rythme de la probabilité de contracter la maladie considérée.
L'échelle de l'axe des ordonnées (probabilité d'être sains) est différentes entres les graphes pour améliorer la lisibilité des courbes.

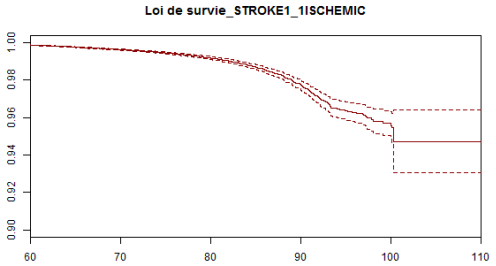


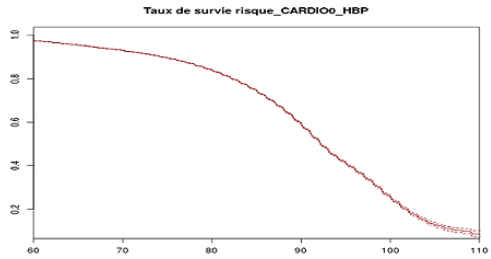












La pathologie la plus représentée, parmi celles identifiées en préambule comme à risque par la profession médicale, est l'hypertension artérielle [1]. Celle-ci atteint une probabilité de survenance de 50% pour les âges les plus avancés (supérieurs à 90 ans). Les autres pathologies sont plus rares (avec des taux d'incidence n'excédant pas plus de 5% même pour les âges les plus avancés)

1. variable CARDIO_HBP

Chapitre 2

Modélisation par un modèle de Cox

2.1 Éléments théoriques - Le modèle de Cox

La régression de Cox est un modèle de régression à risque proportionnel. Elle doit son nom au statisticien britannique David Cox. La régression de Cox appartient à la classe des modèles de survie dont la fonction est d'étudier le temps écoulé avant qu'un événement ne se produise. Historiquement, l'évènement modélisé par la régression de Cox était le décès de l'individu. C'est pour cette raison que l'on parle de « modèles de survie » pour caractériser ces modèles. Cependant, l'utilisation du modèle s'est étendue à tout type d'évènements comme, dans le cadre de notre étude, l'entrée dans l'état de dépendance physique ou cognitive.

Le premier obstacle rencontré dans la modélisation d'une durée réside dans la nature des données utilisées pour la mener. En effet, il n'est pas envisageable de suivre les individus sur une période telle que la vie humaine. De ce fait, les observations pour lesquelles l'évènement que l'on cherche à modéliser ne s'est pas produit, ne fournissent au modèle qu'une information partielle. Nous ne pouvons en effet pas affirmer si l'individu entrera dans l'état de dépendance ou non une fois sortie du périmètre d'étude. Et si tel est le cas, au bout de combien de temps il entrera dans l'état. On parle ainsi, comme nous l'avons abordé dans la partie précédente de données tronquées ou censurées.

L'avantage du modèle de Cox est qu'il permet de prendre en compte ce type de données même si elles ne sont pas « complètes ». Notre étude portant sur un périmètre d'observation de 4 ans et celle-ci cherchant à modéliser un phénomène dont l'incidence peut intervenir à n'importe quel moment de la vie humaine, l'utilisation d'un modèle gérant les données tronquées ou censurées est alors nécessaire. Nous rappellerons comment se calcule la vraisemblance en prenant en compte le phénomène de censure.

2.1.1 Rappels sur le modèle de Cox

La régression de Cox fait partie de la famille des modèles multiplicatifs. En effet, conditionnellement aux caractéristiques de l'individu i donné, la fonction d'incidence de cet individu est construite de la façon suivante :

$$\alpha(t|\mathbf{Z}_i(t)) = \alpha_0(t)exp^{(\mathbf{Z}_i^T\beta)}$$

Avec :

- $\alpha_0(t)$ le risque de base, une fonction non spécifiée (intensité en l'absence d'effet des covariables) ;
- $\mathbf{Z}_i$ le vecteur des covariables pour un individu i ;
- β les mesures d'influence des covariables sur l'intensité ;

On construit ainsi la fonction d'incidence de l'individu i conditionnellement à ses caractéristiques propres. Celles-ci sont modélisées au travers du vecteur de covariables $\mathbf{Z}_i$ qui regroupe l'ensemble des valeurs prises par chacune

des variables explicatives choisie pour l'étude. La forme de la fonction d'incidence conditionnelle ainsi écrite fait alors apparaitre deux facteurs :

Le premier appelé fonction de hasard de base est une fonction uniquement dépendante du temps.

Le second est un facteur ne dépendant quant à lui que de l'individu i. Celui-ci est construit avec une exponentielle dont l'argument s'exprime sous forme vectorielle comme le produit scalaire de la transposée du vecteur des covariables par le vecteur des coefficients à déterminer par la régression. On retrouve dans l'exponentielle la forme classique d'une régression linéaire avec une somme de variables explicatives chacune affectée par son propre coefficient de régression. Le rôle de ces coefficients étant de pondérer l'effet de la variable explicative sur la variable à expliquer.

La fonction d'incidence ainsi défini pour un individu i donné est alors le produit d'un facteur véhiculant l'information liée au vieillissement (ce dernier étant commun à tous les individus) et d'un facteur propre à chaque individu. Ce dernier ayant la particularité de s'exprimer de façon exponentielle, ce qui implique que pour deux individus i et j ayant des vecteurs de covariables $\mathbf{Z}_i$ et $\mathbf{Z}_j$, le rapport de leurs fonctions d'incidence s'écrit :

$$\frac{\alpha(t|\mathbf{Z}_i(t))}{\alpha(t|\mathbf{Z}_j(t))} = \frac{exp^{\mathbf{Z}_i^T\beta}}{exp^{\mathbf{Z}_j^T\beta}} = exp^{(\mathbf{Z}_i-\mathbf{Z}_j)^T\beta}$$

- ne dépend que des individus i et j ;
- est constant au cours du temps.

Ceci confère au modèle la caractéristique de modèle multiplicatif. En effet, l'effet de chaque covariable est caractérisé de façon linéaire dans l'exponentielle. On peut alors sortir chacun des termes sous forme d'un produit d'exponentielles. Par exemple, pour deux individus dont les covariables sont toutes strictement identiques sauf la $k$ieme, on aura :

$$\vec{Z}_i \begin{pmatrix} z_{i,1} \\ z_{i,2} \\ \dots \\ z_{i,k} \\ \dots \\ z_{i,n} \end{pmatrix}$$

,

$$\vec{Z}_j \begin{pmatrix} z_{j,1} = z_{i,1} \\ z_{j,2} = z_{i,2} \\ \dots \\ z_{j,k} \neq z_{i,k} \\ \dots \\ z_{j,n} = z_{i,n} \end{pmatrix}$$

ainsi :

$$\frac{\alpha(t|\mathbf{Z}_i(t))}{\alpha(t|\mathbf{Z}_j(t))} = \frac{exp^{\mathbf{Z}_i^T\beta}}{exp^{\mathbf{Z}_j^T\beta}} = exp^{(\mathbf{Z}_i-\mathbf{Z}_j)^T\beta} = exp^{(z_{i,k}-z_{j,k})\beta}$$

On a constaté également que l'évolution de l'incidence dans le temps est indépendante des caractéristiques des individus. Par exemple, en ne faisant varier que le sexe, le modèle supposerait que toutes choses égales par ailleurs, l'évolution de l'incidence dans le temps serait similaire pour un homme et une femme. Graphiquement ceci se traduirait par des courbes de survie parallèles pour chacun des individus. Cette hypothèse forte est

appelée hypothèse de risques proportionnels.

Le modèle ainsi défini peut alors être estimé à partir des données. Comme dans toute étude de régression, il s'agit alors de trouver le vecteur de coefficients β qui fait correspondre au mieux le modèle avec la réalité. Pour se faire, on rencontre classiquement deux méthodes pour estimer le vecteur des coefficients dans le cadre d'une régression.

La première méthode, dite méthode des moindres carrés ordinaires, cherche à minimiser l'erreur totale résultante de la différence quadratique entre chaque observation de l'échantillon et chaque estimation associée. Le problème, dans le cas présent, est que cette méthode ne peut s'appliquer que pour des modèles paramétriques. En effet, pour pouvoir calculer l'erreur pour une observation donnée il faut pouvoir être en mesure d'évaluer une estimation de celle-ci au moyen de ces paramètres. Or la forme de la fonction d'incidence que l'on cherche à modéliser est le produit d'un facteur qui dépend bien des paramètres du modèle mais d'un autre facteur qu'il n'est pas possible d'estimer au moyen de ces seuls paramètres. On parle dans ce cas de modèle semi-paramétrique.

L'autre méthode classiquement utilisée pour paramétrer un modèle de régression est la méthode du maximum de vraisemblance. Elle a pour vocation initiale de maximiser le produit des densités de la loi que l'on cherche à paramétrer en chaque point observé. En d'autres termes on cherche les paramètres qui maximisent le poids des densités de probabilité des observations les plus fréquentes par rapport à celles moins fréquentes. Ce qui pourrait s'interpréter comme définir le jeu de paramètres pour une loi donnée qui maximiserait les chances de retirer exactement le même échantillon si l'on souhaitait le reproduire aléatoirement avec cette loi.

2.1.2 Calcul de vraisemblance et prise en compte de censure dans les modèles de durée

Nous allons par la suite utiliser la technique de maximum de vraisemblance pour paramétrer nos modèles. Nous rappelons ci-après son expression :
PLANCHET et THEROND [2000] démontrent que la vraisemblance du modèle associé aux observations $(t_1, d_1), ..., (t_n, d_n)$ s'écrit :

$$L(\beta, h_0) = \{\prod_{i=1}^{n} h_0(t_i) e^{\beta^T z_i}\}^{d_i} exp(-\int_0^{T_i} h_0(t) e^{\beta^T z_i} dt)$$

La maximisation de la vraisemblance étant impossible dans le cadre d'un modèle semi-paramétrique, nous rappelons l'expression de la vraisemblance partielle. Celle-ci est obtenue en considérant qu'aucune information ne peut être donnée sur β sur les intervalles pendant lesquels aucun événement n'a eu lieu. Il est alors supposé que h_0 est nulle dans ces intervalles. Il est ainsi supposé que les moments où se produisent les censures n'apportent peu ou pas d'information sur β. On travaille alors conditionnellement à l'ensemble des instants où un décès a lieu.

A l'instant T_i, il y a $R(T_i)$ patients encore à risque, la probabilité d'occurrence de l'événement en T_i de chaque sujet j est donné par la formule :

$$L_{partielle}(\beta) = \prod_{i}^{D} \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)}$$

L'estimateur du maximum de vraissemblance est l'estimateur qui associe aux observations la valeur pour laquelle la probabilité de l'observation est la plus importante. C'est la valeur $\hat{\beta}$ de β qui maximise $L_{partielle}(\beta)$, c'est à dire, telle que :

$$\frac{\partial L_{partielle}(\hat{\beta})}{\partial \hat{\beta}} = 0$$

2.2 Étude pratique - Modélisation des données de l'étude à l'aide d'un modèle de Cox

2.2.1 Validation des hypothèses du modèle

Avant d'interpréter le modèle, il est utile de voir si l'hypothèse d'un modèle à risques proportionnels est vérifiée (forme multiplicative et covariables indépendantes du temps). Cette vérification est à faire sur le modèle global, avant même d'interpréter les tests (qui ne sont pas valables lorsque l'hypothèse n'est pas vérifiée), et avant de sélectionner des variables : il se peut qu'une covariable ait un effet non significatif lorsque cet effet est moyenné dans le temps mais qu'elle ait une interaction significative avec le temps. C'est pourquoi il vaut mieux tester l'hypothèse des risques proportionnels avant d'interpréter la significativité des effets

Test analytique - Hypothèse de nullité des coefficients - Qualité d'ajustement globale du modèle

Un test du χ^2 global est effectué en plus des tests du $\chi 2$ variable par variable. Le test du χ^2 pour la j^{me} covariable consiste à tester l'hypothèse nulle :

$$\text{H0} : \beta^{(j)}(t) = \beta^{(j)} \text{ contre H1} : \beta^{(j)}(t) \neq \beta^{(j)}$$

tandis que le test global du χ^2 consiste à tester l'hypothèse nulle :

$$\text{H0} : \beta(t) = \beta \text{ contre H1} : \beta(t) \neq \beta$$

où $\beta(t) = (\beta^{(1)}(t), ..., \beta^{(p)}(t))'$ représente le vecteur des paramètres que l'on autorise à dépendre du temps t.

Les sorties R présentées en annexe donnent les résultats des tests du $\chi 2$ pour chacune des variables disponibles :

Le test du χ^2 permet de rejeter ou non l'hypothèse H0 en s'intéressant à la probabilité (appelée p-value) que l'hypothèse ne soit pas vérifiée.
La p-value associée à l'hypothèse nulle : $\beta(t) = \beta$ s'élève ici à $4,74.10^{-11}$. Le critère de rejet de l'hypothèse usuellement retenu est une p-value supérieur à 5%. Dans le cas présent, on accepte alors l'hypothèse selon laquelle $\beta(t) = \beta$, cependant, plus le modèle a de variable, plus il est probable que ce score soit très faible. Il est donc difficile de valider l'hypothèse par cette méthode. De plus, on constate que l'hypothèse, selon laquelle les coefficients sont constants au cours du temps, ne se vérifie pas pour chaque variable.

Hypothèses de proportionnalité

Test graphique et analyse des résidus

Pour valider l'hypothèse de proportionnalité, une méthode consiste à comparer les courbes d'incidence constatées sur plusieurs sous-populations et vérifier que celle-ci s'obtiennent au moyen d'une simple translation. La difficulté est de trouver des sous-populations suffisamment volumineuses (pour que les taux bruts constatés sur l'échantillon soient suffisamment stables) et suffisamment hétérogènes en termes de risques pour que les courbes ne se confondent pas. La variable alcool [1] étant a priori significative et la population d'alcooliques étant suffisamment importante (respectivement 15 845 et 10 791 Parisiennes et Picards alcooliques), nous nous intéressons à ces 4 sous-populations pour les deux sorties démence et dépendance physique pour valider l'hypothèse HP :

1. variable fdr aud all

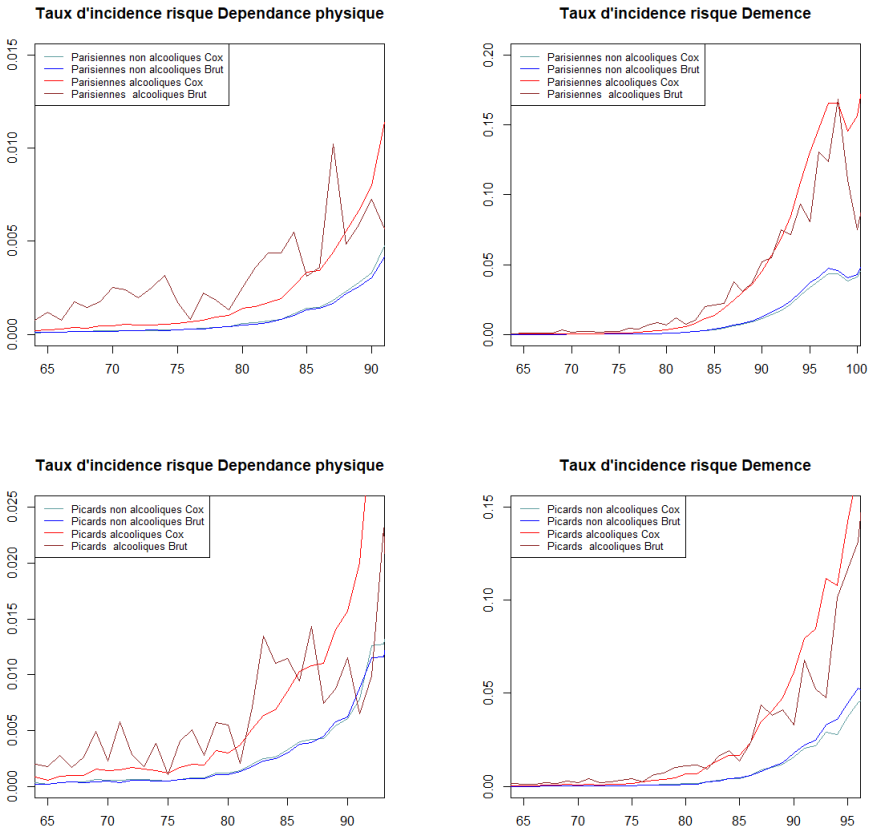


FIGURE 2.1 – Taux bruts et taux prédits par Cox pour différentes sous-populations

Le graphique le plus explicite est celui des Parisiennes pour la démence, on constate que les incidences brutes et les prédictions de Cox sont proches. De plus, les deux sous-populations alcooliques et non alcooliques semblent relativement "parallèles". On peut faire les mêmes constats sur les autres graphiques bien que la forte volatilité des taux bruts des populations alcooliques ainsi que la moins forte hétérogénéité des sous-populations à leurs risques respectifs rendent la lecture moins évidente.
Par ailleurs, de nombreuses procédures de vérification des hypothèses du modèle de Cox sont fondées sur les résidus. Leurs valeurs sont calculées pour chaque individu. Ces résidus ont la particularité d'avoir un comportement connu, du moins approximativement, lorsque les hypothèses du modèle sont remplies.

Les résidus de Schönefeld calculés par le logiciel R se présentent sous la forme d'une matrice à p colonnes qui a autant de lignes qu'il y a d'observations non-censurées dans les données. Les lignes sont ordonnées par durées de vie croissantes. Les résidus de Schönefeld mesurent la distance entre le vecteur covariable des sujets et la moyenne pondérée des vecteurs covariables des sujets à risque. Les résidus de Schönefeld servent à évaluer la tendance au cours du temps et donc à tester l'hypothèse des hasards proportionnels (qui dit que le log-ratio ne doit pas dépendre du temps).

	fdr_aud_all	fdr_smoker	fdr_obesity_all	cp_dep	...	TRAUMA1_2SEVERE
50.0996851471595	0.67412892	-0.25965802	-0.133223971	-0.9873518938	...	-0.0074368740
50.1546680355921	0.67430800	-0.25951043	-0.133308330	-0.9874793118	...	-0.0074504566
50.1695482546201	0.67440986	0.74059358	-0.133305717	-0.9876080851	...	-0.0074541441
50.6562970568104	-0.32395972	-0.25752964	-0.132511402	1.0129335420	...	-0.0074920442
51.1418412046543	0.67517065	0.74239482	-0.133882316	0.0114244182	...	-0.0076578608
51.2742847364819	0.67602674	0.74318546	-0.134037622	0.0115621570	...	-0.0076602097
51.2921423682409	-0.32378916	-0.25667078	-0.134012283	-0.9883051304	...	-0.0076558288
51.3962491444216	0.67695482	-0.25587026	-0.133837794	-0.9887996673	...	-0.0076585298
51.6403080082136	0.67768830	0.74490707	-0.133837592	-0.9888140727	...	-0.0076515766
52.2362491444216	0.67906404	0.74667540	-0.134690992	0.0086166277	...	-0.0066800445
52.2696098562628	0.67942295	0.74700366	-0.134664450	1.0085321203	...	-0.0066453258
52.5748254620123	0.68134150	0.74918080	-0.134626435	1.0098158298	...	-0.0065278619
52.6313689253936	0.68139467	0.74929848	-0.134625721	1.0098501789	...	-0.0065280001
52.6558590006845	-0.31840253	-0.25043202	-0.134661593	0.0100308029	...	-0.0065321404
52.7459137577002	0.68213480	0.75023675	-0.134511662	0.0097380797	...	-0.0065133532
52.8627515400411	0.68256748	-0.24938262	-0.134319409	0.0095889845	...	-0.0064002114
52.9020876112252	0.68291935	0.75096793	-0.134385774	-0.9905104751	...	-0.0064050012
53.0352566735113	0.68092986	-0.25059018	0.864197940	0.0090557601	...	-0.0065237438
53.2111019849418	0.68244000	0.75032969	-0.135727313	1.0087452507	...	-0.0064699575
53.2796098562628	0.68303295	-0.24904618	0.864366199	0.0089172765	...	-0.0064850955
53.5507871321013	0.68476221	0.75158768	-0.135165364	1.0081878334	...	-0.0063903286
53.8904380561259	0.68609191	0.75317696	-0.134955418	-0.9928040893	...	-0.0063339843
...	...	...	...	...	...	...
101.070298836413	0.0000000000	0.0000000000	0.0000000000	0.9402234455		0.0000000000

Ensemble des parisiennes dépendantes classées par âge de survenance

Distance entre le vecteur covariable des sujets et la moyenne pondéerée des vecteurs covarariables des sujets dépendants

FIGURE 2.2 – Matrice des résidus de Schönefeld pour les Parisiennes pour le risque dépendance

Le résidu correspondant à la covariable j et à la $i$ème durée non-censurée est donné par :

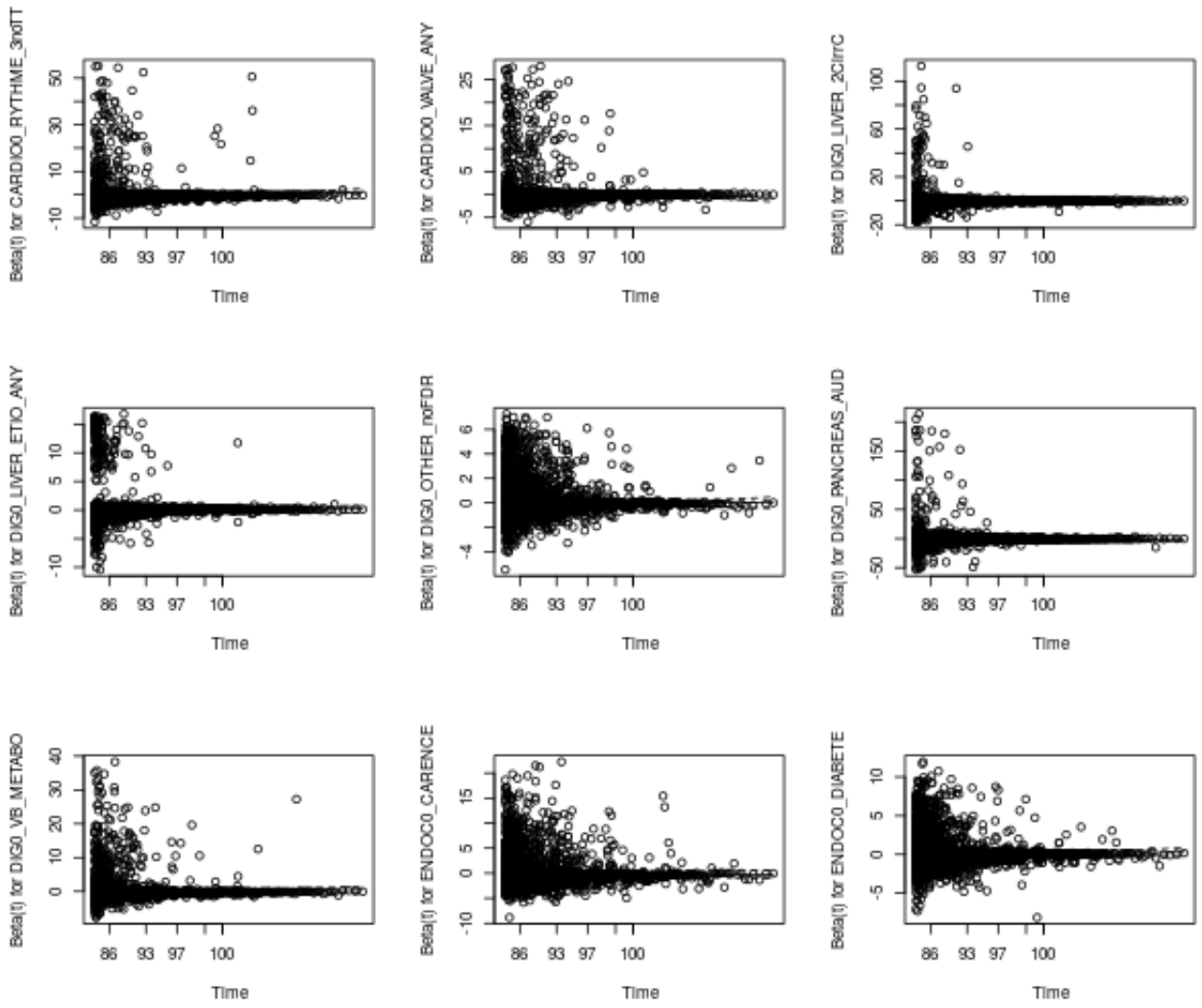
$$\widehat{Scho}_i^{(j)} = Z_i^{(j)} - \bar{Z}^{(j)}(\widehat{\beta}, T_i)$$

où : $\bar{\mathbf{Z}}(\beta, t) = (\bar{Z}^{(1)}(\beta, t), ..., \bar{Z}^{(p)}(\beta, t))$

est le vecteur moyenne pondérée à l'instant t des vecteurs de covariables des sujets à risque à l'instant t dont la $j$ème coordonnée est donnée par :

$$\bar{Z}^{(j)}(\beta, t) = \sum_{k=1}^{n} \frac{I(T_k \geq t) exp(\beta' \mathbf{Z}_k)}{\sum_{l=1}^{n} I(T_l \geq t) exp(\beta'.\mathbf{Z}_l)} Z_k^{(j)}$$

Les graphes ci-dessous représentent les tracés des résidus de Schönefeld pour quelques covariables de la base des parisiennes pour le risque dépendance.

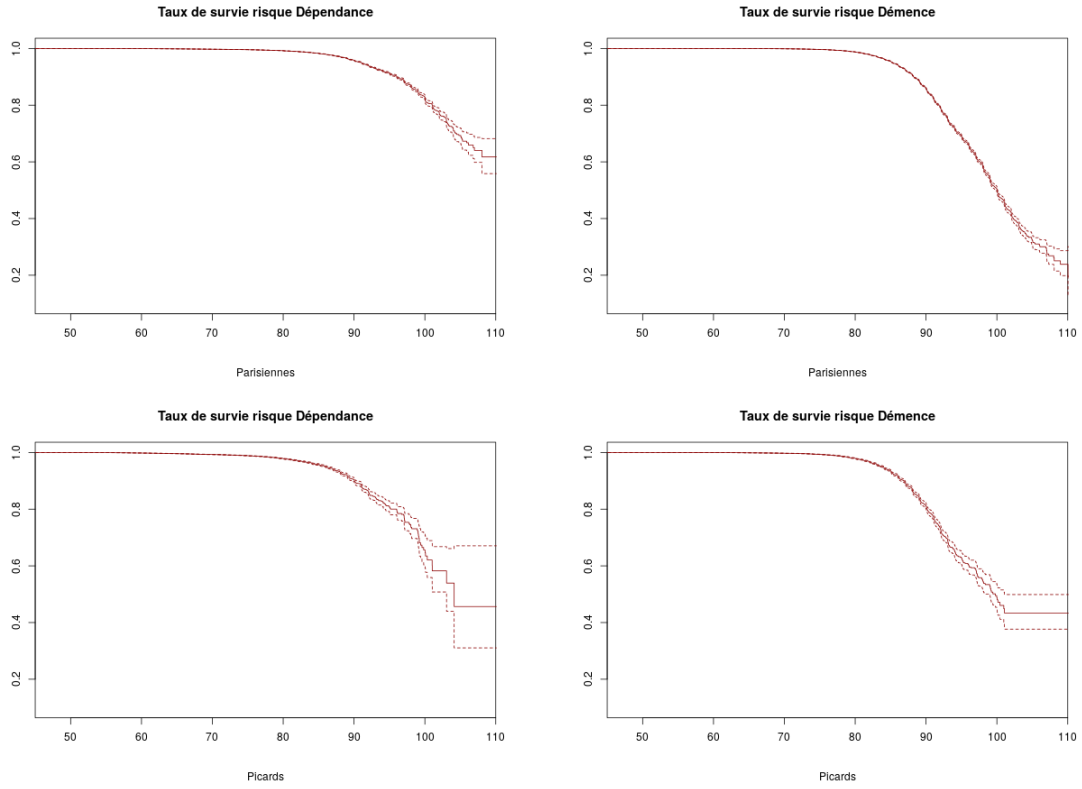


On constate que les log-ratios sont relativement uniformément répartis autour de l'axe des abscisses qui représente le temps. Au vu de ces graphes, on peut considérer que l'hypothèse de proportionnalité est bien vérifiée. Ceci a été vérifié pour les 89 variables des bases Picards et Parisiennes pour les risques dépendance et démence.

Il existe par ailleurs des modèles de régression comme le CART fondés sur des arbres de régression qui permettent de s'affranchir de l'hypothèse de proportionnalité. Les travaux de Lopez [2016] traitent par exemple la modélisation de données censurées par le biais de ces méthodes.

2.2.2 Fonction de survie et taux d'incidences brutes

Les graphes ci-dessous représentent les fonctions de survie des phénomènes de dépendance physique et de démence pour les populations picardes et parisiennes :



L'intervalle de confiance associé à la fonction de survie S(t), sous l'hypothèse que l'erreur de S(t) est distribuée normalement, est donné par la formule :

$$IC = \hat{S}(t) \pm Z_{\alpha/2}\sqrt{Var(\hat{S}(t))}$$

avec $Z_{\alpha/2}$ la valeur critique de la loi normale centrée réduite qui donne une probabilité d'erreur de α pris ici à 5%.

Les courbes représentant les taux d'incidence s'en déduisent. Ci-dessous sont représentés les taux d'incidence brutes constatés sur les bases des populations parisiennes et picardes pour les phénomènes de dépendance physique et de démence :

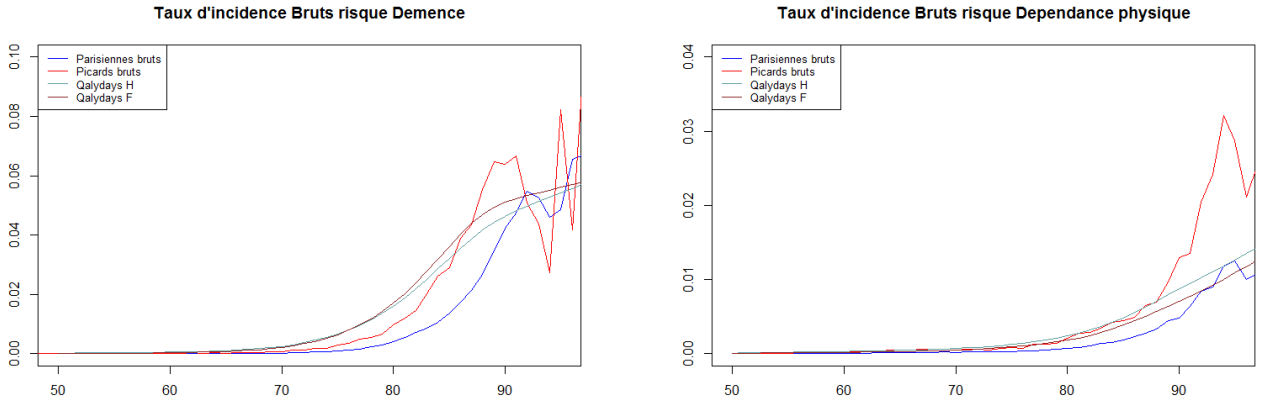


FIGURE 2.3 – Taux bruts démence et dépendance physique

Comme nous nous y attendions, nous constatons que la population picarde est plus à risque que la population parisienne et ce, aussi bien pour le risque dépendance physique que le risque démence. Ceci s'explique a priori

par l'influence du sexe et par celle du niveau de vie mais aussi par les disparités en termes d'accès aux soins entre ces deux populations. Pour la dépendance physique comme pour la démence, l'incidence est presque nulle avant 70 ans mais croit sensiblement pour les âges plus avancés. L'incidence de la démence augmente en effet quasi linéairement après 70 ans pour atteindre une incidence d'environ 6% à 100 ans. Le constat est identique pour la dépendance physique avec une évolution à peu près linéaire qui commence à 70 ans et qui atteint une incidence de 1.5% à 100 ans. Les courbes bleues claires et marron correspondent aux lois d'incidence construites par le groupe de travail Qalydays sur la France entière et constitueront pour la suite des travaux une base de référence pour juger de la bonne qualité de prédiction des modèles construits. Comme mentionné précédemment, le fait de ne pas tenir compte du caractère informatif de la censure aléatoire engendre un biais sur l'estimation des coefficients de régression. Par ailleurs, les taux d'incidence bruts présents dans les graphes ci-dessus concernent une population de personnes à risque puisqu'elles sont déjà présentes à l'hôpital et surestiment donc l'incidence sur la population française. Les travaux de Q.GUIBERT [2018] fournissent une estimation du risque de base pour la démence et la dépendance sans ce bais.
L'étude des coefficients de Cox qui s'en suit va permettre de mettre en lumière les facteurs les plus influents quant à l'apparition de ces deux phénomènes.

2.2.3 Sélection des covariables

Nous sommes maintenant en mesure d'estimer la valeur des coefficients de Cox pour un ensemble de covariables données à partir de nos populations. Il se pose alors la question du choix des variables les plus pertinentes. Le critère d'information d'Akkaike plus communément appelé critère AIC permet de mesurer la qualité d'un modèle prédictif tout en pénalisant le nombre de variables de celui-ci afin de respecter le principe de parcimonie. Le critère d'information d'Akaike s'écrit comme suit :

$$AIC = 2k - 2ln(L)$$

où k est le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle.
Si l'on considère un ensemble de modèles candidats, le modèle choisi est celui qui aura la plus faible valeur d'AIC.
L'AIC est d'autant plus faible que la vraisemblance est élevée et que le nombre de variables est restreint. Le critère AIC représente donc un compromis entre le biais, diminuant avec le nombre de paramètres libres, et la parcimonie, volonté de décrire les données avec le plus petit nombre de paramètres possibles.

Lorsque le nombre k de variables explicatives disponibles n'est pas trop élevé, il est envisageable de considérer tous les modèles possibles :
On a :

$$C_r^k = \frac{k!}{r!(k-r)!}$$

modèles différents faisant intervenir r variables explicatives.

Ce qui fait :

$$\sum_{r=0}^{k} C_r^k = 2^k$$

Il y a dans notre cas $2^{89} \simeq 10^{26}$ modèles possibles, soit l'ordre de grandeur du nombre d'atomes dans le corps humain.
La méthode dite "Best Subsets Regression" visant à ajuster l'ensemble des modèles possibles est donc inenvisageable en pratique avec notre jeu de données.
Pour mettre en œuvre la sélection de modèles au moyen du critère AIC et pour éviter de calculer l'exhaustivité des modèles possibles, on utilise généralement les méthodes pas à pas dite ascendantes, descendantes ou une combinaison des deux.

Méthode AIC forward

La méthode AIC forward consiste à partir du modèle avec comme unique variable celle ayant le critère AIC le plus faible et d'ajouter pas à pas les variables qui respectent la contrainte d'AIC la plus faible.

La méthode ascendante est très satisfaisante pour l'utilisateur préférant avoir toutes les variables possibles afin de ne rien ignorer mais elle a un inconvénient majeur, il n'est plus possible de réintroduire une variable une fois qu'elle a été supprimée.

Méthode AIC backward

La méthode descendantes part quant à elle du modèle complet et cherche à éliminer pas à pas les variables en gardant la logique d'avoir le critère AIC toujours le plus faible.
La méthode ascendante évite de travailler avec plus de variables que nécessaire, améliore l'équation à chaque étape. Mais l'inconvénient majeur de la méthode descendante réside dans le fait qu'une variable introduite dans le modèle ne peut plus être éliminée. Le modèle final peut alors contenir des variables non significatives. Ce problème est alors résolu par la procédure stepwise qui combine les deux méthodes.

Méthode AIC stepwise

La méthode stepwise est une amélioration de la méthode ascendante. En effet à chaque étape, nous réexaminons toutes les variables introduites précédemment dans le modèle. En effet, une variable considérée comme significative à une étape de l'algorithme peut à une étape ultérieure devenir non significative en raison de ces corrélations avec d'autres variables introduites après coup dans le modèle.
La procédure stepwise propose après l'introduction d'une nouvelle variable dans le modèle de réexaminer le critère AIC pour chaque variable explicative anciennement admise dans le modèle. Après réexamen, si des variables ne sont plus significatives, alors elle retire du modèle la moins significative d'entre elles. Le processus continue jusqu'à ce que plus aucune variable ne puisse être introduite ni retirée du modèle.

2.2.4 Comparaison des résultats

Ci-dessous, on présente les résultats pour les différentes méthodes AIC pas à pas appliquée aux deux bases pour chacune des deux sorties dépendance et démence.

	Parisienne Demence			Picard Demence			Parisienne Dependance			Picard Dependance		
Covariable	Stepwise	Backward	Forward	Stepwise	Backward	Forward	Stepwise	Backward	Forward	Stepwise	Backward	Forward
fdr_aud_all	**4,14**	**4,14**	**4,12**	**4,13**	**4,13**	**4,14**	3,77	3,77	3,77	4,02	4,02	4,04
fdr_smoker	2,94	2,94	2,94	2,50	2,50	2,49	3,27	3,27	3,25	2,60	2,60	2,58
CANCER1_SMOKER	2,13	2,13	2,14	1,75	1,75	1,74	2,15	2,15	2,15	1,77	1,77	1,75
CANCER1_PC_POOR	2,03	2,03	2,02	1,60	1,60	1,60	2,09	2,09	2,10	1,63	1,63	1,63
fdr_obesity_all	1,77	1,77	1,77	1,26	1,26	1,27	2,01	2,01	2,01	1,36	1,36	1,36
CANCER1_AUD_SMOKER	1,66	1,66	1,66	1,41	1,41	1,41	1,68	1,68	1,68	1,42	1,42	1,42
CANCER1_PC_GOOD	1,60	1,60	1,59	1,22	1,22	1,22	1,63	1,63	1,64	1,24	1,24	1,24
CANCER1_BREAST	1,44	1,44	1,44	-	-	1,13	1,51	1,51	1,51	-	-	1,12
NEURO1_DEM_FDR	**1,36**	**1,36**	**1,36**	**1,42**	**1,42**	**1,42**	1,34	1,34	1,34	1,32	1,32	1,31
NEURO1_PARKINSON	**1,29**	**1,29**	**1,28**	**1,11**	**1,11**	**1,12**	1,22	1,22	1,22	-	-	1,01
DIG1_LIVER_1CirrD	1,24	1,24	1,22	1,21	1,21	1,24	1,25	1,25	1,25	1,22	1,22	1,26
CANCER1_COLORECTAL	1,24	1,24	1,24	1,23	1,23	1,23	1,30	1,30	1,30	1,23	1,23	1,23
CANCER1_HEMATO	1,24	1,24	1,24	1,24	1,24	1,24	1,31	1,31	1,31	1,31	1,31	1,30
STROKE1_1HEMO	**1,22**	**1,22**	**1,22**	**-**	**-**	**1,11**	1,22	1,22	1,22	-	-	1,12
NEURO1_EPILEPSIE	**1,21**	**1,21**	**1,21**	**1,21**	**1,21**	**1,21**	1,13	1,13	1,13	1,21	1,21	1,21
INFECT1_SEPSIS	1,20	1,20	1,20	1,16	1,16	1,16	1,19	1,19	1,19	1,18	1,18	1,18
DIG1_STOMIE	1,16	1,16	1,17	1,32	1,32	1,33	1,21	1,21	1,21	1,30	1,30	1,31
BLOOD0_2OTHER	1,13	1,13	1,12	1,05	1,05	1,06	1,15	1,15	1,15	-	-	1,04
cp_dipl0	1,13	1,13	1,12	1,05	1,05	1,05	1,16	1,16	1,16	1,06	1,06	1,06
TRAUMA0_SUICIDE	1,10	1,10	1,10	-	-	1,03	1,08	1,08	1,08	-	-	1,02
DIG0_LIVER_ETIO_ANY	1,09	1,09	1,09	1,05	1,05	1,05	1,10	1,10	1,10	1,05	1,05	1,05
RESP1_2INSUF_AIGUE	1,09	1,09	1,09	-	-	1,03	1,11	1,11	1,10	-	-	1,03
BLOOD0_2ANEMIA_LYSE	1,08	1,08	1,08	1,21	1,21	1,22	1,10	1,10	1,10	1,22	1,22	1,23
BLOOD1_1TRANSFUSION	1,08	1,08	1,08	1,10	1,10	1,10	1,11	1,11	1,11	1,10	1,10	1,10
RHEUM0_SYSTEME	1,08	1,08	1,08	-	-	1,02	1,12	1,12	1,12	-	-	1,05
ENDOC0_DIABETE	1,08	1,08	1,08	1,03	1,03	1,03	1,07	1,07	1,07	1,04	1,04	1,04
ENDOC1_METABO	1,06	1,06	1,06	-	-	1,01	1,06	1,06	1,06	-	-	1,02
ENDOC1_GLD_OTHER	1,06	1,06	1,06	1,16	1,16	1,17	1,06	1,06	1,06	1,12	1,12	1,13
RESP0_2OTHER	1,06	1,06	1,06	-	-	0,97	1,09	1,09	1,09	-	-	0,96
CARDIO0_VALVE_ANY	1,05	1,05	1,05	1,03	1,03	1,03	1,07	1,07	1,07	1,04	1,04	1,04
NEURO0_SNP_METABO	1,05	1,05	1,05	1,10	1,10	1,10	1,06	1,06	1,06	1,08	1,08	1,08
DIG1_OCCLUSION	1,05	1,05	1,05	-	-	1,00	1,03	1,03	1,03	-	-	1,00
RESP1_1INSUF_CHRO	1,05	1,05	1,04	0,92	0,92	0,90	1,08	1,08	1,08	0,94	0,94	0,92
KIDNEY1_2INSUF_AIGUE	1,02	1,02	1,03	1,12	1,12	1,12	1,03	1,03	1,03	1,11	1,11	1,11
cp_dep	1,01	1,01	1,02	-	-	1,01	1,01	1,01	1,01	-	-	0,99
ENDOC0_DYSLIPIDEMIA	0,99	0,99	0,99	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,96	0,96
ENDOC0_THYROIDE	0,99	0,99	0,99	-	-	0,97	0,97	0,97	0,97	0,94	0,94	0,94
RESP0_2COPD	0,98	0,98	0,98	-	-	1,02	-	-	1,01	-	-	1,02
STROKE0_AIT_noSEQ	0,98	0,98	0,98	-	-	0,99	0,97	0,97	0,97	-	-	0,98
CANCER0_BENIN	0,97	0,97	0,97	0,90	0,90	0,90	0,98	0,98	0,98	0,90	0,90	0,89
CARDIO1_RYTHME_1ACFA	0,97	0,97	0,97	0,87	0,87	0,88	-	-	1,01	0,87	0,87	0,88
CANCER0_INSITU	0,97	0,97	0,97	-	-	1,03	0,98	0,98	0,98	-	-	1,02
ENDOC0_CARENCE	0,97	0,97	0,97	0,89	0,89	0,88	0,91	0,91	0,91	0,83	0,83	0,83
STROKE0_3noTTT	0,97	0,97	0,97	-	-	0,97	0,96	0,96	0,95	-	-	0,94
DIG0_VB_METABO	0,97	0,97	0,97	0,89	0,89	0,89	0,98	0,98	0,98	0,87	0,87	0,87
CARDIO0_IHD_3noTTT	0,96	0,96	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,96	0,96	0,96
RESP0_LRI_ATCD	0,96	0,96	0,96	-	-	0,98	0,97	0,97	0,97	-	-	0,98
RHEUM0_AUD	0,96	0,96	0,96	0,93	0,93	0,93	0,95	0,95	0,95	0,94	0,94	0,94
CARDIO1_INSUF_CHRO	0,96	0,96	0,96	0,94	0,94	0,95	-	-	1,00	0,95	0,95	0,96
TRAUMA0_CHUTE	0,96	0,96	0,96	0,93	0,93	0,93	0,92	0,92	0,92	0,88	0,88	0,88
CARDIO1_IHD_1MI	0,96	0,96	0,96	0,88	0,88	0,88	-	-	1,01	0,85	0,85	0,84
TRAUMA0_3FRACTURE	0,95	0,95	0,95	0,93	0,93	0,93	0,93	0,93	0,93	0,94	0,94	0,94
RHEUM0_ARTHROSE_OTHE	0,95	0,95	0,95	0,91	0,91	0,91	0,92	0,92	0,92	0,90	0,90	0,91
CARDIO0_HBP	0,95	0,95	0,95	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92	0,92
DIG0_OTHER_noFDR	0,95	0,95	0,95	0,93	0,93	0,93	0,93	0,93	0,94	0,93	0,93	0,93
CARDIO0_RYTHME_3noTT	0,94	0,94	0,94	0,93	0,93	0,93	0,92	0,92	0,92	0,93	0,93	0,93
RESP0_APNEE_SOM	0,93	0,93	0,93	0,92	0,92	0,92	0,91	0,91	0,91	0,89	0,89	0,89
CARDIO0_RYTHME_2TTT	0,93	0,93	0,93	0,90	0,90	0,90	0,94	0,94	0,94	0,91	0,91	0,91
SENSE0_CATARACTE	0,93	0,93	0,93	0,90	0,90	0,90	0,93	0,93	0,93	0,90	0,90	0,90
KIDNEY1_1INSUF_CHRO	0,93	0,93	0,92	-	-	0,95	0,93	0,93	0,93	-	-	0,96
DIG1_HEMORRAGIE	0,93	0,93	0,93	-	-	0,95	0,92	0,92	0,91	-	-	0,97
RHEUM0_ARTHROSE_HIP	0,91	0,91	0,91	0,92	0,92	0,92	0,90	0,90	0,90	0,92	0,92	0,92
RHEUM0_ARTHROSE_KNEE	0,90	0,90	0,90	0,87	0,87	0,87	0,89	0,89	0,89	0,88	0,88	0,87
TRAUMA1_1CRANE	0,89	0,89	0,89	-	-	0,94	0,85	0,85	0,85	-	-	0,92
DIG1_PERITONITE	0,89	0,89	0,89	-	-	0,97	0,89	0,89	0,89	-	-	0,99
CANCER0_SKIN	0,88	0,88	0,88	0,90	0,90	0,90	0,89	0,89	0,89	0,92	0,92	0,92
DIG0_PANCREAS_AUD	0,85	0,85	0,85	0,89	0,89	0,89	0,84	0,84	0,84	0,91	0,91	0,91
NEURO1_OTHER	0,72	0,72	0,72	1,14	1,14	1,14	0,68	0,68	0,68	-	-	1,11
RESP0_2INTERSTI	-	-	1,04	1,27	1,27	1,26	1,08	1,08	1,08	1,30	1,30	1,29
DIG0_LIVER_2CirrC	-	-	1,02	-	-	0,99	1,05	1,05	1,05	-	-	0,98
CV1_MTE	-	-	1,01	1,07	1,07	1,07	1,04	1,04	1,04	1,07	1,07	1,08
KIDNEY0_2GN	-	-	1,01	-	-	1,03	1,04	1,04	1,04	-	-	0,99
CARDIO0_IHD_2TTT	-	-	1,01	0,97	0,97	0,97	1,03	1,03	1,03	0,97	0,97	0,97
cp_immi	-	-	1,00	-	-	1,00	0,99	0,99	0,99	-	-	0,99
BLOOD0_2ANEMIA_IRON	-	-	0,99	0,93	0,93	0,94	0,98	0,98	0,98	0,94	0,94	0,95
KIDNEY0_CYSTITE_ATCD	-	-	0,99	0,90	0,90	0,89	0,98	0,98	0,98	0,91	0,91	0,91
TRAUMA0_OSTEOPOROSE	-	-	1,01	1,11	1,11	1,11	0,98	0,98	0,98	1,08	1,08	1,08
CANCER0_ATCD_FAM	-	-	0,99	0,95	0,95	0,95	-	-	0,99	0,94	0,94	0,94
KIDNEY0_2OTHER	-	-	0,99	0,93	0,93	0,93	-	-	1,00	0,92	0,92	0,92
KIDNEY0_2PYELONEPHRI	-	-	1,01	0,85	0,85	0,86	-	-	0,99	0,88	0,88	0,89
CARDIO0_OTHER	-	-	1,01	-	-	1,01	-	-	0,99	-	-	1,00
KIDNEY0_2UROLITHIASE	-	-	0,97	-	-	0,97	-	-	0,99	-	-	0,95
RESP0_2ASTHMA	-	-	1,00	-	-	0,99	-	-	1,01	-	-	1,01
RESP0_BRONCHITE_ATCD	-	-	1,00	-	-	1,02	-	-	1,02	-	-	1,06
STROKE0_2TTT	-	-	1,04	-	-	1,03	-	-	1,03	-	-	1,06
CANCER1_PROSTATE	-	-	0,00	-	-	1,03	-	-	-	-	-	1,04
CV1_PVD	-	-	0,98	-	-	1,01	-	-	1,01	-	-	1,01
STROKE1_1ISCHEMIC	**-**	**-**	**1,02**	**-**	**-**	**0,97**	-	-	1,00	-	-	0,95
TRAUMA1_2SEVERE	-	-	1,03	-	-	0,97	-	-	1,03	-	-	0,94

FIGURE 2.4 – Comparaison des méthode pas à pas

Un problème majeur rencontré pour l'établissement de ces méthodes réside dans le temps nécessaire pour exécuter ces calculs. En effet, la procédure backward a mis plus de 24h à être exécutée sur la base des Parisiennes et ce malgré l'utilisation d'une machine avec un puissance de calcul élevée qui comporte 16 cœurs de 2.6 GHz par processeur pour une mémoire vive totale de 64 GB. Cependant, le principe de la procédure ne permet pas d'utiliser chacun des 16 cœurs de façon simultanée. En effet, chaque pas étant déterminé par le pas précédent, cela empêche d'utiliser une méthode de parallélisation pour cet algorithme.

Les coefficients obtenus sont classés par ordre décroissant par rapport au périmètre de la population parisienne pour le risque démence avec la méthode stepwise. On remarque que les résultats obtenus pour les trois méthodes sont très proches, cependant le processus de sélection est très mauvais pour la méthode forward.
Les variables identifiées comme à risque pour la démence dans la première partie apparaissent en gras. On remarque que le facteur de risque "Alcool" apparait en première position avec un coefficient multiplicatif de 4,14. Ceci implique que les personnes alcooliques auraient plus de 4 fois plus de chance de contracter une démence que les non alcooliques. Les facteurs de risque "Tabac" et "Obésité" ainsi que certains cancers complètent la liste des causes principales de démence. Les facteurs en gras identifiés dans la première partie ressortent bien comme causes explicatives d'entrée en démence puisque celles-ci sont toutes supérieures à 1. (mis à part AVC ischémiques qui n'est pas sélectionné). On constate en effet qu'une Parisienne qui contracte la maladie de Parkinson a 29% de chance en plus de devenir démente qu'une Parisienne n'ayant pas cette maladie.
On remarque par ailleurs que pour une même cause de sortie, l'intensité des coefficients est très proche entre les deux populations parisienne et picarde, ce qui montre la robustesse du modèle. Cependant les différences intrinsèques entre ces deux populations (sexe, niveau de vie, inégalité d'accès aux soins) expliquent qu'il y ait des écarts. On vérifie cela avec l'exemple du cancer du sein [2], exclusivement féminin (qui sur représente respectivement le risque dépendance et le risque démence de facteur de 1,51 et 1,44 pour les femmes alors qu'il n'est pas sélectionné pour les hommes.)
Enfin, on remarque que les causes de démence sont proches de celles de la dépendance physique. En effet les 4 facteurs les plus influents ressortent dans le même ordre pour les deux pathologies et avec des intensités proches. Cependant certains facteurs comme l'obésité [3] ont des intensités sensiblement différentes pour les deux causes de sortie : pour les parisienne on observe une intensité de 1,77 pour la démence versus 2,01 pour la dépendance ; pour les picards on observe une intensité de 1,26 pour la démence versus 1,36 pour la dépendance.

2. CANCER1_BREAST
3. fdr_obesity_all

Chapitre 3

Modélisation par un modèle de Cox et introduction de méthodes d'apprentissage

Introduction de la partie III

Dans la section précédente, nous avons vu que les étapes de paramétrage du modèle de Cox et de sélection des variables étaient séparées dans le cadre d'une procédure classique et que cela avait un effet très chronophage lorsque la dimension de la base d'étude était importante. Le cadre classique obligeant en effet à effectuer le calcul d'optimisation de la vraisemblance partielle autant de fois que la procédure de sélection l'imposait. Bien que les procédures de sélections de variables pas à pas n'imposent pas de calculer tous les modèles possibles, ces méthodes restent toutefois très lourdes à mettre en œuvre dans le cas de données de très grandes dimensions avec par exemple un temps de calcul supérieur à 24h dans notre cas.

L'idée de la régression pénalisée est d'inclure l'étape de pénalisation directement dans le problème d'optimisation de la vraisemblance en ajoutant une contrainte sur la norme des coefficients de régression.
Une famille de procédures de sélection de variables s'appuyant sur le concept de pénalisation de la vraisemblance a été proposée dans un premier temps en 2001 par Fan et Li pour des modèles paramétriques comme la régression linéaire, la régression linéaire robuste et les modèles linéaires généralisés. A travers ces travaux, Fan et Li ont démontré que ces procédures étaient aussi performantes que si l'ensemble des données significatives était déjà connu à l'avance. Une telle propriété est appelé une propriété Oracle. Par la suite, en 2002, Fan et Li se sont attardés au cas de la pénalisation de la vraisemblance partielle dans le cas d'un modèle de Cox.

Dans ce mémoire, nous nous intéresserons en particulier aux pénalisations de type Ridge, LASSO, E-net et Adaptive-LASSO. Avant d'appliquer ces différentes pénalisation à notre jeu de données, nous allons d'abord tester de façon empirique la convergence des estimateurs issus de chacune de ces régressions pénalisées sur un jeu de données simulées. Nous mesurerons également la qualité de sélection de chacune de ces procédures et comparerons par rapport à la procédure classique fondée sur une méthode pas à pas s'appuyant sur le critère d'Akkaike.

Enfin nous appliquerons ces différentes méthodes à notre jeu de données et tacherons de comparer leurs performances respectives avec celles de la méthode classique décrite dans le chapitre précédent.

3.1 Éléments théoriques - Présentation des concepts de pénalisation et d'apprentissage supervisé

3.1.1 Principe de la régression pénalisée

L'idée de la régression pénalisée est d'introduire l'étape de pénalisation directement dans le problème d'optimisation de la vraisemblance en ajoutant une contrainte sur la norme des coefficients de régression de la forme :

$$\sum_{j=1}^{p} |\beta_j|^{\delta} \leq C$$

La région des valeurs possibles pour les coefficients de régression dépende alors de la valeur du paramètre δ. Par exemple, en dimension 2, les régions autorisées des coefficients β_1 et β_2 en fonction du paramètre δ ont les formes suivantes :

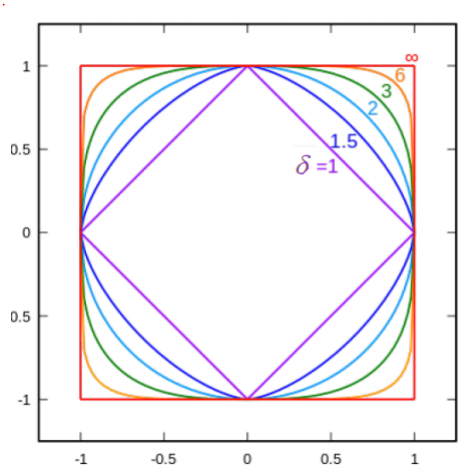


FIGURE 3.1 – Régions des contraintes en fonction de δ

Les travaux de VERWEIJ [1996] permettent de transposer la méthode de pénalisation, couramment rencontrée pour la méthode d'optimisation des moindres carrés ordinaires, au calcul d'optimisation de la vraisemblance. Le fait d'imposer une contrainte sur la norme des coefficients revient alors à exprimer directement la vraisemblance partielle pénalisée de la façon suivante :

$$L_{partielle}(\beta) = \prod_{i}^{D} \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - p(\beta, \lambda)$$

Où $p(\beta, \lambda)$ est la fonction de pénalisation. λ est le paramètre d'intensité ou paramètre de régularisation. Ce paramètre modère l'intensité de la pénalisation de façon linéaire.

En faisant varier la valeur de la constante de régularisation, on obtient alors l'ensemble de solutions suivant :

$$\{\hat{\beta}(\lambda) \in [0, +\infty[\text{ où } \hat{\beta}(\lambda) = argmin_{\beta} \prod_{i}^{D} \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - p(\beta, \lambda)\}$$

Cet ensemble de solutions s'appelle le chemin de régularisation. Le chemin de régularisation est une représentation de l'ensemble des valeurs prises par chacun des coefficients de régression en fonction de la valeur du paramètre de régularisation λ. Le nouveau problème revient alors à trouver la valeur optimale du compromis λ.

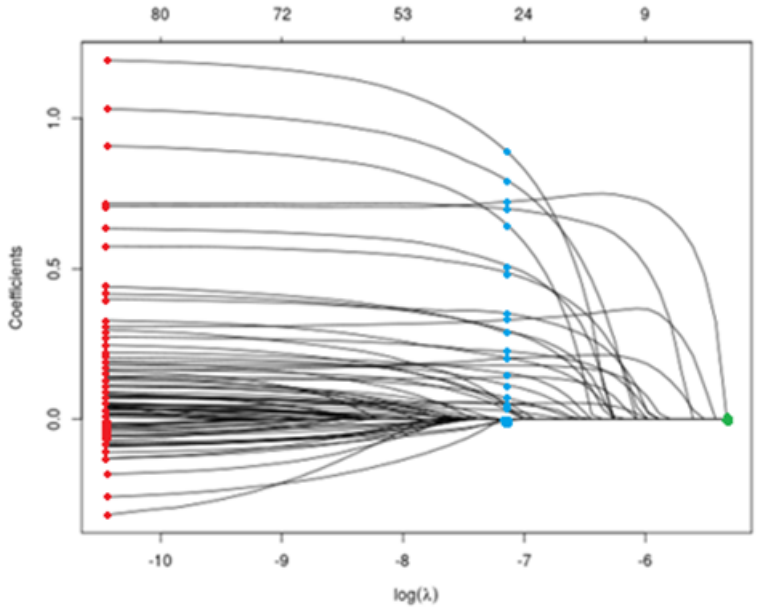


FIGURE 3.2 – Chemin de régularisation

Le chemin de régularisation ci-dessus est donné à titre illustratif. On remarque qu'en faisant varier l'intensité du paramètre de pénalisation on obtient des valeurs différentes pour le vecteur $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$ défini comme l'intersection des n courbes avec la droite verticale $x = \lambda$.

Pour $\lambda = 0$ ($\log(\lambda) \to -\infty$), en rouge sur le graphe
il n'y a aucune pénalisation. Nous sommes dans le cadre d'une régression classique avec un modèle complet qui sélectionne les n variables.

Pour $\lambda \to \infty$ ($\log(\lambda) \to 0$), en vert sur le graphe
la pénalisation est infinie. Le modèle sélectionné est le modèle trivial à 0 variable.

Entre ces deux cas extrêmes, il existe une valeur du paramètre de régularisation optimale. Cette valeur doit être suffisamment petite pour ne pas trop pénaliser le modèle et ainsi le décrire par un nombre suffisant de variables. Cette valeur doit par ailleurs ne pas être trop importante au risque de sélectionner trop de variables et de dégrader le caractère parcimonieux du modèle.

Pour trouver $\lambda_{optimal}$, en bleu sur le graphe
nous utiliserons des méthodes d'apprentissage qui seront décrites par la suite.

L'un des gros avantages de l'utilisation de cette technique réside dans la très grande rapidité de mise en œuvre des calculs et de la sélection. En effet, comme nous le verrons plus en détails par la suite, en utilisant la même machine que pour la sélection pas à pas du chapitre précédent, les temps de calcul nécessaires à la génération des chemins de régularisation et le temps nécessaire pour trouver le paramètre de régularisation optimale est d'environ 15 min sur la base de données du PMSI. La méthode classique pas à pas utilisée dans la deuxième section prenait quant à elle plus de 24h pour effectuer la sélection. Le gain de temps est considérable (la régression pénalisée est environ 100 fois plus rapide)
Comme évoqué en introduction, il existe plusieurs types de pénalisation. Parmi celles-ci seront introduites les pénalisations Ridge, LASSO, Elastic-net et Adaptative LASSO.

Pénalisation Ridge

La régression Ridge pénalise la vraisemblance par la somme des carrés des paramètres de régression.

La pénalisation Ridge s'écrit :

$$p(\beta, \lambda) = \lambda . \sum_{j=1}^{p} \beta_j^2$$

Dans le cadre de variables fortement corrélées, l'identification des coefficients de régression est difficile et a tendance à engendrer une grande variance. En effet, un coefficient particulièrement grand et positif sur une certaine variable peut être compensé par un autre coefficient également important et négatif sur une variable corrélée avec la première. Ainsi, la contrainte Ridge imposant une pénalité sur le carré des coefficients de régression, le phénomène décrit ci-dessus disparait. Les variables corrélées seraient ainsi toutes deux fortement pénalisées.

Le problème de la régression Ridge est que, du fait de sa forme, elle ne permet pas d'annuler les coefficients de régression. Le schéma [1] suivant illustre la difficulté de la contrainte Ridge à annuler les coefficients de régréssion dans le cas d'une régression linéaire classique avec deux variables explicatives.

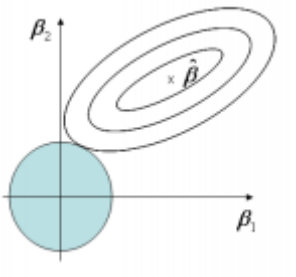


FIGURE 3.3 – Illustration de la contrainte Ridge

La zone bleue est la région de la contrainte définie en figure 3.1 pour $\delta = 2$. Dans le cas bi-dimensionnel facilement modélisable graphiquement, on voit que la contrainte sur la somme des carrés des deux coefficients est l'équation d'un disque :

$\beta_1^2 + \beta_2^2 \leq t^2$

Les courbes iso-erreur pour une régression classique sont des ellipses (ou plus généralement des coniques si la dimension excède 2) d'équation $\| X\hat{\beta} - Y \|^2 = c$. Plus les β_i sont élevés, plus la vraissemblance est importante, la solution du problème d'optimisation est donc l'intersection des courbes iso-erreur avec le cercle $\beta_1^2 + \beta_2^2 = t^2$.

On observe que la solution optimale donne un β_1 et un β_2 non nuls. La probabilité que l'intersection se fasse sur l'un des deux axes est très faible. C'est le cas seulement pour les familles d'ellipses dont l'un des rayons est parallèle à l'un des axes. Rien n'impose d'être dans ce cas très particulier.
Le principe est le même en augmentant le nombre de dimensions, bien que la représentation graphique devienne très difficile voire impossible au delà de 3 dimensions.

Dans le cadre d'une régression de Cox, les courbes iso-erreur ne vérifient pas l'équation d'une conique du fait de l'exponentielle qui déforme celles-ci.
Par ailleurs l'erreur dépend également du temps :

$$\| \alpha_0(t) exp(X^t \hat{\beta}) - Y \|^2 = c$$

Cependant on admettra que cette région de contrainte est dans ce cas également peu favorable pour une intersection sur l'un des deux axes.

1. Statistical Learning with sparcity Hastie and Tibshirani [2016]

Cette pénalité est donc intéressante dans un univers fortement corrélé mais elle ne permet pas d'annuler les coefficients β associés à chaque covariable. Par conséquent cette pénalité ne permet pas d'effectuer de sélection.

Pénalisation LASSO

La méthode Lasso a été proposée par Tibshirani en 1996. C'est certainement la pénalisation la plus répandue à ce jour. Initialement, celle-ci a été introduite dans le but de pénaliser la somme des carrés résiduels en imposant une contrainte sur la norme l_1 des coefficients dans la méthode des moindres carrés ordinaires. Cette pénalisation s'adapte également au problème d'optimisation de la vraisemblance comme nous allons le voir par la suite.
La pénalisation LASSO introduit une contrainte sur la norme des coefficients de régression. Contrairement à la régression Ridge qui pénalisait l'expression par une norme l_2, la régression LASSO pénalise la vraisemblance par la norme l_1 du vecteur des coefficients de régression. Elle s'écrit :

$$p(\beta, \lambda) = \lambda . \sum_{j=1}^{p} |\beta_j|$$

La pénalisation LASSO a donc elle aussi tendance à rétrécir les coefficients β. En effet, plus la valeur du coefficient β_j sera importante, plus la vraisemblance sera pénalisée. Par ailleurs, plus le paramètre de lissage λ sera grand et plus la pénalisation sera puissante. Comme évoqué précédemment, on voit que dans le cas extrême où le paramètre de régularisation λ est très grand, tous les coefficients seront fixés à 0 puisque l'apport en termes de vraisemblance des variables restera insuffisant par rapport à ce que la pénalisation engendrera en contrepartie sur la vraisemblance. A l'opposé, pour un paramètre de régularisation λ nul, il n'y a pas de pénalisation, toutes les variables sont exprimées et participent à améliorer la vraisemblance du modèle. Par ailleurs, la forme de la contrainte permettant d'annuler les coefficients de régression, on déduit qu'en choisissant convenablement λ on peut obtenir un modèle parcimonieux. Plus le paramètre de régularisation est faible plus le modèle sélectionnera un nombre de variables important. L'avantage de la pénalisation LASSO est alors de pouvoir sélectionner les variables pertinentes du modèle directement en résolvant le problème d'optimisation de la vraisemblance pénalisée. Le problème devenant alors de trouver le paramètre de régularisation qui optimise le compromis entre minimisation du biais et augmentation de la parcimonie.

Pour mieux visualiser pourquoi les coefficients peuvent s'annuler avec cette forme de contrainte, réinteressons nous à la zone de contrainte dans le cas bi-dimensionnel [2].

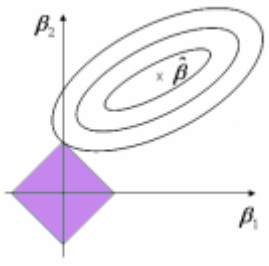


FIGURE 3.4 – Illustration de la contrainte LASSO

La région des contraintes, en violet sur le graphique, est définie dans le cadre du LASSO comme la somme des valeurs absolues des deux coefficients. Cette zone est un losange et son équation est :

$$|\beta_1| + |\beta_2| \leq t$$

2. Statistical Learning with sparcity Hastie and Tibshirani [2016]

En reprenant le cas d'une régression classique, les courbes iso-erreur sont encore des ellipses d'équation $\|X\hat{\beta} - Y\|^2 = c$
La solution du problème d'optimisation est l'intersection des ellipses iso-erreur avec le losange $|\beta_1| + |\beta_2| = t$. En augmentant la taille des ellipses pour trouver l'intersection avec la zone de contrainte on remarque qu'il est bien plus probable que l'intersection se fasse sur un coin de la zone de contrainte que dans le cas du disque avec la norme l_2 où la probabilité était très faible. Dans l'illustration ci-dessus, la solution optimale impose par exemple à β_2 d'être nul.
Encore une fois, nous notons que la forme des courbes iso-erreurs dans le cadre d'une régression de Cox ne vérifie pas l'équation d'une ellipse. Nous admettrons que la forme de la contrainte en losange augmente la probabilité d'intersection sur l'un des deux axes.

Pénalisation Elastic-net

Une alternative aux régressions pénalisées par la norme l_1 et l_2 consiste à les combiner pour avoir d'une part l'avantage de gérer les problèmes de colinéarité et d'autres part à l'avantage de gérer le problème de parcimonie. C'est la pénalisation Elastic-net. Celle-ci s'exprime de la façon suivante :

$$p(\beta, \lambda_1, \lambda_2) = \lambda_1 . \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

ou encore en introduisant $\alpha = \dfrac{\lambda_2}{\lambda_1 + \lambda_2}$ la pénalisation s'écrit (à un facteur $\dfrac{1}{\lambda_1 + \lambda_2}$ près) :

$$p(\beta, \lambda, \alpha) = \lambda(\alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|)$$

La pénalisation Ridge rétrécie les coefficients des variables corrélées entre elles alors que le LASSO tend à choisir l'un des deux et à fixer l'autre à 0. La pénalité Elastic-net combine les deux ; si les covariables sont correlées par groupe, un coefficient $\alpha = 0.5$ tends à sélectionner l'ensemble du groupe ou à l'exclure totalement. La valeur α peut, au même titre que le coefficient de régularisation être choisi par une méthode d'apprentissage en comparant in fine ses performances sur un échantillon test.

A noter que, l'Elastic net avec $\alpha = 1 - \epsilon$, pour un ϵ positif très petit, sera en théorie aussi performant que le LASSO, mais pourra cependant retirer toutes dégénérescences possibles causés par des corrélations extrêmes.

Pénalisation Adaptive-LASSO

La méthode Adaptive-Lasso a été proposée par Zou en 2006. C'est une alternative à la pénalisation LASSO classique qui introduit une pondération pour chacun des différents coefficients de la régression. La pénalisation Adaptive-LASSO s'écrit :

$$p(\beta, \lambda) = \lambda . \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|$$

Il est d'usage de choisir le vecteur de poids sous la forme suivante :

$$\hat{\omega}_j = (|\hat{\beta}_j(LASSO)| + \frac{1}{n})^{-\gamma}$$

Où n représente le nombre de covariables. De plus, comme dans le cadre de l'Elastic-net, on remarque qu'un nouveau paramètre de lissage intervient. Il est courant de choisir un nombre fini et raisonnable de valeur pour γ et de les tester.

3.1.2 Principe de la théorie de l'apprentissage

Dans la section précédente, un certain nombre de pénalités a été présenté. Il a par ailleurs été expliqué que celles-ci ont pour but de pénaliser la vraisemblance et que selon l'intensité du facteur de pénalisation, nous obtenions un ensemble de solutions pour l'ensemble des valeurs des coefficients de régression. L'objectif est alors de déterminer le paramètre de régularisation optimal, celui qui offre le meilleur compromis entre biais et parcimonie. Cette étape n'a pas été décrite précédemment. En effet, toutes les méthodes introduites dans ce chapitre donnent un ensemble de solutions $\hat{\beta}(\Theta)$ pour un ensemble de valeurs du paramètre Θ ($\Theta = \lambda$ pour le LASSO et le Ridge, $\Theta = (\lambda, \alpha)$ pour l'E-net, $\Theta = (\lambda, \mu)$ pour l'Adaptative Lasso) donnant à l'utilisateur le choix de la solution particulière $\hat{\beta}(\Theta)$ selon le compromis biais/parcimonie qu'il envisage. Une approche générale consiste à chercher le paramètre de régularisation qui minimise l'erreur de prédiction.

Une famille d'algorithmes, souvent rencontrée sous le nom de Machine Learning, est de plus en plus utilisée pour répondre à ce type de problème et modifie quelque peu la logique d'estimation du modèle. En effet, dans une démarche statistique « classique », on construit un estimateur sur un jeu de données unique et une théorie asymptotique permet de juger de sa qualité et de construire des intervalles de confiance. Cependant, dans la logique de la théorie de l'apprentissage, les données sont séparées en deux, avec des échantillons d'apprentissage et de validation ; la qualité des estimateurs n'est alors plus jugée par l'intermédiaire des critères asymptotiques, mais en fonction de l'adéquation à l'échantillon de validation par l'intermédiaire d'une mesure de l'erreur de prédiction.

Plusieurs nouvelles problématiques se posent alors. L'impact des tailles respectives des bases, utilisées pour l'apprentissage et la validation, sur la qualité de la prédiction sera ainsi mesuré et des techniques d'échantillonnages seront introduites.
Par la suite, plusieurs mesures qui permettent de juger la qualité de prédiction sur l'échantillon de validation seront présentées. Le calcul d'optimisation entrant en jeu faisant intervenir un calcul de vraisemblance, la mesure de la déviance sera naturellement utilisée. Il est par ailleurs classique dans le cadre des problèmes de classification de comparer les taux de vrais et faux positifs ainsi que ceux de vrais et faux négatifs pour juger de la bonne adéquation de l'estimateur sur la base de validation. Cependant dans le cas de données de survie le problème est plus complexe du fait de la nature dépendante du temps et censurée des données. La mesure AUC de Chambless et Dio (2006) sera alors introduite. Celle-ci a en effet été construite pour répondre à ce type de problème avec des données dépendantes du temps.

Choix des échantillons - Compromis biais-variance - Validation croisée

Nous avons évoqué en introduction que la théorie de l'apprentissage nécessitait la séparation de la base disponible en un échantillon d'apprentissage sur lequel on ajuste les modèles et en un échantillon de validation sur lequel on teste leurs performances. Le meilleur modèle étant alors celui construit sur le premier échantillon et qui affiche la meilleure performance sur le second. Le cours suivant expose les bases théoriques du compromis biais-variance et le choix de la taille respective des bases d'apprentissage et de test :http ://eric.univ-lyon2.fr/ ricco/cours/slides/resampling_evaluation.pdf
En effet, l'erreur en test est un estimateur non biaisé du modèle construit sur la partie apprentissage mais c'est un mauvais estimateur de l'erreur commis par le modèle construit sur l'ensemble des données. La subdivision « apprentissage-test » est alors nécessaire mais elle peut engendrer le dilemme biais variance si la taille des échantillons est limitée.
La figure suivante illustre bien le compromis biais variance.

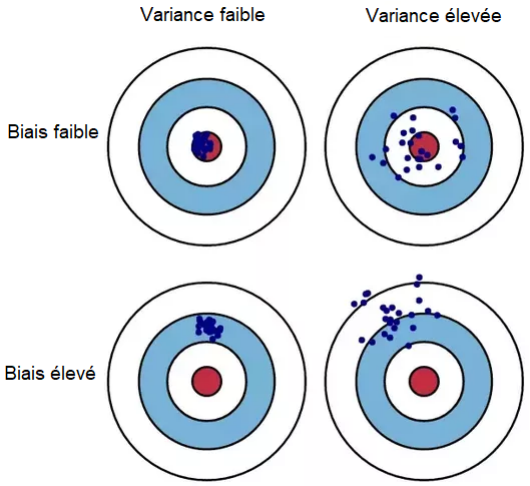


FIGURE 3.5 – Compromis biais variance

Plus le biais de l'estimateur est faible, plus le centre de gravité des estimations est proche du centre de la cible. Plus la variance de l'estimateur est faible plus les estimations sont concentrées.

Une manière d'optimiser la contrainte biais-variance est de découper l'échantillon en k parties égales. Chacune de ces k parties est utilisée à tour de rôle comme jeu de validation. Le reste étant utilisé pour entrainer le modèle. En procédant ainsi, chaque observation sert une fois comme validation et $k-1$ fois à entrainer le modèle. De plus chaque prédiction associée à chacune des observations a été construite avec une base d'apprentissage contenant cette observation. On évite le phénomène de sur-apprentissage. La performance du modèle peut alors être mesurée à partir des prédictions faites sur l'ensemble des données (il y a bien autant de prédictions que d'observations).

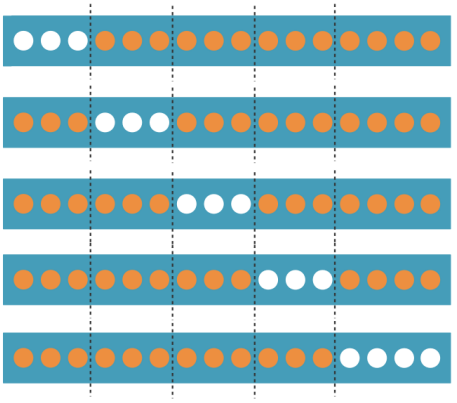


FIGURE 3.6 – Principe de la k-folds Validation croisée (source : openclassrooms)

La performance du modèle peut alors être mesurée à partir des prédictions faites sur l'ensemble des données (il y a bien autant de prédictions que d'observations).
On note qu'en augmentant le nombre de subdivisions k, on augmente le nombre de modèles ajustés et donc mécaniquement le temps de calcul pour leurs mises en oeuvre. La technique d'échantillonnage consistant à choisir autant de subdivisions qu'il y a de données disponibles, est appelée Leave-one-out. Celle-ci est en pratique peu utilisée car elle est très chronophage.
Cependant nous seront peu confrontés au dilemme biais-variance dans cette étude puisque chacun des deux échantillons d'apprentissage et de validation pourra être choisi suffisamment grand au regard de l'abondance des données disponibles.

Choix de la mesure de prédiction

L'introduction d'une mesure permettant d'évaluer les performances d'un modèle de prédiction est primordiale. En effet, pour comparer les modèles candidats et pour trouver le paramètre de régularisation optimal, il faut pouvoir comparer les prédictions faites sur l'échantillon de test avec les vraies valeurs sur cet ensemble. Les mesures les plus utilisées pour quantifier la qualité de l'apprentissage sont le taux d'erreur et la mesure AUC. Cependant, ces mesures sont adaptées à des problèmes de classification ce qui n'est pas le cas de notre présente étude. En effet, le caractère censuré des données ne permet pas de classer chaque observation de façon binaire comme "dependant"/"non dépendant" (resp. "dément"/"non dément"). Il faudrait en effet pouvoir observer chaque patient sur sa vie entière pour pouvoir mesurer simplement les sorties de façon binaire et pouvoir établir un taux de bons ou mauvais classés.
La mesure AUC nous intéresse cependant car HEAGERTY ZHENG [2005] et CHAMBLESS & DIAO [2006] ont introduit une mesure AUC dépendante du temps que nous pouvons tester sur notre jeu de donnée. Nous rappelons rapidement le principe de la mesure AUC pour introduire le concept dépendant du temps

Courbe AUC

La fonction d'efficacité du récepteur, plus fréquemment désignée sous le terme « courbe ROC » (de l'anglais receiver operating characteristic, pour « caractéristique de fonctionnement du récepteur ») est souvent utilisée pour mesurer la qualité de prédiction d'un classificateur binaire. Celle-ci est représentée graphiquement comme le taux de vrais positifs en fonction du taux de faux positifs selon que le seuil, qui permet la classification, varie entre 0 et 1.

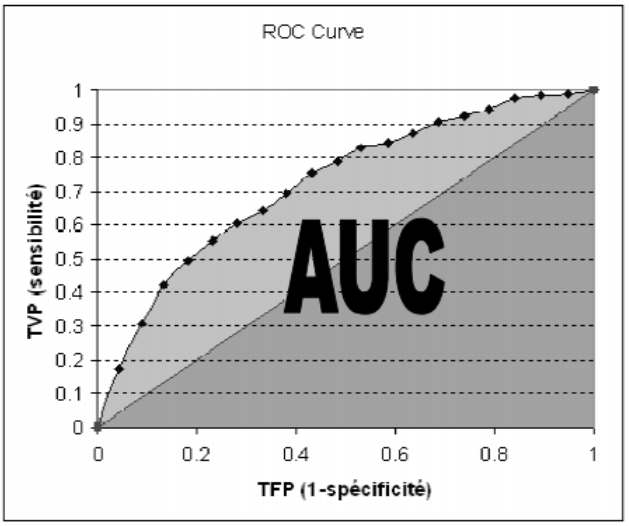


FIGURE 3.7 – Exemple de courbe ROC et de mesure AUC (Wikipedia)

En supposant que les données disponibles pour notre étude soient complètes et en nous ramenant à un problème de classification binaire, l'interprétation de la courbe ROC serait la suivante :
En abscisse, on observerait ce qu'on appelle la "spécificité", c'est à dire la probabilité de mal classer un individu sain. En ordonnée, on observerait la "sensibilité", c'est à dire la probabilité de bien classer un individu malade. En représentant la sensibilité en fonction du complémentaire à 1 de la spécificité et en intégrant entre 0 et 1 on obtiendrait l'aire sous la courbe qui correspondrait à la probabilité pour que la fonction score place un individu positif mieux qu'un individu négatif. Dans le meilleur des cas cette valeur est égale à 1 et la fonction score ne se trompe jamais.

Courbe AUC dépendante du temps

Par essence, les données de survie posent problème pour juger de la qualité des estimateurs par l'intermédiaire de la valeur AUC issue de la courbe ROC décrite précédemment car la réponse binaire de la variable de sortie peut évoluer en fonction du temps et peut être censurée. Les travaux de Heagerty et Zheng (2005) et de Chambless and Diao (2006) permettent d'adapter cette mesure à des données de survie en intégrant les concepts de sensibilité et spécificité dépendantes du temps qui permettent de définir les notions de courbes ROC et mesure AUC dépendant du temps. Chambless and Diao ont ainsi présenté une approche de calcul récursif suivant l'ordre des dates d'événements, analogue à l'approche de Kaplan-Meier pour l'estimation des fonctions de survie.

Supposons que l'on ait un score s qui permet de prédire si un patient a subi un événement avant un temps t si il dépasse un seuil K et de prédire que l'événement n'a pas eu lieu en t si le score s est inférieur à ce même seuil K. La valeur de l'AUC au temps t devient alors la probabilité que le score d'un individu ayant subit l'evenement en t soit supérieure à celui d'un individu n'ayant pas subi l'événement au temps t :

$$\begin{aligned} AUC(t) &= \mathbb{P}(s_i > s_j | D_i(t) = 1, D_j(t) = 0) \\ &= \frac{\mathbb{P}(s_i > s_j, D_i(t) = 1, D_j(t) = 0)}{\mathbb{P}(D_i(t) = 1)\mathbb{P}(D_i(t) = 0))} \end{aligned}$$

où i et j représentent des observations indépendantes. Cependant, le problème d'une estimation directe de $AUC(t)$ et de façon similaire pour la sensibilité et la spécificité dépendantes du temps, est que l'on ne connait pas le statut à t d'une personne censurée avant t. Ainsi, Chambless et Diao ont raisonné de manière equivalente à la méthode de Kaplan-Meier en divisant l'intervalle de temps en autant de sous intervalles que le nombre d'événements n, $t_1 < t_2 < ... < t_n$ alors pour $1 \leq m \leq n$ l'AUC au temps t_m devient :

$$\sum_{k \leq m} \gamma_k \alpha_k (1 - \alpha(t_k)(1 - \alpha(t_k)) S(t_{k-1})^2 - \sum_{k \leq m} \tau_k \alpha(t_k) x \frac{(1 - S(t_{k-1}))S(t_{k-1})}{(S(t_m))(1 - S(t_m))}$$

où S et α sont les fonctions de survie et d'incidence et

$$\begin{aligned} \gamma_k &= \mathbb{P}(s_i > s_j | D_i(t_k) = 1, D_i(t_{k-1}) = 0, D_j(t_k) = 0) \\ \tau_k &= \mathbb{P}(s_i > s_j | D_i(t_{k-1}) = 1, D_j(t_{k-1}) = 0, D_j(t_k) = 1) \end{aligned}$$

Les auteurs démontrent alors que la sensibilité et la spécificité dépendantes du temps au seuil K s'expriment alors ainsi :

$$sens(t_m, K) = \mathbb{P}(s_i > K | D_i(t) = 1) = \frac{\sum_{k \leq m} \rho_k(K) \alpha(t_k) S(t_{k-1})}{1 - S(t_m)}$$

$$spec(t_m, K) = \mathbb{P}(s_i > K | D_i(t) = 0) = \frac{(\mathbb{P}(s_i > K) - \sum_{k \leq m} (1 - \rho_k(K) \alpha(t_k) S(t_{k-1}))}{S(t_m)}$$

où $\rho_k(K) = \mathbb{P}(s_i > K | D_i(t_k) = 1, D_i(t_{k-1}) = 0)$

Déviance

Dans le cadre d'un problème faisant intervenir des calculs de maximum de vraisemblance comme c'est le cas pour la régression de Cox, une mesure naturelle s'impose, la déviance. Celle-ci étant en effet directement calculée à partir de la vraisemblance.
Nous rappelons la définition et la signification de la déviance :

$$Deviance = -2 * (L_m - L_s)$$

L_m représente la valeur maximal de la vraisemblance du modèle étudié et L_s représente la vraisemblance du modèle saturé, c'est à dire le modèle le plus complexe

La technique du maximum de vraisemblance implique que la loi prédéfinie pour expliquer la variable de sortie à partir des données soit correctement choisie. Le fait de sélectionner le paramètre de régularisation en fonction d'une mesure basée sur le calcul de vraisemblance pourrait renforcer le biais engendré par l'erreur de modèle. C'est pourquoi il est souvent intéressant de s'intéresser à des mesures fondée sur la comparaison directe des sorties simulées avec les sorties attendues..

3.2 Étude empirique - Modélisation sur jeux de données simulées

3.2.1 Présentation

Dans cette partie, des jeux de données ont été simulés afin d'étudier les propriétés statistiques de différents types de pénalisations ainsi que les concepts présentés lors de l'introduction des méthodes d'apprentissage. Nous tâcherons ainsi de décrire la qualité de prédiction de chacune d'elles, la justesse de leur sélection de variables, leurs propriétés de convergence ainsi que leur vitesse d'exécution. En simulant convenablement les jeux de données, les résultats obtenus pour chaque type de pénalisation pourrons en effet être comparés avec les résultats attendus. Pour mesurer les effets précités, nous ferons ainsi varier la taille des bases, leurs dimensions ou encore les liens de corrélations pré-établis entre covariables.

Pour l'ensemble des simulations, nous allons simuler des jeux de données censurées dont la variable d'intérêt vérifie la relation de Cox :

$$\alpha(t|\mathbf{Z}_i(t)) = \alpha_0(t) exp^{\mathbf{Z}_i^T \beta}$$

où :

- Nous prenons pour la fonction de hasard de base $\alpha_0(t) = 2.t$ afin d'avoir une expression simple et dépendante du temps (le facteur 2 a été choisi pour qu'il se simplifie lors de l'intégration pour le calcul de la loi de survie). Ce choix revient à considérer qu'en l'absence de l'effet des autres covariables, le taux d'incidence croit linéairement au cours du temps ;
- Pour le vecteur des covariables, nous prenons pour chaque individu i une valeur aléatoire suivant une loi uniforme sur l'intervalle [-1 ;1] : $\forall\ i \in [\![1; n]\!]\ \mathbf{Z}_i \sim \mathcal{U}[-1; 1]$
  Pour certaines simulations, nous pourrons imposer des corrélations plus ou moins fortes entre certaines variables afin de mesurer certains effets ;
- β le vecteur des coefficients mesurant l'influence des covariables sur l'intensité sera spécifié pour chaque simulation. Le fait de fixer certains coefficients à zero nous permettra par ailleurs de mesurer la qualité de sélection des estimateurs puisque cela signifie d'imposer que les variables associées n'ont aucune influence sur la variable de sortie. Un bon estimateur devra alors les exclure. Ayant fixé nous même les covariables influentes pour décrire la variable d'intéret, nous pouvons alors paramétrer le modèle Oracle qui est le modèle seulement composé des covariables associées à des coefficients non nuls ;

Nous rappelons la forme de la loi de survie en fonction du temps pour un individu i :

$$S(t, \mathbf{Z}_i) = exp[-\int_0^t \alpha(s, \mathbf{Z}_i) ds]$$

L'intensité de la loi de survie ayant été choisie de la forme $\alpha_0(t) = 2.t$, on obtient alors :

$$S(t, \mathbf{Z}_i) = exp[-\int_0^t 2s.exp^{\mathbf{Z}_i^T \beta} ds]$$

$$S(t, \mathbf{Z}_i) = exp(-t^2.e^{\mathbf{Z}_i^T \beta})$$

Pour simuler les temps T_i de survenance pour chaque individu i on utilise la méthode de Monte-Carlo en inversant la fonction de survie conditionnelle et en simulant des variables aléatoires uniformes $Y_i \in \mathcal{U}[0; 1]$

On obtient alors pour chaque individu i :

$$T_i = \sqrt{-exp^{-\mathbf{Z}_i^T \beta} log(Y_i)}$$

(remarque : $log(Y_i) < 0\ \forall\ Y_i \in [0, 1]$ et $-exp^{-\mathbf{Z}_i^T \beta} < 0$ car la fonction exponentielle est à valeur dans $\mathbb{R}^+$)

Nous imposons par ailleurs un taux de censure τ. On simule les temps de censure par une loi exponentielle $\varepsilon(\frac{1}{\tau})$ et on choisira τ afin d'obtenir le taux de censure souhaité.

On donne ci-après la courbe de la fonction de survie simulée ainsi que le graphe de ses résidus.

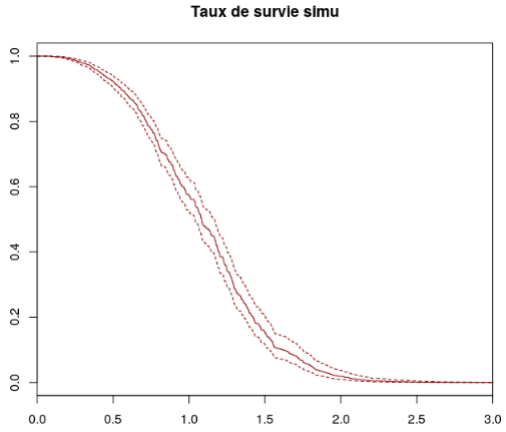


FIGURE 3.8 – Fonction de survie simulée

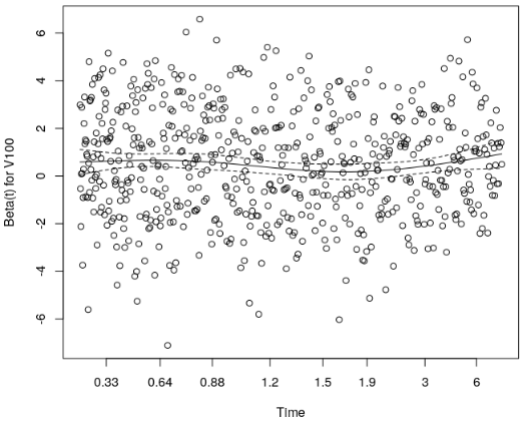


FIGURE 3.9 – Résidus de la fonction de survie simulée

L'hypothèse des risques proportionnels se vérifie simplement par construction de la variable de sortie. La forme de la répartition des résidus autour de l'axe temporel la confirme.

3.2.2 Simulation de référence

Une première simulation est menée avec des conditions de références. L'influence de chacun de ces paramètres sera analysée dans les sections suivantes.

- Le nombre de covariables p est fixé à 100 ;
- Le nombre d'observations n est fixé de manière à être dans le cas de la moyenne dimension où p=$\sqrt{n}$, n est ainsi fixé à n = 10 000. En effet, augmenter (baisser) le nombre d'observations est équivalent à baisser (augmenter) le nombre de covariables. On parle de grande dimension lorsque le nombre de covariable p et le nombre d'observations n sont proches ;
- Le taux de censure est fixé à τ=0% ;
- Les covariables sont identiquement distribuées ($Y_i \in \mathcal{U}[-1;1]$) et indépendantes deux à deux ;
- La mesure de prédiction prise pour optimiser le paramètre de régression est la déviance ;

- La méthode de ré-échantillonnage sélectionnée est la 10-folds validation croisée. Le paramètre k=10 est le plus souvent rencontré (voir en annexe pour une étude de sensibilité au nombre de subdivisions).

Simulons tout d'abord la matrice de taille 100 x 10 000 qui représente les observations des 100 covariables pour 10 000 individus. Chaque covariable est tirée aléatoirement entre -1 et 1.

```
             [,1]         [,2]         [,3]        ...        [,98]         [,99]         [,100]
[1,]   0.7354253707 -8.416804e-01 -0.9569972851   ...    -0.7394540370  3.447313e-01 -8.877249e-03
[2,]   0.7605394167 -7.488734e-01 -0.4661058718   ...    -0.9425640055  1.064807e-01  6.394265e-01
[3,]   0.6343670487 -8.153919e-01 -0.9788567200   ...     0.5132720312  8.843121e-02 -6.744135e-01
[4,]   0.1045835065  6.186364e-01  0.2097916119   ...     0.0554244323 -1.334036e-01 -1.278277e-01
[5,]  -0.2781742350  9.196883e-01 -0.7275017872   ...     0.9588664463  5.116883e-01 -8.909493e-01
[6,]  -0.0781335379 -9.181527e-01  0.9680496519   ...     0.8251434364 -7.870437e-01  8.394683e-01
[7,]  -0.7134180130 -9.336514e-01  0.6512429598   ...    -0.1335316161  1.160541e-01 -3.964417e-01
[8,]   0.6516658366 -1.211254e-01 -0.4114049864   ...    -0.2408059598  2.425775e-01 -6.674863e-01
[9,]   0.9545955053 -2.482859e-01 -0.6517490665   ...     0.6784048327  6.183033e-01  9.217066e-01
[10,]  0.6105476646  3.815215e-01  0.7448392869   ...    -0.4172092350 -9.202700e-02  7.056895e-01
           ...          ...          ...         ...            ...           ...           ...
[10000,] -0.8596901069  9.922216e-01  5.667664e-01 ...    -5.445491e-01  0.7695885920  8.674432e-01
```

FIGURE 3.10 – Simulation des 100 covariables des 10 000 individus

Nous fixons par ailleurs le vecteur β qui pondère l'influence de chaque covariable par rapport à la variable de sortie, qui est ici la durée avant que l'événement simulé ne se produise. Plus la valeur absolue d'un coefficient est élevée, plus la covariable associée contribue (signe positif) ou empêche (signe négatif) l'événement de se produire. Un coefficient nul implique que la covariable associée n'a pas d'influence sur l'apparition du phénomène simulé.

Nous choisissons les valeurs suivantes pour les 100 covariables :

- les 20 premières covariables ont un poids de -0.5,
- les 20 suivantes un poids de -0.1,
- les 20 suivantes un poids nul,
- les 20 suivantes un poids de 0.1,
- les 20 dernières un poids de 0.5.

Un bon modèle devrait prédire des coefficients proches de ceux ainsi pré-fixés et devraient également rejeter les 20 covariables de coefficients nuls.

Régression LASSO

Le graphe suivant montre le chemin pris pour chaque coefficient en fonction de la valeur du paramètre de régularisation λ pour la pénalisation LASSO.

$$\{\hat{\beta}(\lambda) \in [0, +\infty[\text{ où } \hat{\beta}(\lambda) = argmin_{\beta} \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - \lambda . \sum_{j=1}^p |\beta_j|\}$$

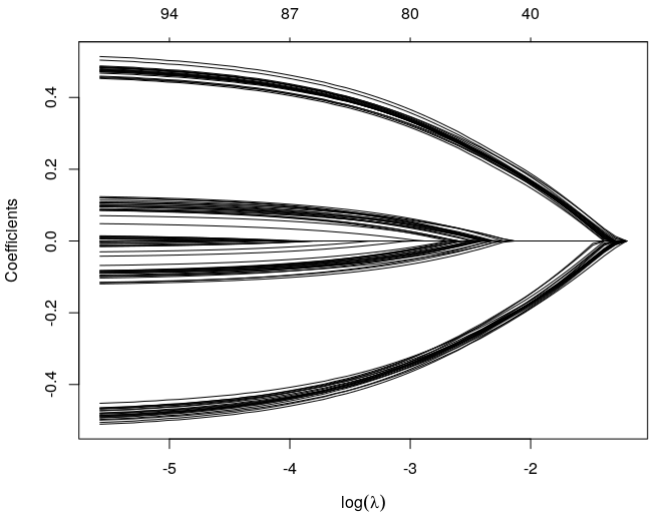


FIGURE 3.11 – Chemin des β

On note tout d'abord que le graphe des chemins des β est symétrique (à l'image des coefficients choisis pour la simulation), ceci indique que l'influence en termes de vraisemblance d'une covariable sur la variable de sortie ne dépend pas de son signe.
On constate que pour un λ suffisamment fort (ici pour $\lambda > 10^{-1} = 0.1$), tous les coefficients sont nuls. A contrario, pour un λ suffisamment faible (ici pour $\lambda < 10^{-6} = 0.000001$) alors tous les coefficients sont non nuls et convergent vers les valeurs pré-fixées a priori.
On identifie bien les 5 groupes de coefficients qui tendent respectivement vers -0.5, -0.1, 0, 0.1 et 0.5.
On remarque par ailleurs, que plus les coefficients sont élevés plus ils se détachent rapidement de l'axe des abscisses. En effet, à partir de $\lambda = 10^{-1} = 0.1$, les coefficients relatifs aux covariables dont les β ont une valeur absolue préfixée à 0.5 se détachent. Suivent ensuite les coefficients relatifs aux covariables dont les β ont une valeur absolue préfixée à 0.1 (pour $\lambda = 10^{-2.5} \simeq 0.003$). Cependant, on remarque que le groupe de coefficients censé être nuls quitte également l'axe des abscisses à partir de $\lambda = 10^{-1} = 0.001$.
Au regard de ce graphe, on souhaiterait donc fixer le λ le plus petit possible (après que les coefficients relatifs aux covariables dont les β ont une valeur absolue préfixée à 0.1 et 0.5 ne se détachent de l'axe, en lisant de droite à gauche) afin de se rapprocher le plus possible des valeurs attendues mais plus grand que la valeur critique à partir de laquelle le groupe de coefficients censé être nuls quitte également l'axe des abscisses de manière à pouvoir effectuer une sélection de variable. Le graphe suivant illustre que le choix est optimal en termes de sélection et de vraissemblance pour un paramètre de régression qui est proche de $\lambda = 10^{-3.2} = 0.00063$

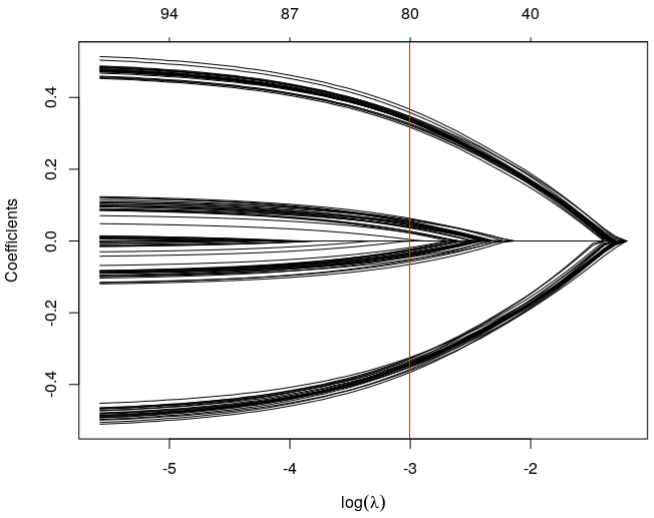


FIGURE 3.12 – Choix de la pénalisation optimale

En effet pour cette valeur du paramètre de régression, le nombre de covariables rejetés est proche de 20 et les coefficients sont proches des éléments attendus. On remarque cependant que pour un paramètre de pénalisation qui tend vers zero, on se rapproche des coefficients attendus mais on perd le caractère sélectif du modèle. Notons que l'on aperçoit bien au regard de ce graphe le compromis entre le biais et la parcimonie. Plus il y a de covariables dans le modèle, plus le modèle est réaliste mais moins il est parcimonieux.

Le graphe suivant indique l'évolution de la déviance en fonction du paramètre de régularisation.

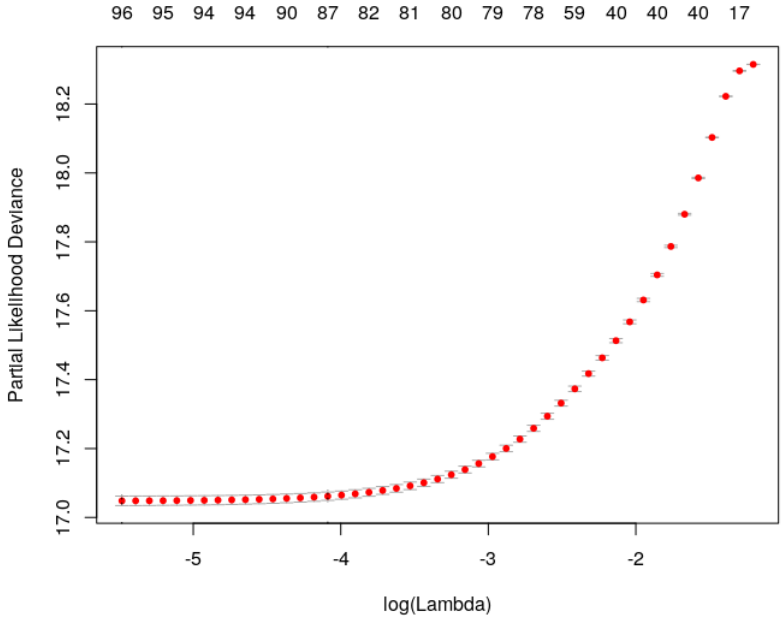


FIGURE 3.13 – Évolution de la déviance du modèle en fonction de la pénalisation

Comme attendu, plus le modèle est pénalisé moins le nombre de variables retenues (en haut sur le graphe) est grand et plus la déviance est forte. On remarque par ailleurs que le graphe se décompose en deux zones :

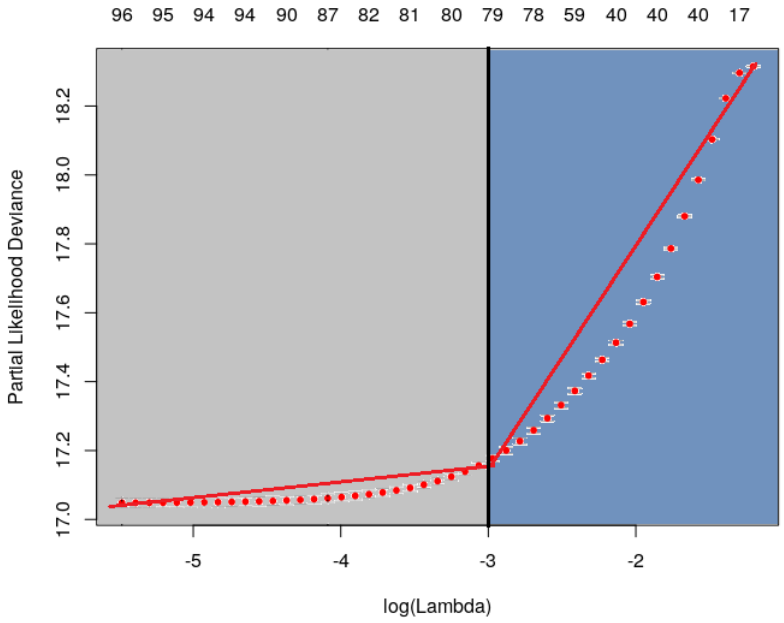


FIGURE 3.14 – Évolution de la déviance du modèle en fonction de la pénalisation

Dans la partie bleue, chaque variable ajoutée au modèle tend à faire baisser sa déviance de façon significative. Sans surprise, la pente de la déviance est forte jusqu'au lambda qui correspond à un nombre de variables sélectionnées proche de 80.
Dans la partie grise, l'ajout de variables a très peu d'influence sur la déviance. Les variables entrant dans le modèles étant en effet celles dont les coefficients avaient été fixés à 0. Théoriquement, la pente devrait être nulle puisque ces variables n'ont pas d'effet sur la variable de sortie. En pratique, on constate une légère pente dans la zone grise. Celle-ci peut s'expliquer à cause de variables significatives qui n'auraient pas été préalablement sélectionnées dans les 80 premières ou bien de variables théoriquement non significatives mais qui présentent en pratique un léger lien avec la variable de sortie. En augmentant la taille de la base d'apprentissage sur laquelle on exerce le modèle, on diminue mécaniquement cet effet.

Pour choisir le λ optimal, une première méthode consiste à repérer le point d'inflexion sur la courbe et choisir le paramètre correspondant.
R.Tibshirani propose de choisir le plus grand λ qui s'écarte d'un écart type du λ minimum. Le graphe suivant (effectué sur des données arbitraires) illustre comment trouver ce $\lambda_{1ecarttype}$.

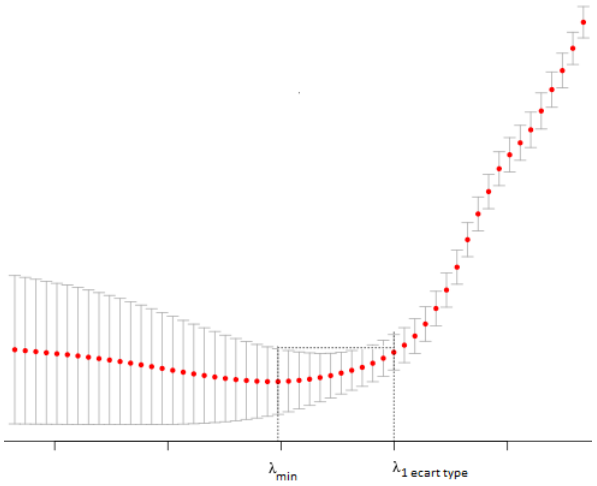


FIGURE 3.15 – Plus grand λ qui s'écarte d'un écart type du λ minimum

En revenant à la simulation, on obtient finalement $log(\lambda) = -3.8$. Avec ce critère, on est proche du $log(\lambda) = -3..2$ choisi graphiquement à la figure 3.12. Le choix du critère par cette méthode n'est néanmoins pas parfait.

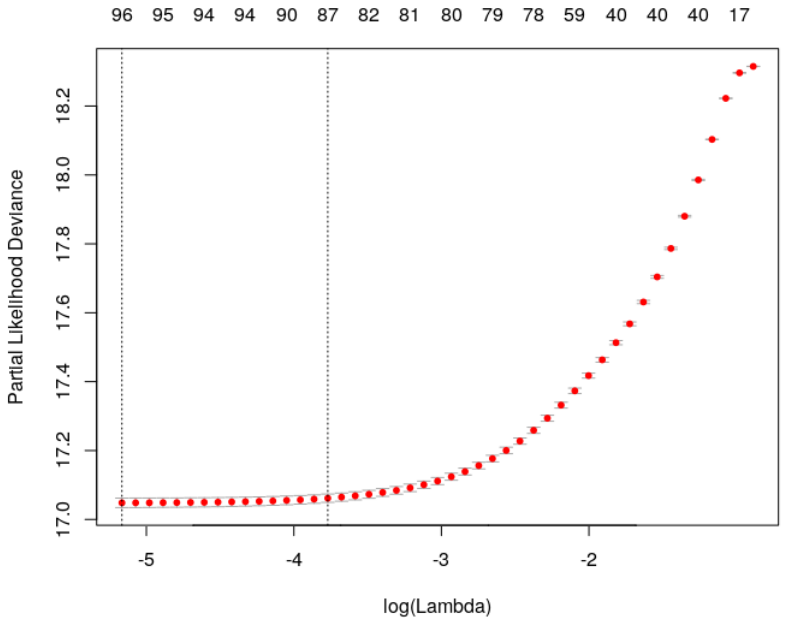


FIGURE 3.16 – Paramètre de régularisation optimal déterminé par 10-folds validation croisée

Les coefficients obtenus pour cette valeur de pénalisation sont exposés ci-après :

V1	-0.4317152414	V21	-0.0857985456	V41	0.0007954711	V61	0.0757011231	V81	0.4318619948
V2	-0.4221259374	V22	-0.0301490141	V42	.	V62	0.0860253244	V82	0.4580169565
V3	-0.4591506384	V23	-0.0693809820	V43	.	V63	0.0955454753	V83	0.4331190782
V4	-0.4510065550	V24	-0.0744464348	V44	.	V64	0.0690913357	V84	0.4280257063
V5	-0.4384794036	V25	-0.0980592909	V45	.	V65	0.0662010243	V85	0.4296070967
V6	-0.4401048048	V26	-0.0656045855	V46	.	V66	0.0880727343	V86	0.4331588499
V7	-0.4561420079	V27	-0.0718893570	V47	.	V67	0.0773784986	V87	0.4287590393
V8	-0.4408852103	V28	-0.0973017349	V48	0.0052974183	V68	0.0351671999	V88	0.4431158127
V9	-0.4355097883	V29	-0.0732320036	V49	0.0226311343	V69	0.0989147426	V89	0.4302120022
V10	-0.4268532451	V30	-0.0796745904	V50	.	V70	0.0836828131	V90	0.4142768667
V11	-0.4275992632	V31	-0.1035505059	V51	.	V71	0.0798714900	V91	0.4396430669
V12	-0.4230793211	V32	-0.0685303406	V52	.	V72	0.0808451492	V92	0.4357004289
V13	-0.4429162438	V33	-0.0663506764	V53	.	V73	0.1010751906	V93	0.4402133407
V14	-0.4263521541	V34	-0.0771663466	V54	-0.0029874163	V74	0.0567201835	V94	0.4392146064
V15	-0.4426841057	V35	-0.0533690610	V55	-0.0126621925	V75	0.0814917602	V95	0.4149033893
V16	-0.4652070701	V36	-0.0850827988	V56	.	V76	0.0914624737	V96	0.4407594384
V17	-0.4124249719	V37	-0.0770311246	V57	.	V77	0.1044155819	V97	0.4673213715
V18	-0.4436056555	V38	-0.0685172384	V58	.	V78	0.0737435926	V98	0.4258434499
V19	-0.4465545371	V39	-0.0806820227	V59	.	V79	0.0674915772	V99	0.4101128797
V20	-0.4458684899	V40	-0.0642667559	V60	.	V80	0.0955123306	V100	0.4173144591

cible -0.5 | cible -0.1 | cible 0 | cible 0.1 | cible 0.5

FIGURE 3.17 – Coefficients obtenus pour la simulation initiale de référence

Régression classique avec selection AIC

Comparons par rapport au modèle de Cox appliqué sur les variables issues d'une procédure de selection AIC stepwise.

V1	-0.48919090	V21	-0.12656120	V41	.	V61	0.09560826	V81	0.47870615
V2	-0.50851023	V22	-0.12921045	V42	.	V62	0.09485571	V82	0.50899092
V3	-0.48611379	V23	-0.07482751	V43	.	V63	0.09554966	V83	0.51109313
V4	-0.46857647	V24	-0.09602093	V44	.	V64	0.15165913	V84	0.51252773
V5	-0.49134419	V25	-0.09040761	V45	-0.03133828	V65	0.12032499	V85	0.53031243
V6	-0.47770422	V26	-0.11606754	V46	-0.02776696	V66	0.11254265	V86	0.49863248
V7	-0.50024793	V27	-0.08576700	V47	.	V67	0.09974605	V87	0.52656615
V8	-0.45889362	V28	-0.06418820	V48	.	V68	0.11304068	V88	0.47809458
V9	-0.50754351	V29	-0.11112792	V49	-0.04474638	V69	0.08560631	V89	0.51152015
V10	-0.51180440	V30	-0.09756558	V50	.	V70	0.08196750	V90	0.48497594
V11	-0.50600938	V31	-0.08530975	V51	.	V71	0.09119195	V91	0.51883174
V12	-0.50707259	V32	-0.08890354	V52	-0.01567894	V72	0.09723145	V92	0.51899265
V13	-0.51832220	V33	-0.07896948	V53	.	V73	0.07878486	V93	0.51962182
V14	-0.51200628	V34	-0.09542747	V54	.	V74	0.09516939	V94	0.51634601
V15	-0.46380952	V35	-0.08676203	V55	.	V75	0.12293075	V95	0.51317595
V16	-0.52174964	V36	-0.08711788	V56	-0.02989279	V76	0.15794190	V96	0.52540383
V17	-0.48899754	V37	-0.09945395	V57	.	V77	0.12386012	V97	0.51373314
V18	-0.53061433	V38	-0.10176582	V58	.	V78	0.10991054	V98	0.50350694
V19	-0.49618074	V39	-0.06394178	V59	.	V79	0.11394357	V99	0.51055966
V20	-0.49796588	V40	-0.07286530	V60	.	V80	0.12728077	V100	0.49370500

cible -0.5 | cible -0.1 | cible 0 | cible 0.1 | cible 0.5

FIGURE 3.18 – Coefficients obtenus pour la simulation initiale de référence avec la méthode AIC

En termes de sélection, les deux méthodes sont équivalentes puisqu'elles sélectionnent 85 variables dont 5 variables non-souhaitées.
En termes de justesse des coefficients la méthode AIC est plus performante. En effet, en calculant la moyenne des erreurs quadratiques des deux ensembles de coefficients, on obtient respectivement $3.5 * 10^{-5}$ et $4.2 * 10^{-3}$.
En termes de temps de calcul, la méthode AIC met $4min$ $40sec$ à tourner alors que la méthode LASSO prend $11sec$.

Pénalisation Adaptive-LASSO

La méthode Adaptive-Lasso va être testée en effectuant une pondération pour chacun des différents coefficients de la régression à partir des coefficients obtenus pour le LASSO. La pénalisation Adaptive-LASSO s'écrit :

$$p(\beta, \lambda) = \lambda . \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|$$

Le vecteur de poids est choisi sous la forme suivante :

$$\hat{\omega}_j = (|\hat{\beta}_j(LASSO)| + \frac{1}{n})^{-\gamma}$$

Où n représente le nombre de covariables. Le paramètre de lissage γ est choisi égal à $0,5$.

On obtient les résultats suivants :

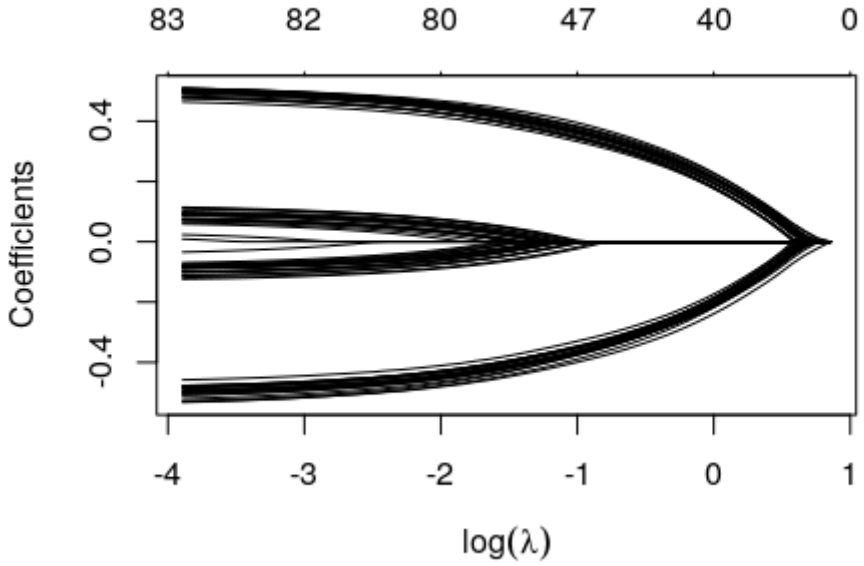


FIGURE 3.19 – Chemin des β

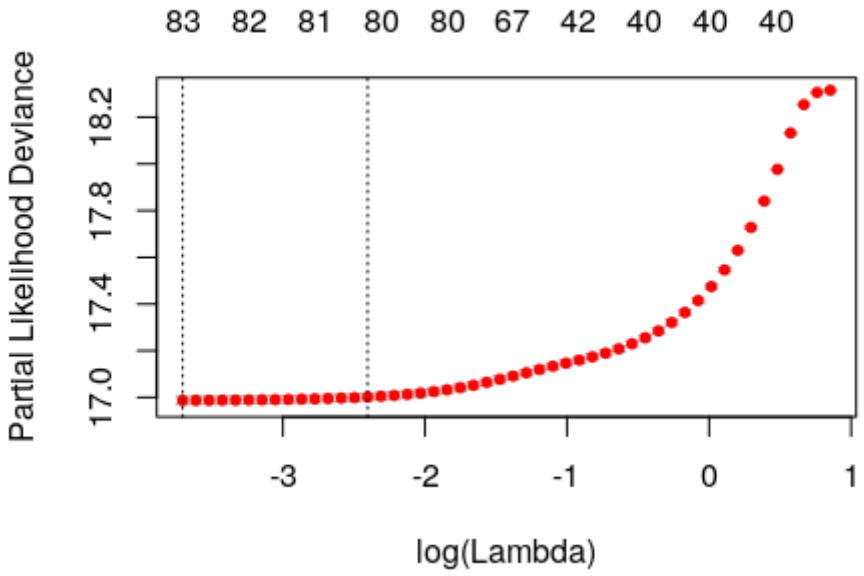


FIGURE 3.20 – Évolution de la déviance du modèle en fonction de la pénalisation

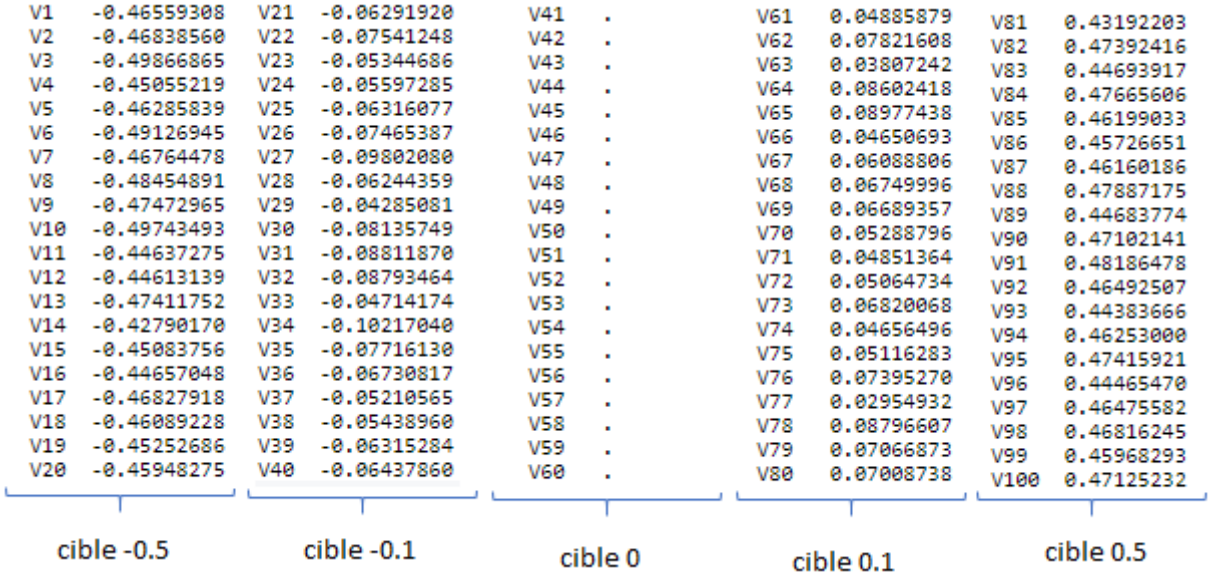


FIGURE 3.21 – Coefficients obtenus

En termes de sélection, la méthode Adaptative Lasso est de loin la plus performante puisque elle sélectionne exactement les 90 variables souhaitées.
En termes de justesse des coefficients la méthode AIC est cependant plus performante. En effet, la moyenne des erreurs quadratiques des coefficients de la méthode Adaptative Lasso est $1.2 * 10^{-3}$ (on obtenait respectivement $3.5 * 10^{-5}$ et $4.2 * 10^{-3}$ pour la méthode AIC et la méthode LASSO).
En termes de temps de calcul, la méthode Adaptative LASSO est de l'ordre de 20. Elle est deux fois plus longue que le LASSO car elle nécessite d'exécuter les deux algorithmes à tour de rôle.

3.2.3 Univers corrélé

Reprenons la même simulation que la simulation de référence mais ajoutons une corrélation entre certaines variables. Les paramètres de la simulations deviennent :

- Le nombre de covariables p est fixé à 100 ;
- Le nombre d'observations n est fixé à n = 10 000 ;
- Le taux de censure est fixé à τ=0% ;
- Les covariables sont identiquement distribuées ($Y_i \in \mathcal{U}[-1;1]$) mais ne sont plus indépendantes deux à deux. On impose la relation suivante à 4 groupes de variables :

$$Y_i = \frac{Y_{i+1} + Y_{i+2}}{2}$$

  pour $i = 1, 20, 60, 80$
- La mesure de prédiction prise pour optimiser le paramètre de régression est la déviance ;
- Nous utilisons une 10-folds validation croisée.

Simulons la matrice de taille 100 x 10 000 qui représente les observations des 100 covariables pour 10 000 individus. Chaque covariable est tirée aléatoirement entre -1 et 1. La figure suivante indique la matrice de corrélation des 100 variables

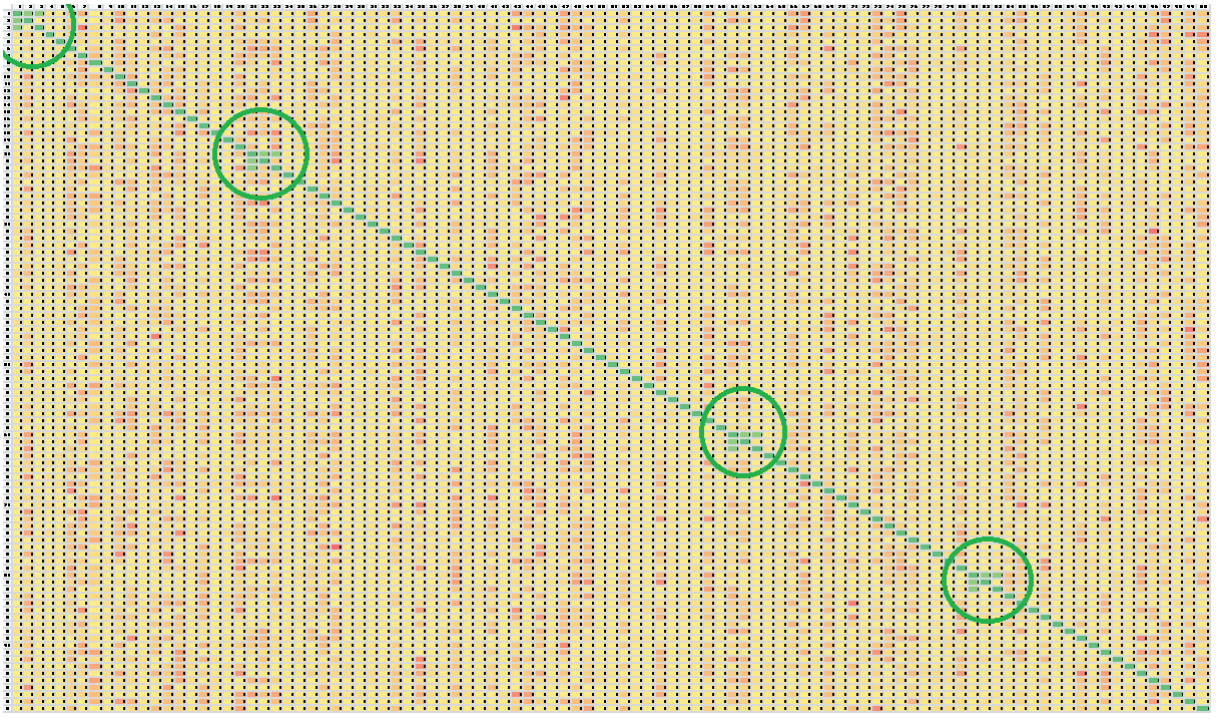
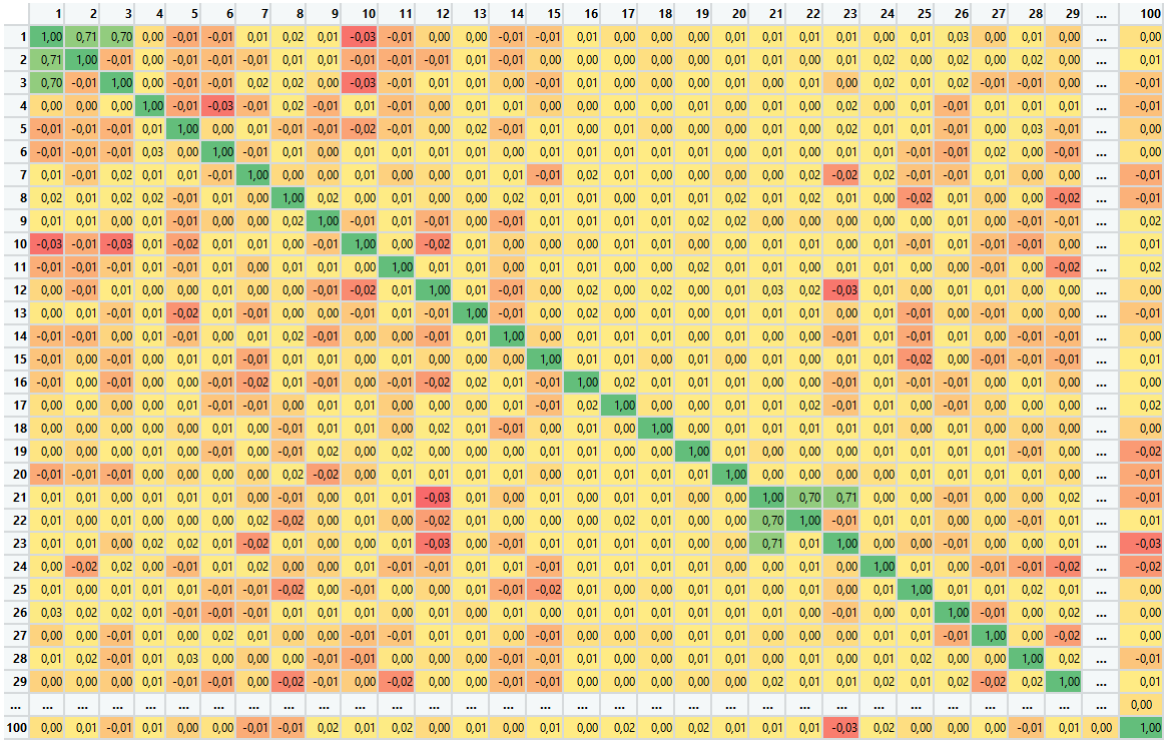


FIGURE 3.22 – Matrice de corrélation des 100 variables

En zoomant sur les 30 premières variables, on observe que les groupes de variables corrélées ont une corrélation proche de 0.7 là où les autres variables sont quasi-indépendantes car leur corrélation est proche de 1% en valeur absolue.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	...	100
1	1,00	0,71	0,70	0,00	-0,01	-0,01	0,01	0,02	0,01	-0,03	-0,01	0,00	0,00	-0,01	-0,01	0,01	0,00	0,00	0,00	0,01	0,01	0,01	0,01	0,00	0,01	0,03	0,00	0,01	0,00	...	0,00
2	0,71	1,00	-0,01	0,00	-0,01	-0,01	-0,01	0,01	0,01	-0,01	-0,01	-0,01	0,01	-0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,01	0,02	0,00	0,02	0,00	0,02	0,00	...	0,01
3	0,70	-0,01	1,00	0,00	-0,01	-0,01	0,02	0,02	0,00	-0,03	-0,01	0,01	0,01	0,00	-0,01	0,01	0,00	0,00	0,00	0,01	0,00	0,01	0,00	0,02	0,01	0,02	-0,01	-0,01	0,00	...	-0,01
4	0,00	0,00	0,00	1,00	-0,01	-0,03	-0,01	0,02	-0,01	0,01	-0,01	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,01	0,00	0,01	0,00	0,02	0,00	0,01	-0,01	0,01	0,01	0,01	...	-0,01
5	-0,01	-0,01	-0,01	0,01	1,00	0,00	0,01	-0,01	-0,01	-0,02	-0,01	0,00	0,02	-0,01	0,01	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,02	0,01	0,01	-0,01	0,00	0,03	-0,01	...	0,00
6	-0,01	-0,01	-0,01	0,03	0,00	1,00	-0,01	0,01	0,00	0,01	0,01	0,01	0,01	0,00	0,01	0,01	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,01	-0,01	-0,01	0,02	0,00	-0,01	...	0,00
7	0,01	-0,01	0,02	0,01	0,01	-0,01	1,00	0,00	0,00	0,01	0,00	0,00	0,01	0,01	-0,01	0,02	0,01	0,00	0,00	0,00	0,00	0,02	-0,02	0,02	-0,01	-0,01	0,01	0,00	0,00	...	-0,01
8	0,02	0,01	0,02	0,02	-0,01	0,01	0,00	1,00	0,02	0,00	0,01	0,00	0,00	0,02	0,01	0,01	0,00	0,01	0,01	0,02	0,01	0,02	0,01	0,00	-0,02	0,01	0,00	0,00	-0,02	...	-0,01
9	0,01	0,01	0,00	0,01	-0,01	0,00	0,00	0,02	1,00	-0,01	0,01	-0,01	0,00	-0,01	0,01	0,01	0,01	0,01	0,02	0,02	0,00	0,00	0,00	0,00	0,00	0,01	0,00	-0,01	-0,01	...	0,02
10	-0,03	-0,01	-0,03	0,01	-0,02	0,01	0,01	0,00	-0,01	1,00	0,00	-0,02	0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,01	0,01	0,00	0,01	-0,01	0,01	-0,01	-0,01	0,00	...	0,01
11	-0,01	-0,01	-0,01	0,01	-0,01	0,01	0,00	0,01	0,01	0,00	1,00	0,01	0,01	0,00	0,01	0,01	0,00	0,00	0,02	0,01	0,01	0,00	0,01	0,01	0,00	0,00	-0,01	0,00	-0,02	...	0,02
12	0,00	-0,01	0,01	0,00	0,00	0,01	0,00	0,00	-0,01	-0,02	0,01	1,00	0,01	-0,01	0,00	0,02	0,00	0,02	0,00	0,01	0,03	0,02	-0,03	0,01	0,00	0,01	0,01	0,00	0,00	...	0,00
13	0,00	0,01	-0,01	0,01	-0,02	0,01	-0,01	0,00	0,00	-0,01	0,01	-0,01	1,00	-0,01	0,00	0,02	0,00	0,01	0,00	0,01	0,01	0,01	0,00	0,01	-0,01	0,00	-0,01	0,00	0,00	...	-0,01
14	-0,01	-0,01	0,00	0,01	-0,01	0,00	0,01	0,02	-0,01	0,00	0,00	-0,01	0,01	1,00	0,00	0,01	0,01	0,01	0,00	0,01	0,00	0,00	-0,01	0,01	-0,01	0,01	0,00	-0,01	-0,01	...	0,00
15	-0,01	0,00	-0,01	0,00	0,01	0,01	-0,01	0,01	0,01	0,00	0,01	0,00	0,00	0,00	1,00	0,01	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,01	-0,02	0,00	-0,01	-0,01	-0,01	...	0,01
16	-0,01	0,00	-0,01	0,00	0,00	-0,01	-0,02	0,01	-0,01	0,00	-0,01	-0,02	0,02	0,01	-0,01	1,00	0,02	0,01	0,01	0,01	0,00	0,00	-0,01	0,01	-0,01	-0,01	0,00	0,01	0,00	...	0,00
17	0,00	0,00	0,00	0,00	0,01	-0,01	-0,01	0,00	0,01	0,01	0,00	0,00	0,00	0,01	-0,01	0,02	1,00	0,00	0,00	0,01	0,01	0,02	-0,01	0,01	0,00	-0,01	0,00	0,00	0,00	...	0,02
18	0,00	0,00	0,00	0,00	0,00	0,01	0,00	-0,01	0,01	0,01	0,00	0,02	0,01	-0,01	0,00	0,01	0,00	1,00	0,00	0,01	0,01	0,01	0,01	0,01	0,00	0,01	0,00	0,00	0,00	...	0,00
19	0,00	0,00	0,00	0,01	0,00	-0,01	0,00	-0,01	0,02	0,00	0,02	0,00	0,00	0,00	0,01	0,01	0,00	0,00	1,00	0,01	0,00	0,00	0,00	0,01	0,01	0,01	0,01	-0,01	0,00	...	-0,02
20	-0,01	-0,01	-0,01	0,00	0,00	0,00	0,00	0,02	-0,02	0,00	0,01	0,01	0,01	0,01	0,00	0,01	0,01	0,01	0,01	1,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,01	0,00	...	-0,01
21	0,01	0,01	0,00	0,01	0,01	0,01	0,00	-0,01	0,00	0,01	0,01	-0,03	0,01	0,00	0,01	0,00	0,01	0,01	0,00	0,00	1,00	0,70	0,71	0,00	0,00	-0,01	0,00	0,00	0,02	...	-0,01
22	0,01	0,00	0,01	0,00	0,00	0,00	0,02	-0,02	0,00	0,01	0,00	-0,02	0,01	0,00	0,00	0,00	0,02	0,01	0,00	0,00	0,70	1,00	-0,01	0,01	0,01	0,00	0,00	-0,01	0,01	...	0,01
23	0,01	0,01	0,00	0,02	0,02	0,01	-0,02	0,01	0,00	0,00	0,01	-0,03	0,00	-0,01	0,01	0,01	0,01	0,01	0,00	0,00	0,71	0,01	1,00	0,00	0,00	-0,01	0,00	0,00	0,01	...	-0,03
24	0,00	-0,02	0,02	0,00	-0,01	0,01	0,02	0,00	0,00	0,01	-0,01	-0,01	0,01	0,01	-0,01	0,01	0,01	0,01	0,01	0,00	0,00	0,01	0,00	1,00	0,01	0,00	-0,01	-0,01	-0,02	...	-0,02
25	0,01	0,00	0,01	0,01	0,01	-0,01	-0,01	-0,02	0,00	-0,01	0,00	0,00	0,01	-0,01	-0,02	0,01	0,00	0,00	0,01	0,01	0,00	0,01	0,00	0,01	1,00	0,01	0,01	0,02	0,01	...	0,00
26	0,03	0,02	0,02	0,01	-0,01	-0,01	-0,01	0,01	0,01	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,01	0,01	0,01	0,01	0,01	0,00	-0,01	0,00	0,01	1,00	-0,01	0,00	0,02	...	0,00
27	0,00	0,00	-0,01	0,01	0,00	0,02	0,01	0,00	0,00	-0,01	-0,01	0,01	0,01	0,00	-0,01	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,01	0,01	-0,01	1,00	0,00	-0,02	...	0,00
28	0,01	0,02	-0,01	0,01	0,03	0,00	0,00	0,00	-0,01	-0,01	0,00	0,00	0,00	-0,01	-0,01	0,01	0,00	0,00	0,01	0,01	0,00	0,01	0,00	0,01	0,02	0,00	0,00	1,00	0,02	...	-0,01
29	0,00	0,00	0,00	0,01	-0,01	-0,01	0,00	-0,02	-0,01	0,00	-0,02	0,00	0,00	-0,01	-0,01	0,00	0,00	0,00	0,00	0,00	0,02	0,01	0,01	0,02	0,01	0,02	-0,02	0,02	1,00	...	0,01
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	0,00
100	0,00	0,01	-0,01	0,01	0,00	0,00	-0,01	-0,01	0,02	0,01	0,02	0,00	0,01	0,00	0,01	0,00	0,02	0,00	0,02	0,01	0,01	0,01	-0,03	0,02	0,00	0,00	0,00	-0,01	0,01	0,00	1,00

FIGURE 3.23 – Zoom de la matrice de corrélation des 30 premières variables

Comme dans la partie précédente, nous fixons le vecteur β de la façon suivante :

- les 20 premières covariables ont un poids de -0.5,
- les 20 suivantes un poids de -0.1,
- les 20 suivantes un poids nul,
- les 20 suivantes un poids de 0.1,
- les 20 dernières un poids de 0.5.

Un bon modèle devrait prédire des coefficients proches de ceux ainsi pré-fixés et devrait rejeter seulement les 20 covariables de coefficients nuls.

Régression LASSO

Les graphes en annexe 3 montrent le chemin pris pour chaque coefficient en fonction de la valeur du paramètre de régularisation λ pour la pénalisation LASSO, l'évolution de la déviance en fonction du paramètre de pénalisation et la valeur des coefficients obtenus pour la plus grande valeur de pénalisation éloignée d'un écart type de la pénalisation minimale.

On s'aperçoit que le modèle diminue fortement le coefficient de deux des trois variables des groupes de variables et attribue au troisième coefficient un poids élevé. Dans trois des quatre groupes ou apparaissent les variables corrélées, le modèle rejette une variable. Par ailleurs, on remarque également que ces rejets se font au détriment d'autres variables non significatives. En effet, le nombre total de variables sélectionnées par le modèle est toujours de 80. On a donc cette fois 3 variables rejetées à tort (contre 0 dans le cas précédent sans corrélation) et 8 variables non-rejetées à tort (contre 5 dans le cas précédent).

Dans un contexte prédictif, le fait que le modèle concentre l'intégralité de l'information sur l'une des variables d'un groupe de variables pose problème puisque l'information relative à plusieurs variables est alors supportée par une seule variable, cela impose une volatilité plus importante que si elle était répartie sur l'ensemble des variables du groupe de variables corrélées. D'un point de vu explicatif, le fait de concentrer l'information de plusieurs variables sur une seule affecte l'ensemble du groupe de covariable ; la variable recevant l'intégralité de l'information du groupe étant systématiquement surestimée ou sous estimée. De plus les autres variables sortent du modèle alors qu'elles influent en réalité sur la variable de sortie.

Pénalisation Adaptive-LASSO

La méthode Adaptive-Lasso va être testée en effectuant une pondération pour chacun des différents coefficients de la régression à partir des coefficients obtenus pour le LASSO.Comme dans la section précédente, on choisit le vecteur de poids sous la forme suivante :

$$\hat{\omega}_j = (|\hat{\beta}_j(LASSO)| + \frac{1}{n})^{-\gamma}$$

Où n représente le nombre de covariables. Le paramètre de lissage γ est choisi égal à $0,5$.

Les graphes en annexe 4 montrent les résultats obtenus.

Comme dans le cas du LASSO, on s'aperçoit que le modèle diminue fortement le coefficient de deux des trois variables des groupes de variables et attribue au troisième coefficient un poids élevé. Cette pénalisation ne semble donc pas non plus bien adaptée à la présence de corrélations entre covariables.

Régression Ridge

Les graphes en annexe 5 montrent le chemin pris pour chaque coefficient en fonction de la valeur du paramètre de régularisation λ pour la pénalisation Ridge.
Comme attendu, la régression Ridge ne permet pas de sélectionner de variables car celle-ci ne permet pas de fixer de coefficients à zéro. On s'aperçoit cependant que les effets néfastes de la corrélation sont diminués. En effet, contrairement à la pénalisation LASSO, l'information de l'ensemble des variables d'un groupe de variables corrélées ne se concentre plus sur une seule d'entre elle.

Régression Elastic-net

Les graphes en annexe 6 montrent le chemin pris pour chaque coefficient et l'évolution de la déviance en fonction de la valeur du paramètre de régularisation λ pour la pénalisation Elastic-net pour différentes valeurs du paramètre α :

$$\hat{\beta}(\lambda) = argmin_\beta \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - \lambda(\alpha \sum_{j=1}^p \beta_j^2 + (1-\alpha) \sum_{j=1}^p |\beta_j|)\}$$
$$\{\hat{\beta}(\lambda) \in [0, +\infty[\text{ où}$$

Les coefficients obtenus par le biais de cette méthode montrent que celle-ci combine les avantages des deux types de pénalisation LASSO et Ridge et se rapprochent de ceux obtenus avec une pénalisation LASSO dans un univers non corrélé.

Régression classique avec sélection AIC

Comparons par rapport au modèle de Cox appliqué sur les variables issues d'une procédure de sélection AIC stepwise.

V1	-1,5214807	V21	-0,2836735	V41	.	V61	0,28557202	V81	1,52112486
V2	.	V22	.	V42	.	V62	.	V82	.
V3	.	V23	.	V43	.	V63	.	V83	.
V4	-0,4924069	V24	-0,0959958	V44	.	V64	0,07636043	V84	0,48952894
V5	-0,4902296	V25	-0,0624731	V45	0,03832374	V65	0,07500789	V85	0,4681035
V6	-0,4918826	V26	-0,0944808	V46	.	V66	0,11816911	V86	0,47951851
V7	-0,4939376	V27	-0,080228	V47	.	V67	0,13066679	V87	0,45058916
V8	-0,5295058	V28	-0,0907455	V48	.	V68	0,11537244	V88	0,49202521
V9	-0,5036114	V29	-0,1102526	V49	.	V69	0,11632867	V89	0,48934378
V10	-0,49159	V30	-0,1176058	V50	.	V70	0,05666613	V90	0,47316331
V11	-0,4645147	V31	-0,0912493	V51	-0,03244726	V71	0,09206072	V91	0,52447457
V12	-0,4868481	V32	-0,0738773	V52	0,03223604	V72	0,07272254	V92	0,51817266
V13	-0,5030866	V33	-0,1131651	V53	.	V73	0,10654247	V93	0,5316543
V14	-0,5313881	V34	-0,1406378	V54	.	V74	0,09988537	V94	0,51347538
V15	-0,5393824	V35	-0,0832529	V55	.	V75	0,10747701	V95	0,49912259
V16	-0,5368761	V36	-0,1139182	V56	.	V76	0,11530271	V96	0,52117535
V17	-0,5436506	V37	-0,1167065	V57	.	V77	0,12056886	V97	0,49518729
V18	-0,4870592	V38	-0,0915019	V58	.	V78	0,08747725	V98	0,47230591
V19	-0,4982169	V39	-0,1292793	V59	.	V79	0,09565027	V99	0,49866586
V20	0,50881005	V40	-0,0807632	V60	.	V80	0,09346127	V100	0,52501312

cible -0.5 | cible -0.1 | cible 0 | cible 0.1 | cible 0.5

FIGURE 3.24 – Coefficients obtenus pour la simulation initiale de référence avec la méthode AIC avec variables corrélées

En termes de sélection, la méthode AIC est plus performante, cependant celle-ci exclut les variables corrélées comme la pénalisation LASSO.
Comme dans le cas non corrélé, la justesse des coefficients est meilleure avec la méthode AIC (hormis les variables corrélées qui accumulent l'information de l'ensemble des variables avec laquelle elle est corrélée).
En termes de temps de calcul, la méthode AIC est environ 5 fois plus lente que lorsqu'il n'y avait pas de corrélation entre les variables. Elle met en effet $24min\ 48sec$ à tourner. La méthode Elastic-net s'exécute en à peine $11sec$.

3.2.4 Univers censuré et univers de petite, moyenne et grande dimension

Des études ont également été menées sur l'influence du taux de censure ainsi que celle du nombre d'observations afin d'étudier l'impact de la dimension de la base sur la qualité des estimateurs. Les résultats théoriques introduits dans la partie précédente sur l'évolution de l'erreur de prédiction en fonction de la taille des échantillons y sont étudiés, l'intérêt de la validation croisée est mis en pratique. Afin de limiter la redondance de la lecture, ces parties sont placées en annexe.

3.3 Étude pratique - Modélisation et comparaison sur la base des données du PMSI

Dans les parties précédentes, les données du PMSI ont été utilisées pour paramétrer un modèle de Cox en opérant une sélection de variables en amont de la phase de paramétrage du modèle. Cette méthode, classiquement utilisée en statistique, s'est avérée particulièrement chronophage du fait de la taille et de la dimension importante de la base d'étude. Par la suite, des techniques de régression pénalisée ont été introduites. Celles-ci ont été testées sur des bases de données simulées de manière à mesurer la qualité du processus de sélection, la justesse des estimateurs obtenus et la rapidité d'exécution de ces méthodes. Il en résulte que les estimateurs obtenus sont presque aussi performants que ceux obtenus par la procédure classique mais qu'ils ont l'avantage d'être extrêmement plus rapides à exécuter et relativement plus simples à mettre œuvre. La nature des données du PMSI étant des

pathologies médicales, il existe naturellement de fortes corrélations entre celles-ci. Dans la partie précédente, il s'est avéré que les méthodes de régression LASSO et Adaptative LASSO étaient peu performantes en présence de corrélations puisque les estimateurs résultants de ces modèles concentraient l'intégralité de l'information sur l'une des variables du groupe de variables corrélées. Ceci introduisant un biais en termes de prédictibilité mais ceci nuisant également fortement à la nature explicative des estimateurs obtenus. La régression Ridge permettait de passer outre les problèmes engendrés par la présence de corrélations entre variables mais avait par ailleurs le gros désavantage de ne pouvoir opérer de sélection de variables. Parmi les types de pénalisations introduits dans la partie précédente, celle qui semble alors la plus adéquate pour paramétrer les modèles de Cox sur les données du PMSI serait donc la régression Elastic-net.

La figure suivante donne la matrice de corrélation de l'ensemble des variables présentes dans la base de données médicales du PMSI.

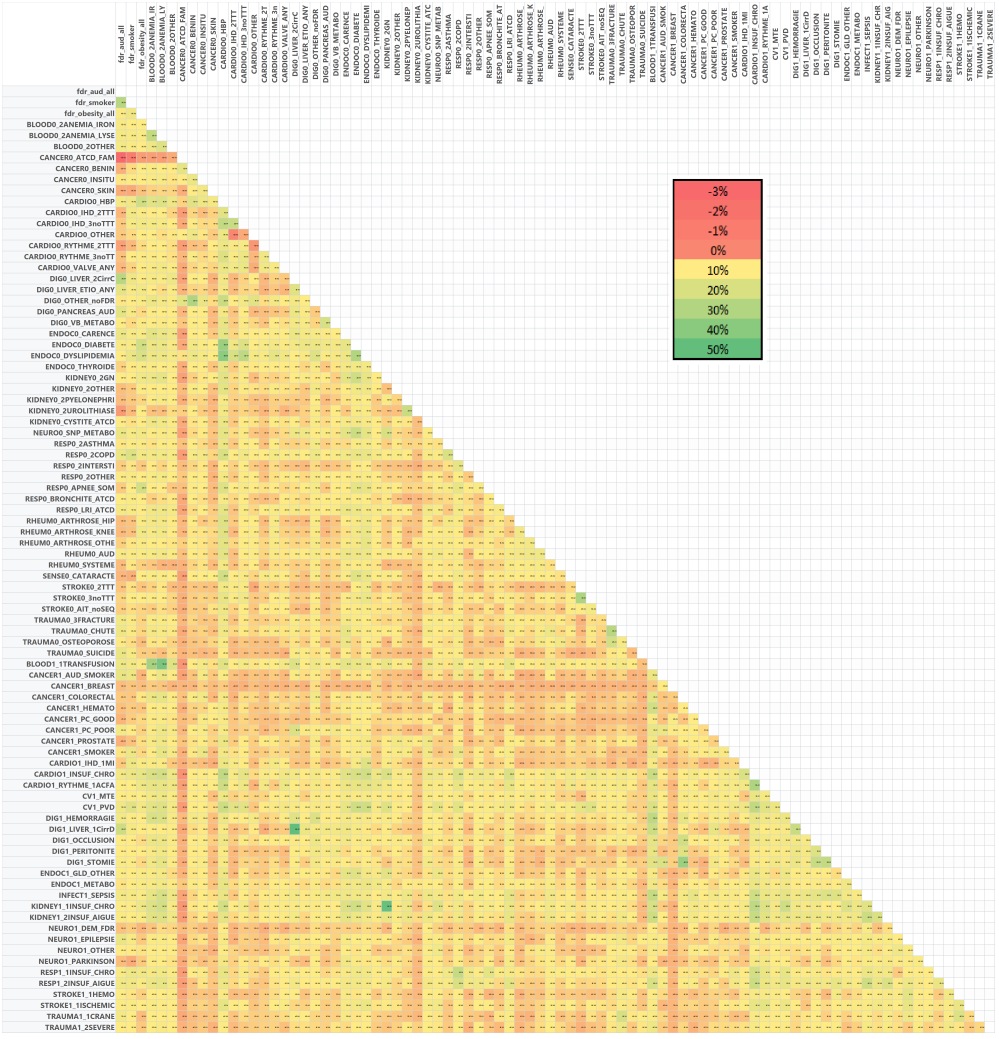


FIGURE 3.25 – Matrice de corrélation

On constate bien qu'il existe de fortes corrélations entre certains groupes de pathologies ou facteurs de risque. On note par exemple une corrélation de 30% entre les patients fumeurs et ceux caractérisés comme dépendants à l'alcool. Les pathologies "Blood1 1 transfusion" et "Blood0 2 Anemia" présentent quant à elles une corrélation de 50%. Il apparait légitime de choisir la pénalisation Elastic-net pour mener l'étude de régression.

3.3.1 Modélisation de la dépendance sur les bases des Picards et des Parisiennes à l'aide de la régression Elastic-net

Détermination du paramètre alpha de l'Elastic-Net

Les graphiques suivants indiquent l'évolution du niveau de déviance des modèles de prédiction de la Dépendance obtenus en faisant varier le paramètre α de l'Elastic net sur les populations parisiennes et picardes.

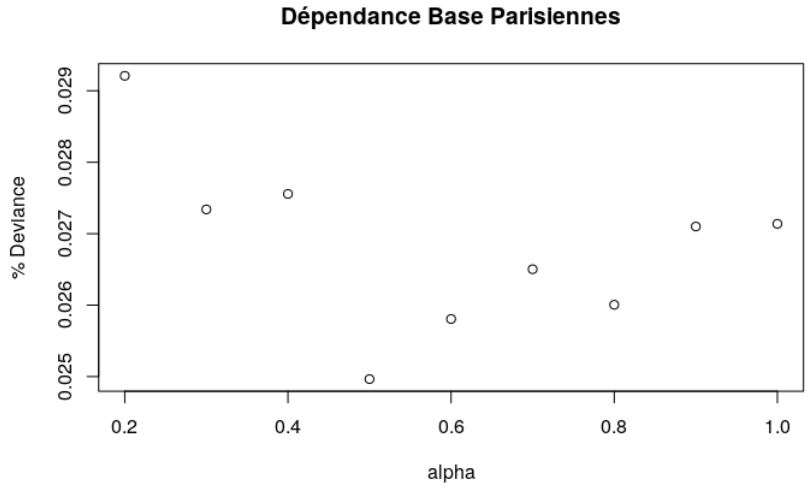


FIGURE 3.26 – Détermination du paramètre α de l'Elastic-net pour la dépendance des Parisiennes

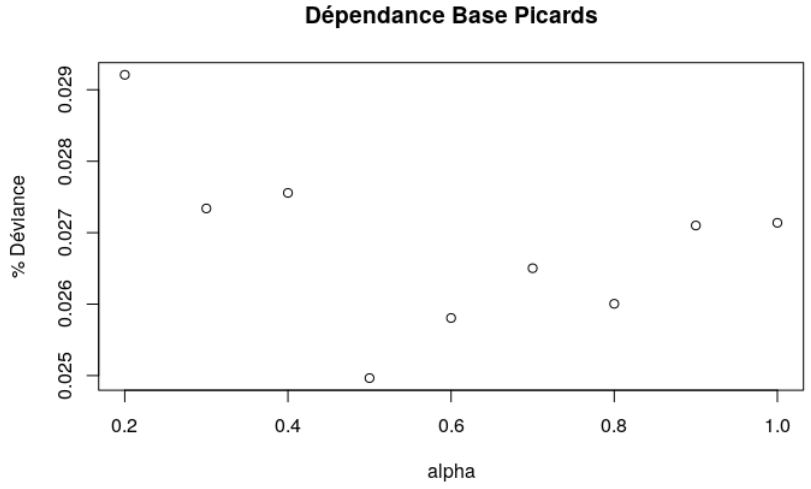


FIGURE 3.27 – Determination du paramètre α de l'Elastic-net pour la dépendance des Picards

En changeant la mesure de prédiction et en s'intéressant à l'aire sous la courbe ROC de Chambless et Diao introduite dans la section précédente on obtient le graphe suivant pour l'evolution de l'AUC de Chambless et Diao en fonction du paramètre α pour la dépendance sur la base des Parisiennes.

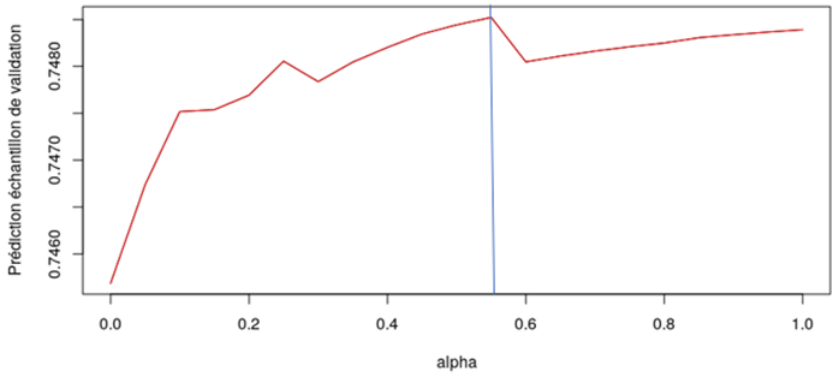


FIGURE 3.28 – Optimisation du facteur de l'Elastic-net avec la mesure AUC de Chambless et Diao sur la base des Parisiennes

On constate que le choix optimal se situe aux environs de α=0.5 pour la Dépendance sur la base des Parisiennes avec la mesure de déviance comme pour la mesure AUC de Chambless et Diao.

La pénalisation Elastic-net suivante a donc été paramétrée sur la base de données des Parisiennes et des Picards afin d'établir la relation entre l'ensemble des covariables et l'état de dépendance :

$$p(\beta, \lambda, \alpha) = \lambda(\frac{1}{2}\sum_{j=1}^{p}\beta_j^2 + \frac{1}{2}\sum_{j=1}^{p}|\beta_j|)$$

La norme l_2 contribue à effacer les effets non désirés de la corrélation alors que la norme l_1 contribue à effectuer la sélection de variable.

Détermination des chemins de régularisation

Le vraisemblance partielle pénalisée s'exprime alors de la façon suivante :

$$L_{partielle}(\beta) = \prod_{i}^{D}\frac{exp(\beta^T Z_i)}{\sum_{j\in R(T_i)} exp(\beta^T Z_j)} - \lambda(\frac{1}{2}\sum_{j=1}^{p}\beta_j^2 + \frac{1}{2}\sum_{j=1}^{p}|\beta_j|)$$

En faisant varier la valeur de la constante de régularisation, on obtient alors l'ensemble de solutions suivant :

$$\{\hat{\beta}(\lambda) \in [0, +\infty[\text{ où } \hat{\beta}(\lambda) = argmin_\beta \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j\in R(T_i)} exp(\beta^T Z_j)} - \lambda(\tfrac{1}{2}\sum_{j=1}^{p}\beta_j^2 + \tfrac{1}{2}\sum_{j=1}^{p}|\beta_j|)\}$$

Le chemin de régularisation ci-dessous est la représentation de l'ensemble des valeurs prises par chacun des coefficients de régression en fonction de la valeur du paramètre de régularisation λ.

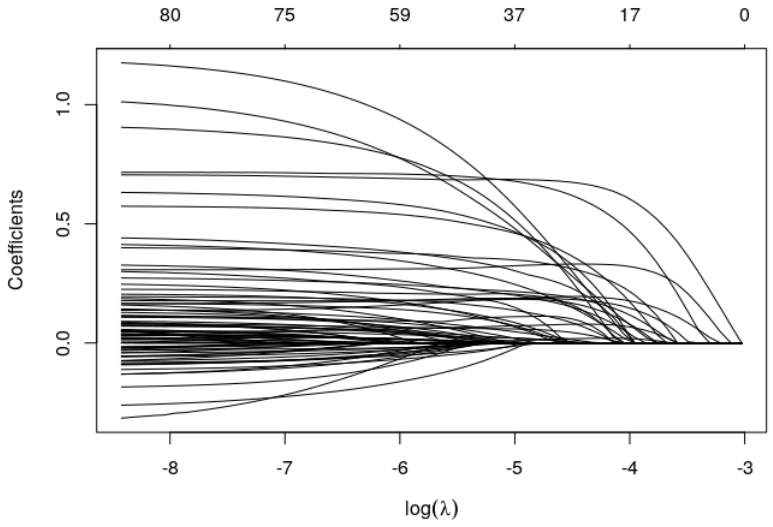


FIGURE 3.29 – Dépendance des Parisiennes

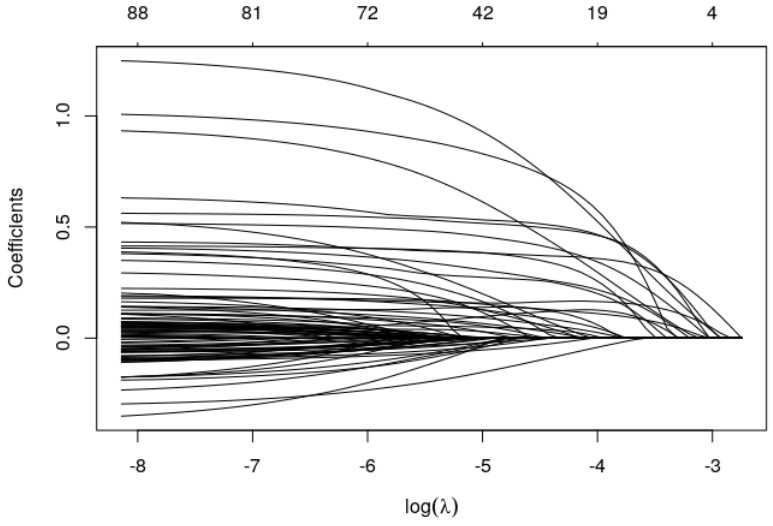


FIGURE 3.30 – Dépendance des Picards

On constate que les deux représentations ont des formes similaires. Bien qu'on ne puisse rien en conclure, des chemins de régularisations totalement différents auraient néanmoins remis en doute la stabilité de la méthode.

Détermination de l'intensité de la pénalisation par validation croisée

En utilisant une 10-folds validation croisée (se référer à l'annexe 7 pour l'étude de sensibilité sur le nombre de subdivisions) avec la déviance comme mesure de prédiction comme utilisée dans la section précédente, on obtient les graphes suivants :

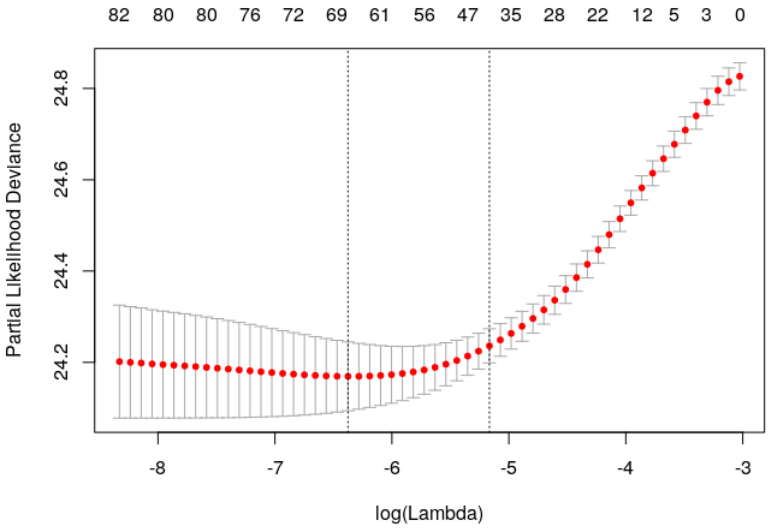


FIGURE 3.31 – Détermination du paramètre λ de l'Elastic-net pour la dépendance des Parisiennes

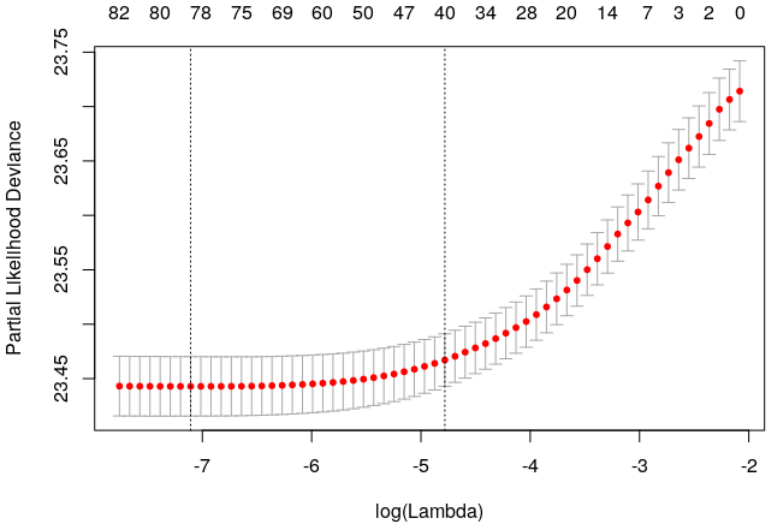


FIGURE 3.32 – Détermination du paramètre λ de l'Elastic-net pour la dépendance des Picards

La pente de la déviance est forte jusqu'à $\log(\lambda)$ proche de -6. Dans cette partie, chaque variable ajoutée au modèle tend à faire baisser sa déviance de façon significative. Pour des λ plus faibles, l'ajout de variables a peu d'influence sur la déviance. Les variables entrants dans le modèle ont donc a priori peu d'influence sur la variable de sortie. En suivant le critère du λ qui s'écarte d'un écart type du λ minimum (la ligne pointillée de droite sur le graphique), les modèles sélectionnés sur la base des Parisiennes et des Picards sont les suivants (avec pour rappel les résultats obtenus dans le cadre de la sélection avec le critère AIC stepwise) :

	Parisienne Demence		Picard Demence	
Covariable	Enet	AIC	Enet	AIC
CANCER1_SMOKER	1,90	2,15	1,55	1,77
CANCER1_PC_POOR	1,88	2,09	1,68	1,63
NEURO1_DEM_FDR	1,56	1,34	2,28	1,32
CANCER1_PC_GOOD	1,55	1,63	1,20	1,24
fdr_aud_all	1,54	3,77	2,12	4,02
fdr_smoker	1,53	3,27	1,65	2,60
DIG1_STOMIE	1,40	1,21	1,32	1,30
fdr_obesity_all	1,39	2,01	1,23	1,36
STROKE1_1HEMO	1,39	1,22	1,67	-
INFECT1_SEPSIS	1,35	1,19	1,44	1,18
CANCER1_AUD_SMOKER	1,27	1,68	1,22	1,42
CANCER1_BREAST	1,24	1,51	1,31	-
STROKE1_1ISCHEMIC	1,23	-	1,37	-
CANCER1_HEMATO	1,23	1,31	1,17	1,31
DIG1_LIVER_1CirrD	1,23	1,25	1,14	1,22
RESP1_1INSUF_CHRO	1,21	1,08	1,10	0,94
RESP1_2INSUF_AIGUE	1,20	1,11	1,11	-
BLOOD1_1TRANSFUSION	1,19	1,11	1,17	1,10
NEURO1_OTHER	1,18	0,68	1,45	-
NEURO1_EPILEPSIE	1,17	1,13	1,18	1,21
ENDOC1_GLD_OTHER	1,11	1,06	1,04	1,12
NEURO1_PARKINSON	1,11	1,22	-	-
NEURO0_SNP_METABO	1,10	1,06	1,07	1,08
CV1_MTE	1,08	1,04	1,09	1,07
KIDNEY1_2INSUF_AIGUE	1,08	1,03	-	1,11
BLOOD0_2OTHER	1,06	1,15	1,04	-
BLOOD0_2ANEMIA_LYSE	1,04	1,10	1,18	1,22
ENDOC1_METABO	1,04	1,06	1,04	-
DIG0_LIVER_2CirrC	1,04	1,05	1,01	-
cp_dipl0	1,04	1,16	1,02	1,06
CV1_PVD	1,04	-	1,01	-
DIG1_OCCLUSION	1,03	1,03	1,05	-
RHEUM0_SYSTEME	1,03	1,12	1,05	-
RESP0_2COPD	1,03	-	1,03	-
CANCER1_COLORECTAL	1,03	1,30	1,08	1,23
KIDNEY0_2PYELONEPHRI	1,03	-	-	0,88
CARDIO1_INSUF_CHRO	1,02	-	-	0,95
DIG0_LIVER_ETIO_ANY	1,02	1,10	1,00	1,05
ENDOC0_DIABETE	1,02	1,07	-	1,04
RESP0_2OTHER	1,01	1,09	1,01	-
KIDNEY0_CYSTITE_ATCD	1,01	0,98	1,12	0,91
cp_dep	1,01	1,01	0,88	-
RESP0_2INTERSTI	-	1,08	1,01	1,30
TRAUMA0_SUICIDE	-	1,08	-	-
CARDIO0_VALVE_ANY	-	1,07	-	1,04

	Parisienne Demence		Picard Demence	
Covariable	Enet	AIC	Enet	AIC
KIDNEY0_2GN	-	1,04	1,03	-
CARDIO0_IHD_2TTT	-	1,03	0,99	0,97
CANCER0_BENIN	-	0,98	0,98	0,90
BLOOD0_2ANEMIA_IRON	-	0,98	0,98	0,94
DIG0_VB_METABO	-	0,98	0,94	0,87
CANCER0_INSITU	-	0,98	1,03	-
ENDOC0_DYSLIPIDEMIA	-	0,97	0,98	0,96
ENDOC0_THYROIDE	-	0,97	0,97	0,94
STROKE0_AIT_noSEQ	-	0,97	1,03	-
RESP0_LRI_ATCD	-	0,97	1,01	-
STROKE0_3noTTT	-	0,96	0,92	-
RHEUM0_AUD	-	0,95	-	0,94
DIG0_OTHER_noFDR	-	0,93	0,98	0,93
KIDNEY1_1INSUF_CHRO	-	0,93	1,00	-
RHEUM0_ARTHROSE_OTHE	-	0,92	1,01	0,90
CARDIO0_HBP	-	0,92	0,96	0,92
DIG1_HEMORRAGIE	-	0,92	-	-
RESP0_APNEE_SOM	-	0,91	0,99	0,89
RHEUM0_ARTHROSE_HIP	-	0,90	-	0,92
RHEUM0_ARTHROSE_KNEE	-	0,89	-	0,88
DIG1_PERITONITE	-	0,89	0,98	-
DIG0_PANCREAS_AUD	-	0,84	-	0,91
CANCER0_ATCD_FAM	-	-	0,99	0,94
KIDNEY0_2OTHER	-	-	-	0,92
CARDIO1_RYTHME_1ACFA	-	-	0,97	0,87
CARDIO1_IHD_1MI	-	-	0,91	0,85
CARDIO0_OTHER	-	-	0,99	-
KIDNEY0_2UROLITHIASE	-	-	-	-
RESP0_2ASTHMA	-	-	-	-
RESP0_BRONCHITE_ATCD	-	-	1,03	-
STROKE0_2TTT	-	-	-	-
CANCER1_PROSTATE	-	-	1,02	-
TRAUMA1_2SEVERE	-	-	1,25	-
cp_immi	1,0002	0,99	1,02	-
SENSE0_CATARACTE	0,9972	0,93	0,97	0,90
CARDIO0_IHD_3noTTT	0,9957	0,97	0,96	0,96
CARDIO0_RYTHME_3noTT	0,99	0,92	0,98	0,93
ENDOC0_CARENCE	0,99	0,91	0,92	0,83
TRAUMA0_OSTEOPOROSE	0,99	0,98	-	1,08
CANCER0_SKIN	0,99	0,89	0,97	0,92
TRAUMA0_3FRACTURE	0,99	0,93	1,02	0,94
CARDIO0_RYTHME_2TTT	0,97	0,94	0,94	0,91
TRAUMA0_CHUTE	0,97	0,92	0,98	0,88
TRAUMA1_1CRANE	0,93	0,85	-	-

TABLE 3.1 – Comparaison des résultats obtenus entre les méthodes AIC et Elastic-net pour la dépendance

Les variables les plus influentes obtenues avec la méthode AIC sont les facteurs de risque de l'alcool, du tabac, de l'obésité et différents cancers. Ces mêmes variables ressortent également parmi les variables les plus influentes pour la régression pénalisée mais l'intensité de celles-ci est plus faible que lors de la première étude. Dans la partie 2, la modélisation impliquait qu'une Parisienne présentant le facteur de risque de l'alcool était 3.8 fois (resp 4.0 pour les Picards) plus susceptible d'être dépendante qu'un individu ne présentant pas ce facteur de risque. Dans le cadre de l'étude avec la régression pénalisée, ce facteur redescend à 1.5 (resp 2.1 pour les Picards). Par analogie avec la partie sur les données simulées, et en constatant que cette covariable est très corrélée aux autres, on déduit que ce coefficient est probablement sur-évalué dans le cas de la méthode de régression effectuée dans la partie 2.
Les résultats obtenus grâce à la régression pénalisée ont ainsi l'avantage de prendre en compte les problèmes liés à la corrélation entre variables et l'étape de sélection avec une relative simplicité de mise en œuvre et avec

une grande rapidité d'exécution. La sélection et le paramétrage du modèle avec la méthode AIC mettait en effet 24h à s'exécuter alors que la méthode pénalisée ne met que 15 minutes tout en prenant en compte les effets dûs à la corrélation.

3.3.2 Modélisation de la démence sur les bases des Picards et des Parisiennes à l'aide de la régression Elastic-net

La même méthodologie que celle appliquée à la modélisation de la dépendance physique est reprise ici. Les variables sélectionnées par le modèle dans cette section pourrons être confrontées aux causes de l'apparition de la démence établies dans la première partie de cette étude par l'intermédiaire des études issues de la communauté médicale.

Détermination du paramètre alpha de l'Elastic-Net

En procédant de la même manière mais pour prédire la démence, les résultats obtenus sont exposés ci-dessous :

Les graphiques suivants indiquent l'évolution du niveau de déviance des modèles de prédiction de la Démence obtenus en faisant varier le paramètre α de l'Elastic net sur les populations parisiennes et picardes.

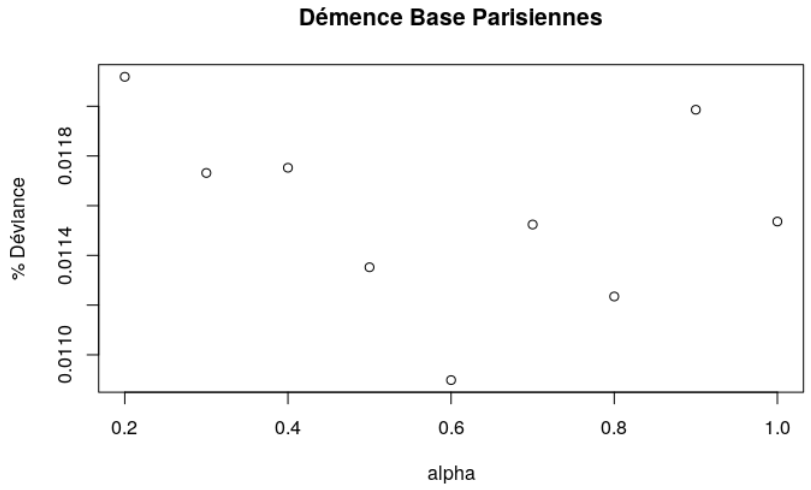


FIGURE 3.33 – Détermination du paramètre α de l'Elastic-net pour la démence des Parisiennes

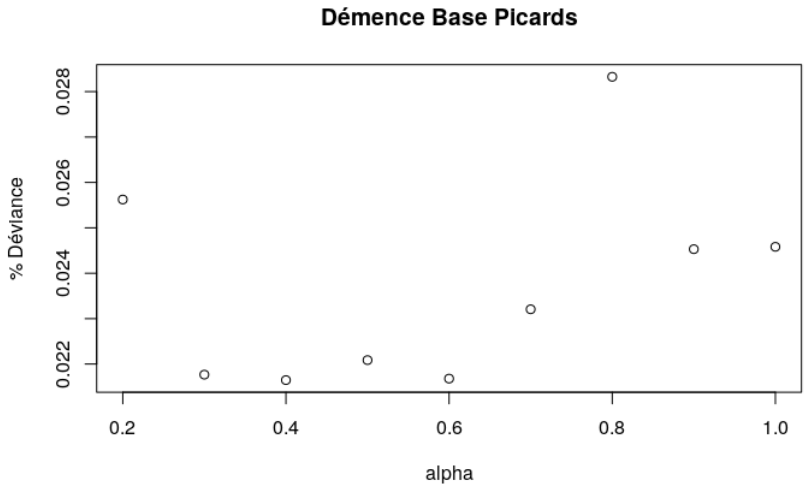


FIGURE 3.34 – Détermination du paramètre α de l'Elastic-net pour la démence des Picards

Nous choisissons un α de 0,6 pour paramétrer l'Elastic-net pour la démence puisque cette valeur du paramètre minimise la déviance.

Nous rappelons l'expression de l'ensemble de solution lorsque la valeur de la constante de régularisation varie :

$$\{\hat{\beta}(\lambda) \in [0, +\infty[\text{ où } \hat{\beta}(\lambda) = argmin_\beta \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - \lambda(\frac{1}{2}\sum_{j=1}^p \beta_j^2 + \frac{1}{2}\sum_{j=1}^p |\beta_j|)\}$$

Le chemin de régularisation ci-dessous représente l'ensemble des valeurs prises par chacun des coefficients de régression en fonction de la valeur du paramètre de régularisation λ.

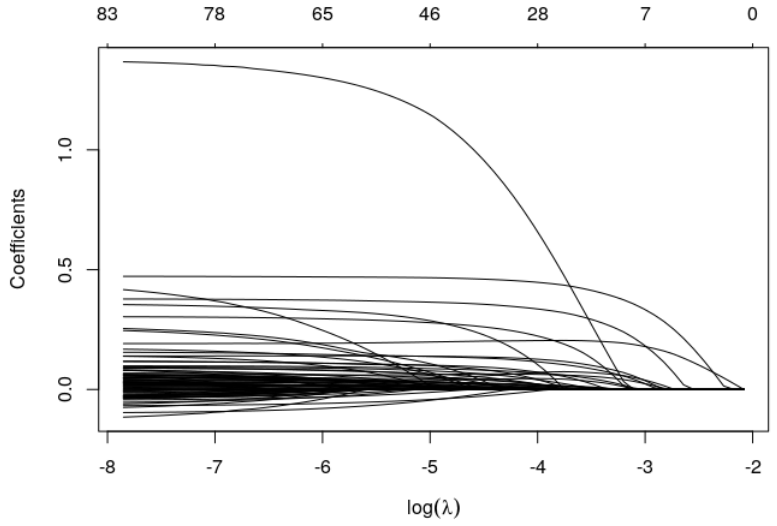


FIGURE 3.35 – démence des Parisiennes

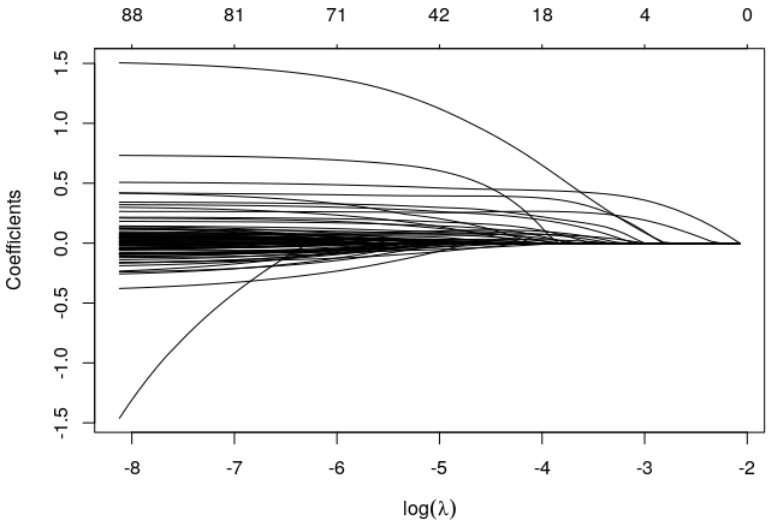


FIGURE 3.36 – démence des Picards

Comme dans le cas de la dépendance physique, les chemins de régularisation ont des formes globalement similaires entre les deux populations.

En utilisant une 10-folds validation croisée avec la déviance comme mesure de prédiction, nous obtenons les graphes suivants :

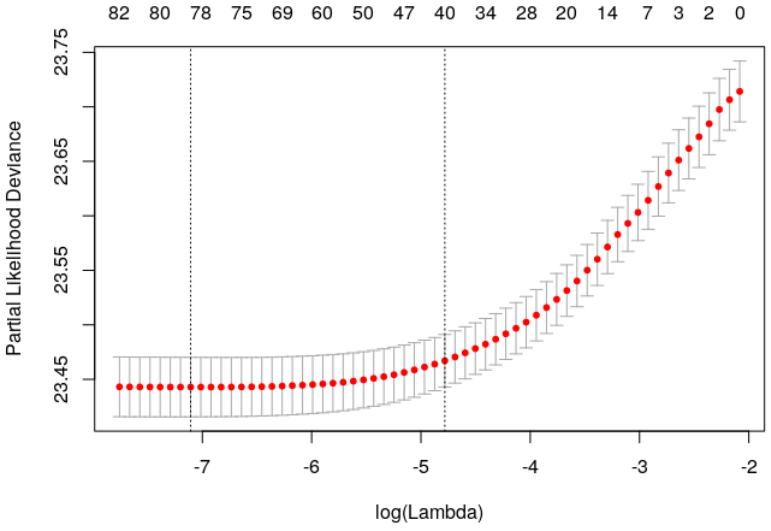


FIGURE 3.37 – Détermination du paramètre α de l'Elastic-net pour la démence des Parisiennes

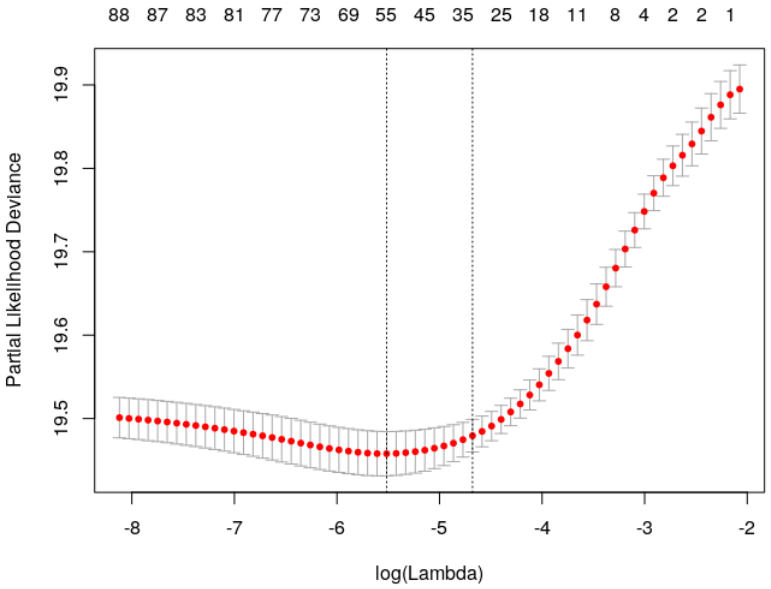


FIGURE 3.38 – Détermination du paramètre α de l'Elastic-net pour la démence des Picards

L'évolution de la déviance en fonction de la pénalisation suscite les mêmes commentaires que pour la dépendance physique. En effet, on constate que la pente de la déviance est forte jusqu'à $\log(\lambda)$ proche de -5. Dans cette partie, chaque variable ajoutée au modèle tend à faire baisser sa déviance de façon significative. Pour des λ plus faibles, l'ajout de variables a peu d'influence sur la déviance. Les variables entrant dans le modèle ont donc a priori peu d'influence sur la variable de sortie. En suivant le critère du lambda qui s'écarte d'un écart type du λ minimum (la ligne pointillée de droite sur le graphique), les modèles sélectionnés sur la base des Parisiennes et des Picards sont les suivants (avec pour rappel les résultats obtenus dans le cadre de la sélection avec le critère AIC stepwise :

Les covariables identifiées comme causes d'apparition de la démence dans la première partie de ce travail sur la base d'études reconnues du milieu médical apparaissent en gras. Les intensités attribuées à chaque cofacteurs sont par ailleurs triées par ordre décroissant sur le périmètre de la démence des Parisiennes. On remarque que toutes les causes identifiées a priori ressortent en tête du classement.

Ceci confirme que la sélection automatique avec une pénalisation Elastic-net est performante, et ce malgré les fortes corrélations qu'il existe dans la base d'étude.

Nous comparons dans le tableau suivants, les seules covariables identifiées dans la première partie comme causes de démence selon le type de méthode employé et sur la base des Picards et des Parisiennes :

	Parisienne Demence		Picard Demence	
Covariable	Enet	AIC	Enet	AIC
fdr_aud_all	**2,46**	**4,14**	**2,21**	**4,13**
NEURO1_PARKINSON	**1,57**	**1,29**	**-**	**1,11**
NEURO1_EPILEPSIE	**1,43**	**1,21**	**1,18**	**1,21**
NEURO1_DEM_FDR	**1,33**	**1,36**	**2,28**	**1,42**
STROKE1_1HEMO	**1,33**	**1,22**	**1,67**	**-**
ENDOC0_CARENCE	1,20	0,97	0,92	0,89
STROKE1_1ISCHEMIC	**1,15**	**-**	**1,37**	**-**
DIG1_OCCLUSION	1,15	1,05	1,05	-
DIG1_STOMIE	1,14	1,16	1,32	1,32
ENDOC1_GLD_OTHER	1,12	1,06	1,04	1,16
NEURO1_OTHER	1,11	0,72	1,45	1,14
INFECT1_SEPSIS	1,11	1,20	1,44	1,16
fdr_smoker	1,10	2,94	1,65	2,50
KIDNEY0_2PYELONEPHRI	1,09	-	-	0,85
fdr_obesity_all	1,09	1,77	1,23	1,26
NEURO0_SNP_METABO	1,08	1,05	1,07	1,10
CV1_MTE	1,08	-	1,09	1,07
ENDOC0_DIABETE	1,07	1,08	-	1,03
KIDNEY0_CYSTITE_ATCD	1,07	-	1,12	0,90
TRAUMA1_1CRANE	1,06	0,89	-	-
RESP1_2INSUF_AIGUE	1,05	1,09	1,11	-
TRAUMA0_SUICIDE	1,05	1,10	-	-
KIDNEY0_2GN	1,04	-	1,03	-
KIDNEY1_2INSUF_AIGUE	1,04	1,02	-	1,12
KIDNEY1_1INSUF_CHRO	1,04	0,93	1,00	-
DIG1_LIVER_1CirrD	1,04	1,24	1,14	1,21
BLOOD0_2ANEMIA_IRON	1,04	-	0,98	0,93
RHEUM0_AUD	1,04	0,96	-	0,93
DIG0_LIVER_2CirrC	1,04	-	1,01	-
TRAUMA0_3FRACTURE	1,03	0,95	1,02	0,93
DIG0_LIVER_ETIO_ANY	1,03	1,09	1,00	1,05
CANCER1_HEMATO	1,03	1,24	1,17	1,24
CARDIO0_OTHER	1,03	-	0,99	-
ENDOC1_METABO	1,03	1,06	1,04	-
TRAUMA0_CHUTE	1,03	0,96	0,98	0,93
ENDOC0_THYROIDE	1,02	0,99	0,97	-
cp_dipl0	1,02	1,13	1,02	1,05
CANCER1_BREAST	1,02	1,44	1,31	-
CARDIO1_RYTHME_1ACFA	1,02	0,97	0,97	0,87
ENDOC0_DYSLIPIDEMIA	1,02	0,99	0,98	0,97
RESP0_LRI_ATCD	1,02	0,96	1,01	-
CARDIO0_HBP	1,02	0,95	0,96	0,92
TRAUMA0_OSTEOPOROSE	1,02	-	-	1,11
cp_dep	1,01	1,01	0,88	-
cp_immi	1,01	-	1,02	-

	Parisienne Demence		Picard Demence	
Covariable	Enet	AIC	Enet	AIC
STROKE0_3noTTT	1,01	0,97	0,92	-
TRAUMA1_2SEVERE	1,00	-	1,25	-
BLOOD1_1TRANSFUSION	1,00	1,08	1,17	1,10
CANCER1_SMOKER	-	2,13	1,55	1,75
CANCER1_PC_POOR	-	2,03	1,68	1,60
CANCER1_AUD_SMOKER	-	1,66	1,22	1,41
CANCER1_PC_GOOD	-	1,60	1,20	1,22
CANCER1_COLORECTAL	-	1,24	1,08	1,23
BLOOD0_2OTHER	-	1,13	1,04	1,05
RHEUM0_SYSTEME	-	1,08	1,05	-
BLOOD0_2ANEMIA_LYSE	-	1,08	1,18	1,21
RESP0_2OTHER	-	1,06	1,01	-
RESP1_1INSUF_CHRO	-	1,05	1,10	0,92
RESP0_2INTERSTI	-	-	1,01	1,27
CARDIO0_VALVE_ANY	-	1,05	-	1,03
CARDIO0_IHD_2TTT	-	-	0,99	0,97
CANCER0_BENIN	-	0,97	0,98	0,90
DIG0_VB_METABO	-	0,97	0,94	0,89
CANCER0_INSITU	-	0,97	1,03	-
STROKE0_AIT_noSEQ	-	0,98	1,03	-
DIG0_OTHER_noFDR	-	0,95	0,98	0,93
RHEUM0_ARTHROSE_OTHE	-	0,95	1,01	0,91
CARDIO0_RYTHME_3noTT	-	0,94	0,98	0,93
DIG1_HEMORRAGIE	-	0,93	-	-
RESP0_APNEE_SOM	-	0,93	0,99	0,92
RHEUM0_ARTHROSE_KNEE	-	0,90	-	0,87
DIG1_PERITONITE	-	0,89	0,98	-
DIG0_PANCREAS_AUD	-	0,85	-	0,89
CARDIO1_INSUF_CHRO	-	0,96	1,00	0,94
CANCER0_ATCD_FAM	-	-	0,99	0,95
KIDNEY0_2OTHER	-	-	-	0,93
KIDNEY0_2UROLITHIASE	-	-	-	-
RESP0_2ASTHMA	-	-	-	-
RESP0_2COPD	-	0,98	1,03	-
RESP0_BRONCHITE_ATCD	-	-	1,03	-
STROKE0_2TTT	-	-	-	-
CANCER1_PROSTATE	-	-	1,02	-
CV1_PVD	-	-	1,01	-
CARDIO1_IHD_1MI	1,00	0,96	0,91	0,88
CARDIO0_IHD_3noTTT	0,99	0,96	0,96	0,97
RHEUM0_ARTHROSE_HIP	0,99	0,91	-	0,92
SENSE0_CATARACTE	0,98	0,93	0,97	0,90
CARDIO0_RYTHME_2TTT	0,95	0,93	0,94	0,90
CANCER0_SKIN	0,95	0,88	0,97	0,90

TABLE 3.2 – Comparaison des résultats obtenus entre les méthodes AIC et Elastic-net pour la démence

	Parisienne Demence		Picard Demence	
Covariable	Enet	AIC	Enet	AIC
fdr_aud_all	**2,46**	**4,14**	**2,21**	**4,13**
NEURO1_PARKINSON	**1,57**	**1,29**	**-**	**1,11**
NEURO1_EPILEPSIE	**1,43**	**1,21**	**1,18**	**1,21**
NEURO1_DEM_FDR	**1,33**	**1,36**	**2,28**	**1,42**
STROKE1_1HEMO	**1,33**	**1,22**	**1,67**	**-**
STROKE1_1ISCHEMIC	**1,15**	**-**	**1,37**	**-**

TABLE 3.3 – Comparaison sur le périmètre des covariables identifiées comme à risque par la profession médicale

On constate que la sélection est meilleure avec les méthodes d'apprentissage qu'avec la méthode classique. On remarque de plus que le poids du facteur de risque "Alcool" est beaucoup plus important avec la méthode AIC qu'avec l'Elastic net. En effet, par analogie avec l'analyse effectuée sur les jeux de données simulées, cette covariable capte l'information de l'ensemble des variables qui lui sont corrélées, chose que l'Elastic net gère grâce à la partie de la pénalisation de type Ridge qui pénalise la norme l_2 des coefficients de régression. On remarque par ailleurs que sur la base des Picards la covariable "Parkinson" n'a pas été sélectionnée par l'Elastic-net alors que celle-ci l'est sur la population Parisienne. Les études médicales indiquaient (cf chapitre 1) que le cas de la démence parkinsonienne se déclarait en moyenne 10 à 15 ans après l'apparition de la maladie de Parkinson qui se déclare à des âges moyens relativement avancés. De ce fait, il est possible que la population masculine picarde, dont l'espérance de vie est moins importante que la population féminine parisienne capte moins bien l'information qui la lie à la démence du fait du phénomène de censure droite probablement plus important.

Les résultats obtenus démontrent le rôle prépondérant de l'alcool sur l'apparition de la démence et de la dépendance physique. Les graphiques suivants comparent les estimations issues des modèles de Cox sur les populations parisiennes et picardes alcooliques et non alcooliques ainsi que les estimations issues de l'étude Qalydays réalisés sur la France entière.

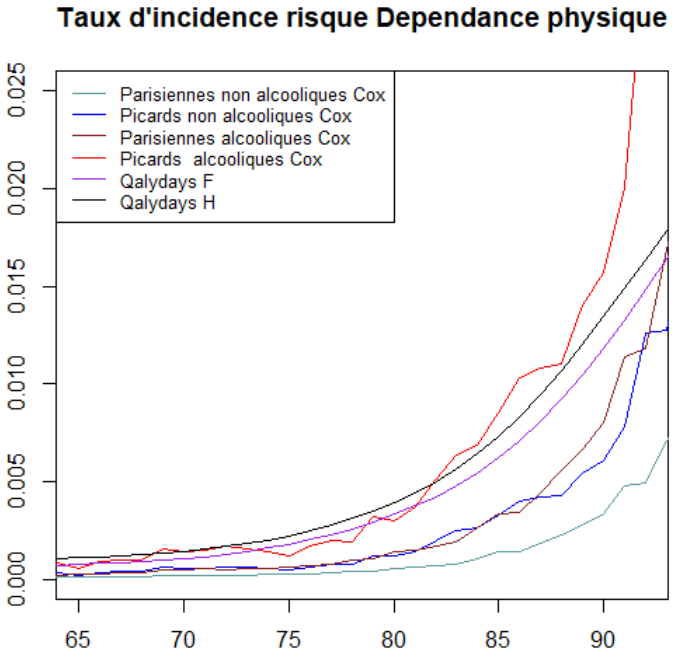


TABLE 3.4 – Comparaison des modèles prédictifs de Cox pour la Dépendance physique

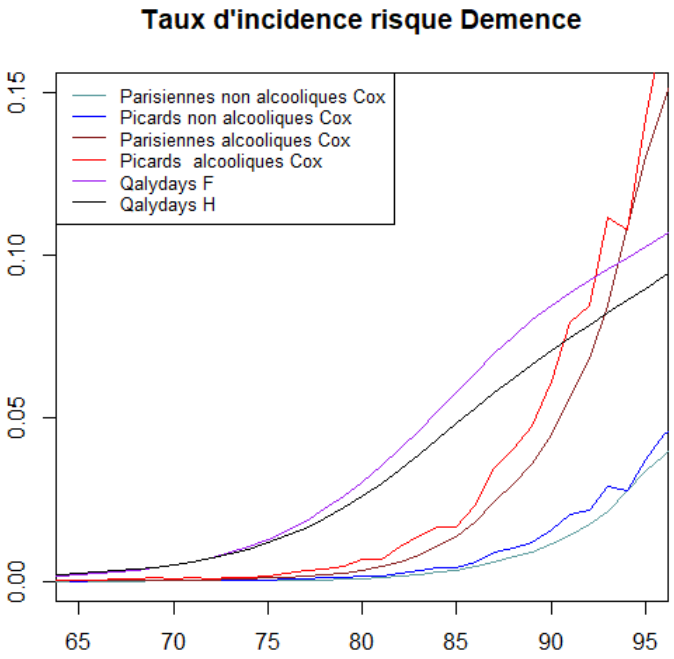


TABLE 3.5 – Comparaison des modèles prédictifs de Cox pour la Démence

Les populations alcooliques ressortent plus risquées que la population française de référence Qalydays et les populations non alcooliques ressortent moins à risque ce qui semble cohérent. De plus, comme attendu, les Picards ont des incidences plus importantes que les Parisiennes du fait de leurs différences de sexe, de niveau de vie et d'accès aux soins comme évoqué dans la première partie.
L'avantage du paramétrage des lois par l'intermédiaire de l'Elastic-net est la grande souplesse et la grande rapidité d'exécution et de sélection tout en gérant les problèmes de colinéarité particulièrement complexes lorsque le nombre de variables est aussi important. Cette méthode permet d'industrialiser la modélisation à un grand nombre de sous-populations avec une grande flexibilité et avec des résultats satisfaisants.

Parisienne démence AIC:

Likelihood ratio test=4965 on 67 df, p=< 2.2e-16
n= 941696, number of events= 18260

Parisienne démence Elastic-net:

Likelihood ratio test=4967 on 39 df, p=< 2.2e-16
n= 941696, number of events= 18260

Parisienne dépendance AIC:

Likelihood ratio test=4717 on 73 df, p=< 2.2e-16
n= 941696, number of events= 6438

Parisienne dépendance Elastic-net:

Likelihood ratio test=4345 on 26 df, p=< 2.2e-16
n= 941696, number of events= 6438

TABLE 3.6 – Comparaison des modèles prédictifs de Cox pour la Dépendance physique

Picards démence AIC:

Likelihood ratio test=1587 on 71 df, p=< 2.2e-16
n= 137355, number of events= 1828

Picards démence Elastic-net:

Likelihood ratio test=1565 on 35 df, p=< 2.2e-16
n= 137355, number of events= 1828

Picards dépendance AIC:

Likelihood ratio test=1902 on 72 df, p=< 2.2e-16
n= 137355, number of events= 1357

Picards dépendance Elastic-net:

Likelihood ratio test=1951 on 36 df, p=< 2.2e-16
n= 137355, number of events= 1357

TABLE 3.7 – Comparaison des modèles prédictifs de Cox pour la Démence

En effectuant des tests de vraisemblance sur les modèles paramétrés par le biais du critère AIC et de l'Elastic-net, on remarque que la qualité prédictive des modèles est globalement égale d'une méthode à l'autre alors que le nombre de degrés de libertés est toujours plus faible pour la sélection Elastic-net. Les modèles de régression pénalisée apparaissent donc comme aussi performant, plus parcimonieux et extrêmement plus rapides à être exécutés ($\sim$ 15 min vs 24h).

Détermination et analyse d'interaction avec un LASSO

L'objet de ce paragraphe est de présenter rapidement la possibilité de détection d'interactions au moyen d'une régression LASSO. Une interaction statistique entre deux variables explicatives apparaît lorsque l'effet de l'une d'entre elles sur la variable réponse dépend de la valeur de l'autre.

Dans le cas simple d'une régression linéaire à deux variables $Y \sim X_1 + X_2$
Il y a interaction entre X_1 et X_2 si l'effet marginal de X_1 sur Y dépend de X_2 :

$$\frac{\partial(\hat{Y})}{\partial X_1} = \beta_1(X_2)$$

En croisant certaines pathologies et en appliquant une méthode de sélection LASSO, il est ainsi possible de détecter des interactions. Nous n'exhiberons ici qu'un exemple afin d'illustrer la méthode :

Covariable	Parisienne Demence
fdr_aud_all	4.22
...	...
STROKE1_1ISCHEMIC	1.7
...	...
TRAUMA0_OSTEOPOROSE	1.14
...	...
NEURO1_PARKINSON	1.58
...	...
fdr_aud_allxNEURO1_PARKINSON	1
fdr_aud_allxTRAUMA0_OSTEOPOROSE	1
fdr_aud_allxSTROKE1_1ISCHEMIC	0.83

TABLE 3.8 – Exemple de détection d'interactions

En croisant la covariable alcool (fdr_aud_all) avec certaines pathologies et en rajoutant ces interactions à la liste initiale il est possible de détecter celles qui sont significatives, c'est à dire celles dont le LASSO n'a pas fixé le coefficient à 0.
En opérant ainsi sur quelques pathologies pour la sortie démence sur la base des Parisiennes, on remarque que le terme d'interaction Alcool*AVC ischémique ressort avec un coefficient de 0.83 alors que les deux pathologies Alcool et AVC ischémique sont des facteurs aggravants en termes de démence (coefficients égaux à 4,2 et 1,7 respectivement). En effectuant une rapide recherche, on apprend que "boire de façon modérée diminuerait le risque d'AVC ischémique, causé par des caillots sanguins qui bloquent les artères cérébrales malades ou endommagées." [3] Le LASSO apparait ainsi comme une méthode simple pour aider à détecter des interactions non évidentes a priori. L'avis d'expert apparait tout de même fondamental pour valider ou invalider ce type d'analyse.

3. https ://www.topsante.com/medecine/troubles-cardiovasculaires/avc/l'alcool-favorise-une-des-deux-formes-d'avc-61445

Conclusion

Avec l'explosion des dispositifs de santé connectés et du stockage massif de données, nul doute que les secteurs de l'assurance santé et de la prévoyance vont connaitre une transformation dans les années à venir. Cette transformation s'accompagne d'une expansion de l'utilisation de méthodes d'apprentissage particulièrement appropriées dans un cadre de données abondantes. Ces travaux ont démontré, à travers l'application d'une forme simple de méthode d'apprentissage, la grande flexibilité et la relative facilité de mise en œuvre d'une de ces méthodes. En effet, le processus de sélection de variables ainsi que la gestion des corrélations étant directement inclus dans l'étape d'optimisation des paramètres, l'utilisation d'une pénalisation de type Elastic-net paramétrée via une technique d'apprentissage a permis l'obtention de résultats satisfaisant avec un temps d'exécution extrêmement plus faible qu'avec une procédure traditionnelle fondée sur des calculs itératifs. En effet, la quasi-totalité des variables pré-identifiées comme responsables de la démence par les études médicales citées dans la première partie ont été sélectionnées par le modèle pénalisé par l'Elastic-net et paramétré grâce aux méthodes d'apprentissage. D'autres méthodes comme les arbres de régression CART donnent de très bons résultats en termes de sélection de variables pour ce type de problèmes mais ces méthodes sont plus complexes à mettre en œuvre, très chronophages et s'avèrent beaucoup moins souples. En effet, la méthode CART ne permet pas de générer un grand nombre de modèles de façon "industrielle" comme le permet la pénalisation Elastic-net.

Cependant, la nécessité de disposer de volumes de données importants représente une contrainte non négligeable pour l'utilisation des méthodes d'apprentissage. En effet, malgré les techniques de ré-échantillonage comme la validation croisée, l'étude montre que le volume d'informations disponibles conditionne fortement la qualité de l'apprentissage. Par ailleurs, l'idée répendue selon laquelle le machine learning permet de s'affranchir de l'intervention de l'homme est erronée. En effet, le choix en amont du modèle, la phase préliminaire essentielle de retraitement et de regroupement des variables ainsi que la phase de validation et d'analyse de la cohérence des résultats sont impossibles sans l'expertise humaine. En effet, il est clair à travers ces travaux que l'obtention des résultats via les méthodes d'apprentissage n'aurait pas été possible sans l'expertise du docteur Schwarzinger et toute la phase indispensable de traitements de la donnée brute pour obtenir une base de données exploitable. En effet cette expertise médicale a par exemple été nécessaire lors de la phase de regroupement de pathologies parmi les quelques 14 000 existantes dans le code CIM-10 de l'OMS ou encore lors des choix d'exclusions de l'étude, pour la gestion des troncatures gauches, des individus présentant certaines pathologies jugés comme sévères. Le choix d'un modèle multiplicatif à risques concurrents, la validation des résultats finaux opérée à partir d'étude médicales reconnues sont autant d'étapes qui démontrent que l'intervention humaine est omniprésente dans ce type de travail statistique malgré l'utilisation de méthodes d'apprentissage. L'apport de ces méthodes se situe au niveau de la sélection de variables, de la gestion des corrélations et de la possibilité de répliquer facilement l'étude à d'autres populations. Les méthodes d'apprentissage apparaissent ainsi comme un complément intéressant mais non comme un substitut aux méthodes traditionnelles.

Le choix pour cette étude d'un modèle relativement simple a entrainé l'application d'hypothèses simplificatrices comme la condition de proportionnalité des effets ou encore la censure simple. Pour aller plus loin il est possible d'introduire un modèle à censures aléatoires en incluant le décès comme une cause de sortie aléatoire ou encore de segmenter le temps en plusieurs intervalles pour prendre en compte le caractère non proportionnel des effets en fonction du temps. L'introduction d'interactions entre variables pourrait également apporter des résultats intéressants. Quelques exemples on été donnés pour lesquels les effets pris seuls avaient une influence positive sur la probabilité d'entrer en dépendance alors que les mêmes effets combinés influaient en sens opposé. Malgré ces simplifications, il a été démontré que la régression pénalisée offrait des résultats satisfaisants pour sélectionner des modèles cohérents sans avis d'expert dans un cadre complexe avec un grand nombre de variables très corrélées.

Par ailleurs, la mise en lumière des effets des différentes pathologies ou facteur de risques sur la probabilité d'entrer en dépendance peut aider les assureurs à affiner les clauses d'exclusion de leurs contrats et améliorer le processus de souscription par l'intermédiaire d'un questionnaire médicale plus fin. Ce type de produit étant particulièrement difficile à tarifer, une sélection rigoureuse est essentielle pour permettre aux assureurs d'améliorer la qualité de leur portefeuille dont les ratios S/P sont pour l'heure souvent élevés. Les résultats obtenus lors de cette étude mettent notamment en lumière la surexposition des personnes identifiées comme alcooliques en matière de dépendance physique ou cognitive.

Enfin, cette étude montre une fois de plus les conséquences ravageuses de la consommation d'alcool sur la santé publique et nul doute que l'une des pistes de travail sur l'actuel dossier du financement de la dépendance par les pouvoirs publics passera par une meilleure sensibilisation de la population aux risques qui découlent d'une consommation excessive d'alcool.

ANNEXES

Annexe 1 : Mise en évidence des facteurs de risque d'entrée en dépendance

Notion de prévalence et d'incidence

En épidémiologie, la prévalence est une mesure de l'état de santé d'une population, dénombrant le nombre de cas de maladies à un instant donné ou sur une période donnée. Pour une affection donnée, on calcule ainsi le taux de prévalence en rapportant à la population considérée, ce nombre de cas de maladies présents dans cette population (source : https ://fr.wikipedia.org/wiki/Prévalence). L'incidence ne tient compte que des nouveaux cas sur une période donnée, alors que la prévalence s'appuie sur le nombre total de cas présents sur cette même période, c'est-à-dire ceux déjà présents plus ceux incidents.
La prévalence est donc toujours supérieure à l'incidence.
A noter que si la prévalence est faible (P $<$ 5 %), alors la relation entre incidence et prévalence sur la même période est :

$$P = I \times D$$

avec P la prévalence, I l'incidence et D la durée moyenne de la maladie.

En cas d'épidémie il y a déstabilisation de l'incidence ; la prévalence peut alors évoluer très différemment, par exemple selon l'impact de la maladie sur le taux de mortalité.
Si la maladie a une influence très faible sur le taux de mortalité de la population, à la fin de l'épidémie la prévalence déclinera lentement, la persistance de la maladie (temps nécessaire à la disparition de son agent étiologique) sera importante et la guérison de la population sera longue (schéma A).
Au contraire, si la maladie augmente de façon importante le taux de mortalité, la prévalence déclinera plus rapidement dès la fin de l'épidémie, sa persistance sera moins longue et la guérison de la population sera plus rapide (schéma B).

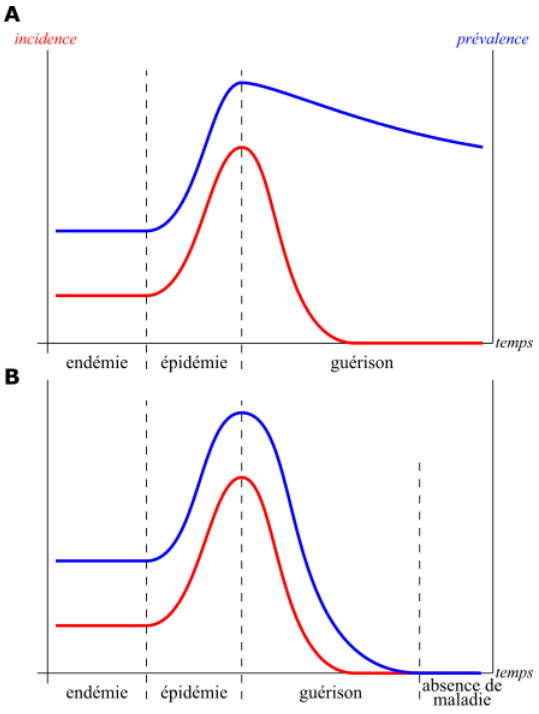


FIGURE 39 – Illustration des concepts de Prévalence et incidence (Wikipedia)

Le cas de la dépendance est particulier puisque il est irréversible. La notion de guérison n'a

donc pas lieu d'être. Seul le décès est la cause de sortie possible de l'état.

Démence de type Alzheimer.

La démence touche 5 à 7% des personnes âgées de 60 ans et plus (Lancet). Plusieurs types de démence existent. La plus connue et la plus rependue est la maladie d'Alzheimer, elle représente environ 50 à 60% des causes de démence (Nyenhuis et coll., 1998). Cette pathologie affecte la mémoire et induit donc inéluctablement la perte d'autonomie.

Le diabète, cause probable de l'apparition d'Alzheimer.

Xu.W and Al [2007] ((Diabetes. 2007 Jan ;56(1) :211-6.) établissent un lien entre le diabète et le risque de contracter la maladie d'Alzheimer. Les auteurs ont en effet étudié une cohorte de 1 173 personnes non diabétiques et non atteintes de démence et âgées d'au moins 75 ans. L'évolution des patients a été analysée à 3 reprises et avec une fréquence de 3 ans. Les données ont été analysées au moyen d'un modèle de Cox à hasard proportionnel. Durant les 9 années de suivis, 397 sujets ont développés de la démence et parmi eux, 307 des cas de maladie d'Alzheimer. 47 sujets ont été diagnostiqués diabétiques. Les cas de diabètes ont été associés à des ratios de Cox de 1.67 (IC à 95% 1,04 – 2.67) pour la démence et 1.77 (IC à 95% 1.06-2.97) pour la maladie d'Alzheimer. Ceci signifie que dans cette population, les personnes ayant présentées un diabète ont eu 77% de chance en plus de contracter la maladie d'Alzheimer que les autres.
Cette étude, bien que fondée sur un échantillon assez faible établit un lien entre la présence de diabète et l'occurrence de la maladie d'Alzheimer. L'étendue de l'intervalle de confiance à 95% est cependant très large du fait de la faiblesse de l'effectif et donc de la grande volatilité des estimateurs obtenus. Cette étude apporte néanmoins des éléments quantitatifs qui pourront être confrontés aux résultats obtenus avec les méthodes utilisées dans ce mémoire.

Démence de type vasculaire [4]

Il existe également une forme de démence appelée démence vasculaire. On admet généralement que celle-ci représente la seconde cause de démence dans les populations européennes et américaines, après la démence dégénérative de type Alzheimer. Elle représente 10 à 20% des cas de démence (Nyenhuis et coll., 1998). Cependant chez les sujets très âgés, ces proportions se modifient considérablement. En effet dans leur étude épidémiologique, Skoog et coll., trouvent une prévalence discrètement plus élevée de démence vasculaire (46,9%) par rapport à la démence de type Alzheimer (43,5%), cela dans une population de patients de 85 ans (Skoog et coll., 1993). A noter également que même si les accidents vasculaires cérébraux sont plus fréquents chez l'homme, la prévalence de la démence vasculaire est semblable pour les deux sexes. L'incidence de la démence vasculaire, est estimée à 6-10 cas par an pour 1000 personnes de plus de 70 ans. L'incidence augmente avec l'âge et il n'y a pas non plus, tout comme pour la prévalence, de différence entre les deux sexes (Hebert et coll., 1995).

Le facteur de risque hypertension artérielle et cardiopathie causes probables de l'apparition de la démence vasculaire.

Le facteur de risque le plus important pour la démence vasculaire est l'hypertension artérielle (Skoog, 1994). Elle serait en effet deux fois plus fréquente chez les patients atteints de démence vasculaire que chez les sujets du même âge non-déments (Gold et coll., 1998). Au Japon, où la démence vasculaire a une incidence importante, une augmentation de la pression artérielle systolique d'un écart-type, s'accompagne d'une augmentation de 60% de risque de survenue d'une démence vasculaire, chez les sujets âgés de plus de 60 ans (Yoshitake et coll., 1995).
Par ailleurs, une hypertension artérielle a été retrouvée dans plus de 90% des cas de maladie de Binswanger décrits dans la littérature (Babikian et coll., 1987).
D'autres facteurs liés au risque vasculaire tels que le diabète sucré, le tabac, la fibrillation auriculaire et la présence d'une cardiopathie sont retrouvés avec une fréquence plus élevée chez les

4. issu de la Thèse Bachetta pour l'université de Médecine de Genève

individus présentant une démence vasculaire que chez les sujets du même âge non-déments. Il est possible que certains de ces facteurs comme l'hypertension et le diabète soient directement associés à la démence vasculaire, alors que d'autres le soient indirectement : c'est le cas pour le tabac et l'élévation du cholestérol, qui constituent des facteurs de risque pour les accidents vasculaires cérébraux (AVC). Les AVC quant à eux, augmentent de façon considérable le risque de survenue d'une démence. Ainsi, le risque serait multiplié par neuf dans l'année qui suit l'AVC (Kokmen et coll., 1996).

Le facteur de risque "Alcool", cause majeure de l'apparition de la démence.

L'article Lancet Public Health 3 : e124–32 [Schwarzinger and al, 2018] établit un lien fort entre l'apparition de la démence (au sens médical du CIM-10) et la consommation d'alcool.
Sur la base des données du PMSI, de l'ensemble des personnes admises à l'hôpital en France entre 2008 et 2013 (31 624 156 observations), les auteurs ont mis en évidence à travers l'utilisation d'un modèle de Cox que le facteur de risque "Alcool" était prépondérant. Le tableau suivant issu de cette étude présente les facteurs de risque majeurs et l'intensité de leurs coefficients de Cox pour les populations féminines et masculines sur le périmètre France entière

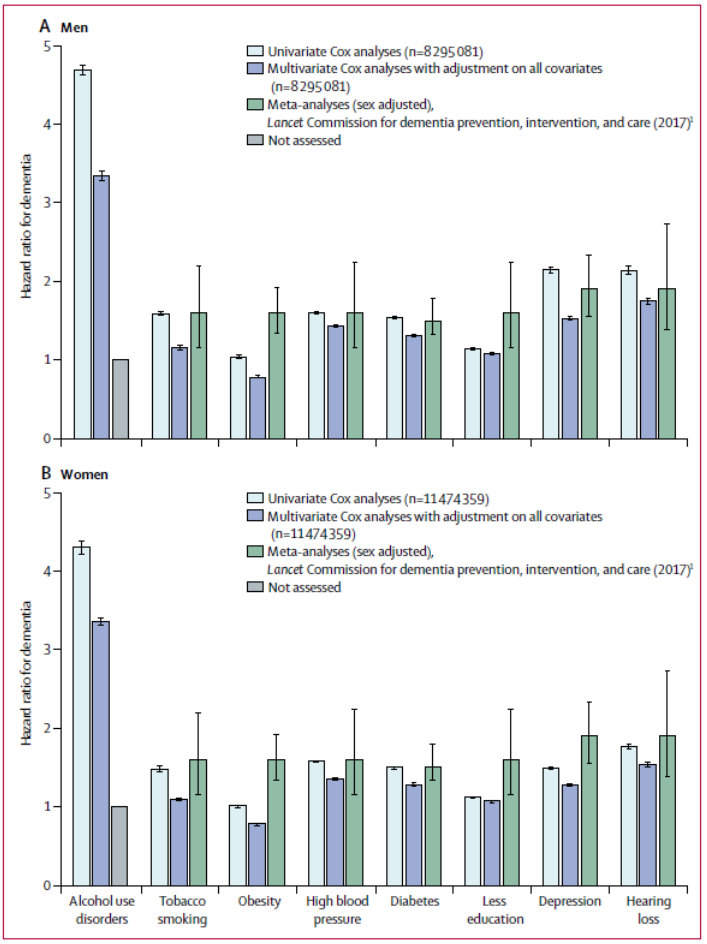


FIGURE 40 – Étude facteur démence (article Lancet Public Health 3 : e124–32 [Schwarzinger and al, 2018])

Les coefficients de Cox des facteurs de risque "Alcool" sont respectivement de 3.34 (IC 95% 3.28 – 3.41) pour les femmes et 3.36 (IC 95% 3.31 – 3.41) pour les hommes. On retrouve l'hypertension, le diabète et le tabac, déjà évoqués dans les études cités dans les précédents paragraphes.

Démence Parkinsonienne et démence à corps de Lewy. [5]

La démence à corps de Lewy est une perte progressive de la fonction cognitive caractérisée par le développement des corps de Lewy dans les cellules nerveuses. La démence de la maladie de Parkinson est une perte progressive de la fonction cognitive caractérisée par le développement des corps de Lewy chez les personnes atteintes de la maladie de Parkinson.

- Les personnes atteintes de démence à corps de Lewy fluctuent entre vigilance et somnolence et peuvent avoir des difficultés à dessiner, à se mouvoir et avoir des hallucinations qui sont identiques à celles dues à la maladie de Parkinson.
- La démence de la maladie de Parkinson se développe habituellement 10 à 15 ans après l'apparition d'autres symptômes de la maladie de Parkinson.
- Le diagnostic repose sur les symptômes.
- Des stratégies sont utilisées pour prolonger le fonctionnement aussi longtemps que possible, et les médicaments utilisés pour traiter la maladie d'Alzheimer peuvent aider.

La démence à corps de Lewy est le troisième type de démence le plus fréquent. La démence à corps de Lewy se développe chez des personnes de plus de 60 ans. L'âge moyen d'apparition de la maladie de parkinson est 70 ans. [6]
Environ 40 % des personnes atteintes de la maladie de Parkinson développent une démence de la maladie de Parkinson.

L'épilepsie influence le pronostic de la démence. [7]

« L'épilepsie et la démence sont deux pathologies fréquentes chez le sujet âgé. L'épilepsie influence le pronostic de la démence : aggravation des troubles cognitifs, perte d'autonomie conduisant à la mise en institution, augmentation de la mortalité et de la morbidité. De plus, les sujets âgés atteints de démence sont plus vulnérables aux effets secondaires des médicaments antiépileptiques. » [8]

5. issue des travaux de Juebin Huang, MD, PhD, Assistant Professor, Department of Neurology, Memory Impairment and Neurodegenerative Dementia (MIND) Center pour l'University of Mississippi Medical Center
6. https ://www.inserm.fr/information-en-sante/dossiers-information/parkinson-maladie
7. issue des travaux du Dr Caroline Hommet et Dr Marc Verny à travers l'ouvrage "Gériatrie et Psychologie Neuropsychiatrie du Vieillissement"
8. https ://www.inserm.fr/information-en-sante/dossiers-information/epilepsie)

Annexe 2 : Test du χ^2

	rho	chisq	p
fdr_aud_all	-0.023965	5.05e+00	2.46e-02
fdr_smoker	-0.008236	6.20e-01	4.31e-01
fdr_obesity_all	-0.011557	1.12e+00	2.91e-01
cp_dep	0.005540	2.58e-01	6.12e-01
cp_immi	-0.034850	9.03e+00	2.65e-03
cp_dipl0	-0.035429	9.54e+00	2.01e-03
BLOOD0_2ANEMIA_IRON	-0.007620	4.02e-01	5.26e-01
BLOOD0_2ANEMIA_LYSE	-0.018813	2.30e+00	1.29e-01
BLOOD0_2OTHER	-0.007105	2.69e-01	6.04e-01
CANCER0_ATCD_FAM	-0.006273	3.17e-01	5.73e-01
CANCER0_BENIN	0.006316	3.12e-01	5.76e-01
CANCER0_INSITU	-0.002604	5.26e-02	8.19e-01
CANCER0_SKIN	0.022265	4.01e+00	4.53e-02
CARDIO0_HBP	0.032542	7.49e+00	6.21e-03
CARDIO0_IHD_2TTT	-0.004293	1.50e-01	6.98e-01
CARDIO0_IHD_3noTTT	0.026160	5.37e+00	2.04e-02
CARDIO0_OTHER	-0.009645	6.96e-01	4.04e-01
CARDIO0_RYTHME_2TTT	0.027193	6.08e+00	1.37e-02
CARDIO0_RYTHME_3noTT	0.008094	5.79e-01	4.47e-01
CARDIO0_VALVE_ANY	0.003361	9.72e-02	7.55e-01
DIG0_LIVER_2CirrC	-0.000487	1.55e-03	9.69e-01
DIG0_LIVER_ETIO_ANY	0.004046	1.47e-01	7.01e-01
DIG0_OTHER_noFDR	0.020326	2.98e+00	8.43e-02
DIG0_PANCREAS_AUD	0.011892	2.18e+00	1.40e-01
DIG0_VB_METABO	0.002759	5.33e-02	8.17e-01
ENDOC0_CARENCE	-0.026850	5.24e+00	2.21e-02
ENDOC0_DIABETE	-0.006022	2.54e-01	6.14e-01
ENDOC0_DYSLIPIDEMIA	0.007428	4.13e-01	5.20e-01
ENDOC0_THYROIDE	0.010694	8.83e-01	3.47e-01
GYNECO0_ALL	-0.004790	1.54e-01	6.95e-01
KIDNEY0_2GN	-0.003281	7.92e-02	7.78e-01
KIDNEY0_2OTHER	-0.009414	5.42e-01	4.61e-01
KIDNEY0_2PYELONEPHRI	-0.015845	1.74e+00	1.87e-01
KIDNEY0_2UROLITHIASE	-0.004202	1.18e-01	7.31e-01
KIDNEY0_CYSTITE_ATCD	0.001323	1.30e-02	9.09e-01
NEURO0_SNP_METABO	0.003858	1.01e-01	7.51e-01
RESP0_2ASTHMA	0.007354	3.96e-01	5.29e-01
RESP0_2COPD	0.002180	3.36e-02	8.55e-01
RESP0_2INTERSTI	-0.011219	1.11e+00	2.92e-01
RESP0_2OTHER	-0.014264	1.38e+00	2.40e-01
RESP0_APNEE_SOM	-0.010760	8.26e-01	3.63e-01
RESP0_BRONCHITE_ATCD	-0.014716	1.68e+00	1.95e-01
RESP0_LRI_ATCD	0.013809	1.29e+00	2.57e-01
RHEUM0_ARTHROSE_HIP	0.005413	2.50e-01	6.17e-01
RHEUM0_ARTHROSE_KNEE	0.014608	1.73e+00	1.88e-01
RHEUM0_ARTHROSE_OTHE	0.012077	1.10e+00	2.95e-01
RHEUM0_AUD	-0.030019	6.19e+00	1.28e-02
RHEUM0_SYSTEME	-0.012101	1.07e+00	3.01e-01
SENSE0_CATARACTE	0.030312	7.56e+00	5.96e-03
STROKE0_2TTT	-0.007513	4.40e-01	5.07e-01
STROKE0_3noTTT	0.019909	3.29e+00	6.97e-02
STROKE0_AIT_noSEQ	0.025063	5.24e+00	2.21e-02
TRAUMA0_3FRACTURE	0.002376	4.46e-02	8.33e-01
TRAUMA0_CHUTE	-0.015259	1.74e+00	1.88e-01
TRAUMA0_OSTEOPOROSE	0.006038	2.83e-01	5.95e-01
TRAUMA0_SUICIDE	-0.004455	1.52e-01	6.96e-01
BLOOD1_1TRANSFUSION	-0.005903	2.07e-01	6.49e-01
CANCER1_AUD_SMOKER	0.008627	5.03e-01	4.78e-01
CANCER1_BREAST	-0.027408	5.42e+00	2.00e-02
CANCER1_COLORECTAL	0.015366	2.07e+00	1.50e-01
CANCER1_HEMATO	-0.021881	3.43e+00	6.40e-02
CANCER1_PC_GOOD	-0.007626	3.48e-01	5.55e-01
CANCER1_PC_POOR	-0.004538	1.34e-01	7.14e-01
CANCER1_SMOKER	-0.045598	1.11e+01	8.59e-04
CARDIO1_IHD_1MI	0.004180	1.07e-01	7.43e-01
CARDIO1_INSUF_CHRO	0.043765	1.28e+01	3.40e-04
CARDIO1_RYTHME_1ACFA	0.026826	5.12e+00	2.37e-02
CV1_MTE	-0.003709	9.35e-02	7.60e-01
CV1_PVD	-0.009660	6.48e-01	4.21e-01
DIG1_HEMORRAGIE	-0.004970	1.65e-01	6.85e-01
DIG1_LIVER_1CirrD	0.003071	5.51e-02	8.14e-01
DIG1_OCCLUSION	-0.009310	6.14e-01	4.33e-01
DIG1_PERITONITE	0.010964	7.81e-01	3.77e-01
DIG1_STOMIE	-0.004810	1.71e-01	6.79e-01
ENDOC1_GLD_OTHER	0.000633	2.57e-03	9.60e-01
ENDOC1_METABO	0.008277	3.86e-01	5.34e-01
INFECT1_SEPSIS	-0.032020	6.15e+00	1.31e-02
KIDNEY1_1INSUF_CHRO	0.010313	7.69e-01	3.80e-01
KIDNEY1_2INSUF_AIGUE	0.000708	3.44e-03	9.53e-01
NEURO1_DEM_FDR	-0.022178	3.22e+00	7.28e-02
NEURO1_EPILEPSIE	-0.026330	4.09e+00	4.32e-02
NEURO1_OTHER	0.006661	4.74e-01	4.91e-01
NEURO1_PARKINSON	-0.007325	3.99e-01	5.27e-01
RESP1_1INSUF_CHRO	0.006704	3.16e-01	5.74e-01
RESP1_2INSUF_AIGUE	0.009108	5.14e-01	4.73e-01
STROKE1_1HEMO	-0.008557	4.68e-01	4.94e-01
STROKE1_1ISCHEMIC	-0.016254	1.75e+00	1.86e-01
TRAUMA1_1CRANE	0.011846	1.16e+00	2.82e-01
TRAUMA1_2SEVERE	0.003348	8.04e-02	7.77e-01
GLOBAL		NA	2.04e+02	4.74e-11

TABLE 9 – Test du χ^2

Annexe 3 : Simulation LASSO - Univers corrélé

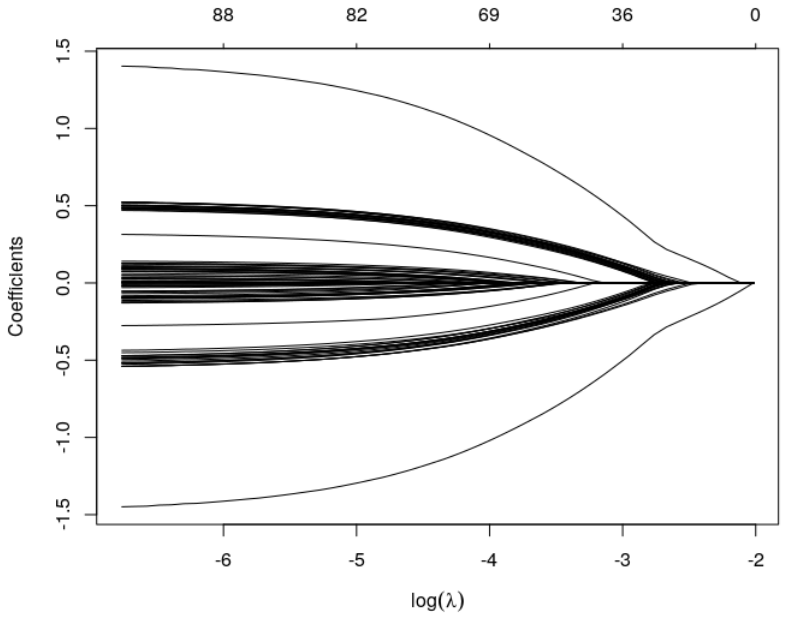


FIGURE 41 – Chemin des β

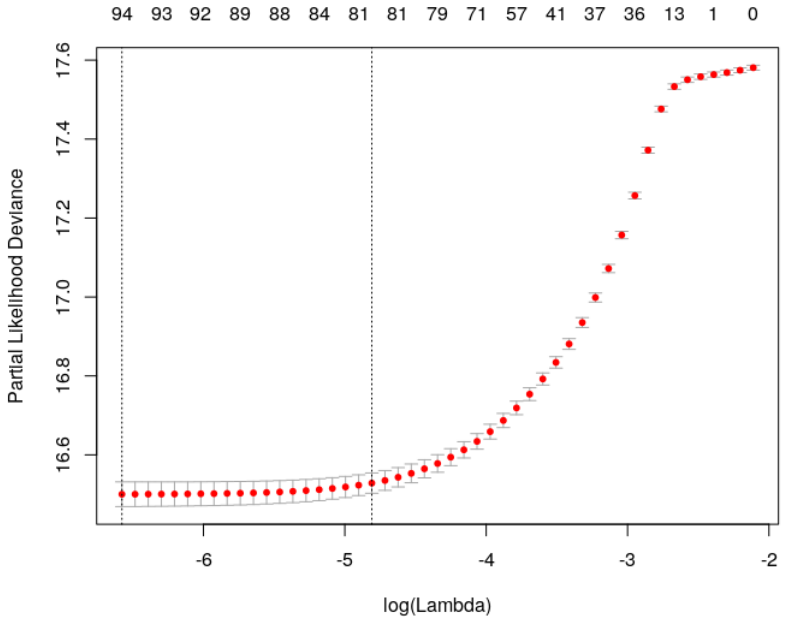
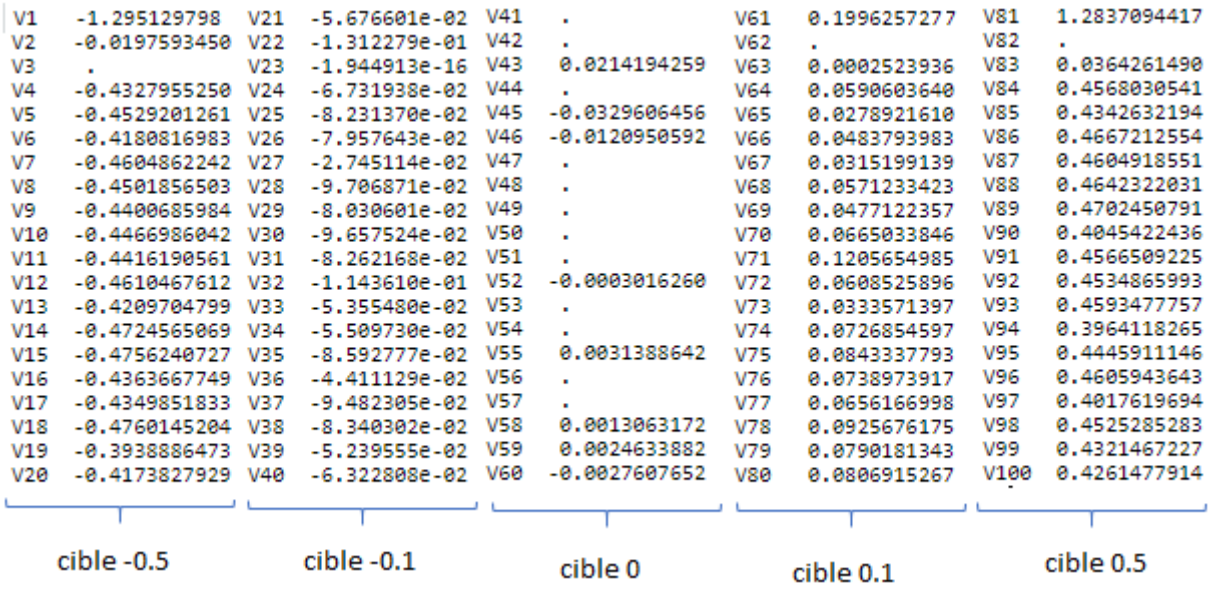


FIGURE 42 – Évolution de la déviance du modèle en fonction de la pénalisation



V1	-1.295129798	V21	-5.676601e-02	V41	.	V61	0.1996257277	V81	1.2837094417
V2	-0.0197593450	V22	-1.312279e-01	V42	.	V62	.	V82	.
V3	.	V23	-1.944913e-16	V43	0.0214194259	V63	0.0002523936	V83	0.0364261490
V4	-0.4327955250	V24	-6.731938e-02	V44	.	V64	0.0590603640	V84	0.4568030541
V5	-0.4529201261	V25	-8.231370e-02	V45	-0.0329606456	V65	0.0278921610	V85	0.4342632194
V6	-0.4180816983	V26	-7.957643e-02	V46	-0.0120950592	V66	0.0483793983	V86	0.4667212554
V7	-0.4604862242	V27	-2.745114e-02	V47	.	V67	0.0315199139	V87	0.4604918551
V8	-0.4501856503	V28	-9.706871e-02	V48	.	V68	0.0571233423	V88	0.4642322031
V9	-0.4400685984	V29	-8.030601e-02	V49	.	V69	0.0477122357	V89	0.4702450791
V10	-0.4466986042	V30	-9.657524e-02	V50	.	V70	0.0665033846	V90	0.4045422436
V11	-0.4416190561	V31	-8.262168e-02	V51	.	V71	0.1205654985	V91	0.4566509225
V12	-0.4610467612	V32	-1.143610e-01	V52	-0.0003016260	V72	0.0608525896	V92	0.4534865993
V13	-0.4209704799	V33	-5.355480e-02	V53	.	V73	0.0333571397	V93	0.4593477757
V14	-0.4724565069	V34	-5.509730e-02	V54	.	V74	0.0726854597	V94	0.3964118265
V15	-0.4756240727	V35	-8.592777e-02	V55	0.0031388642	V75	0.0843337793	V95	0.4445911146
V16	-0.4363667749	V36	-4.411129e-02	V56	.	V76	0.0738973917	V96	0.4605943643
V17	-0.4349851833	V37	-9.482305e-02	V57	.	V77	0.0656166998	V97	0.4017619694
V18	-0.4760145204	V38	-8.340302e-02	V58	0.0013063172	V78	0.0925676175	V98	0.4525285283
V19	-0.3938886473	V39	-5.239555e-02	V59	0.0024633882	V79	0.0790181343	V99	0.4321467227
V20	-0.4173827929	V40	-6.322808e-02	V60	-0.0027607652	V80	0.0806915267	V100	0.4261477914

FIGURE 43 – Coefficients obtenus pour la simulation initiale de référence

Annexe 4 : Simulation Adaptative LASSO - Univers corrélé

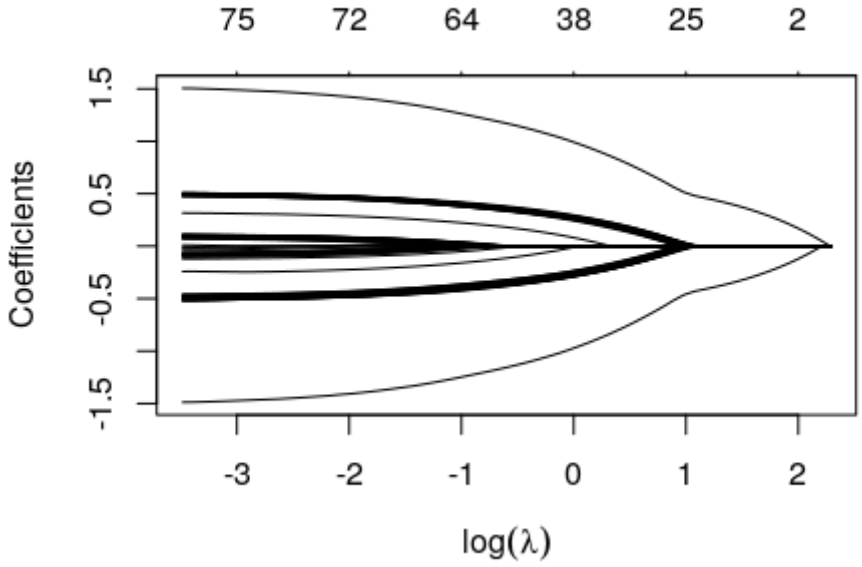


FIGURE 44 – Chemin des β

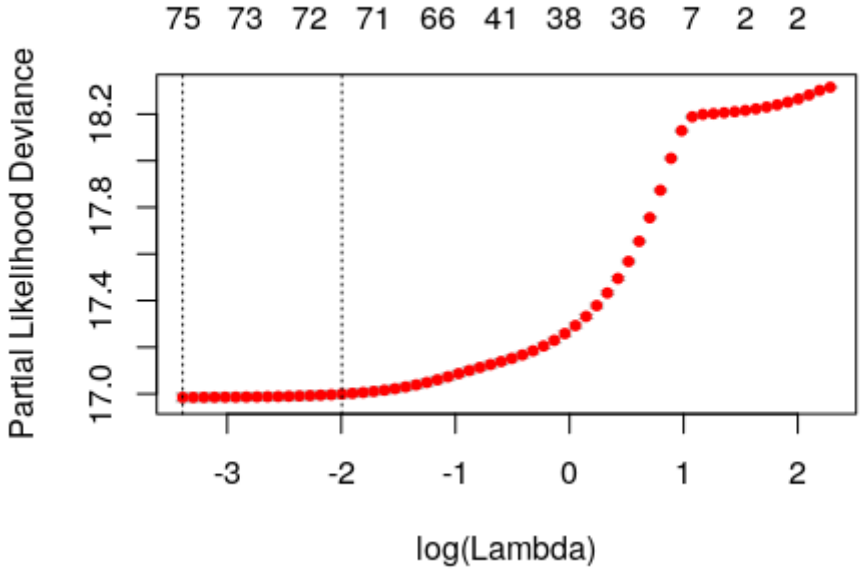


FIGURE 45 – Évolution de la déviance du modèle en fonction de la pénalisation

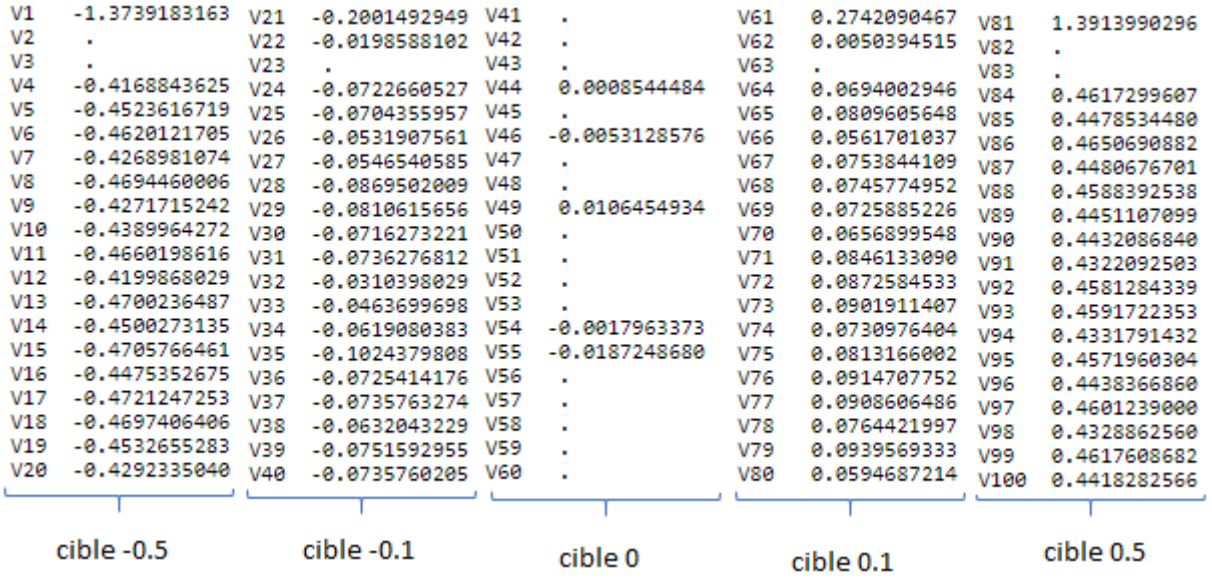


FIGURE 46 – Coefficients obtenus

Annexe 5 : Simulation Ridge - Univers corrélé

$$\{\hat{\beta}(\lambda) \in [0, +\infty[\text{ où } \hat{\beta}(\lambda) = argmin_{\beta} \prod_i^D \frac{exp(\beta^T Z_i)}{\sum_{j \in R(T_i)} exp(\beta^T Z_j)} - \lambda . \sum_{j=1}^p \beta_j^2\}$$

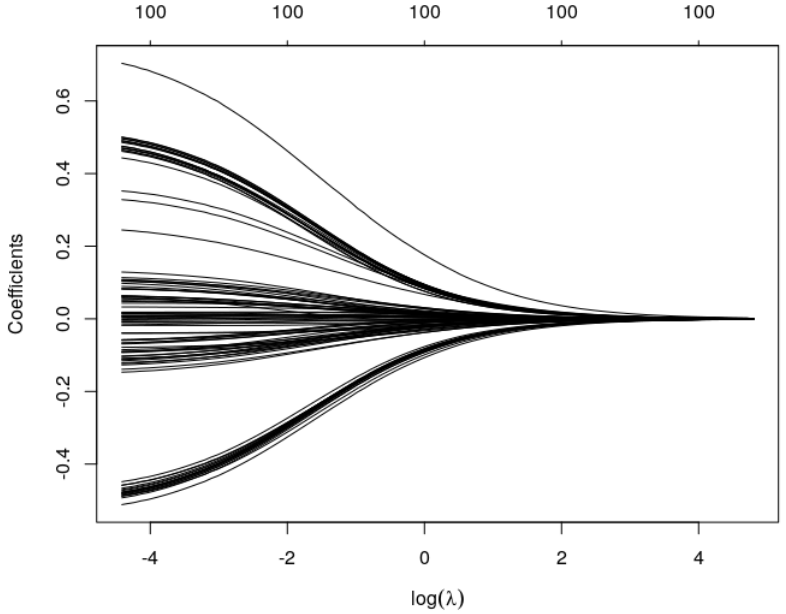


FIGURE 47 – Chemin des β
L'évolution de la déviance en fonction du paramètre de pénalisation et la valeur des coefficients obtenus pour la plus grande valeur de pénalisation éloignée d'un écart type de la pénalisation minimale :

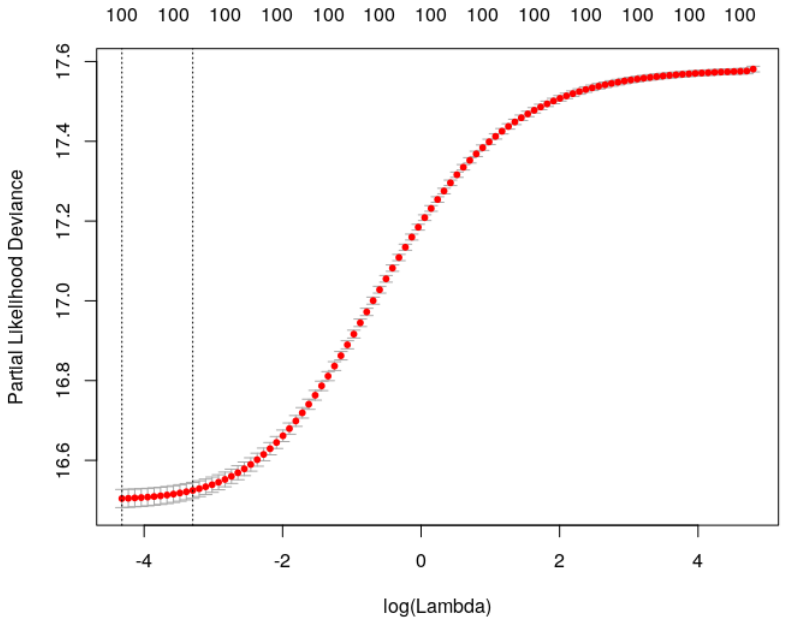


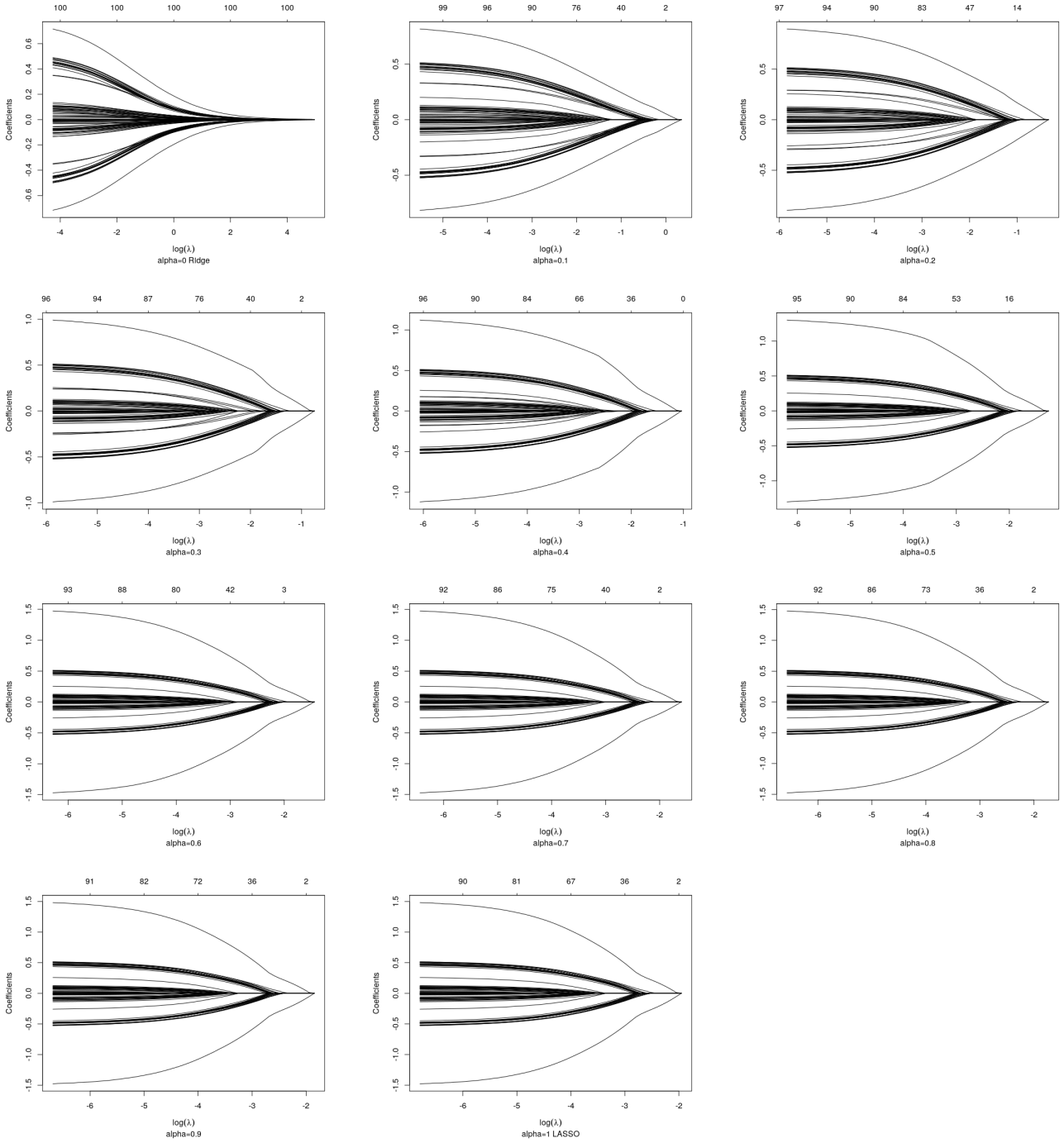
FIGURE 48 – Évolution de la déviance du modèle en fonction de la pénalisation

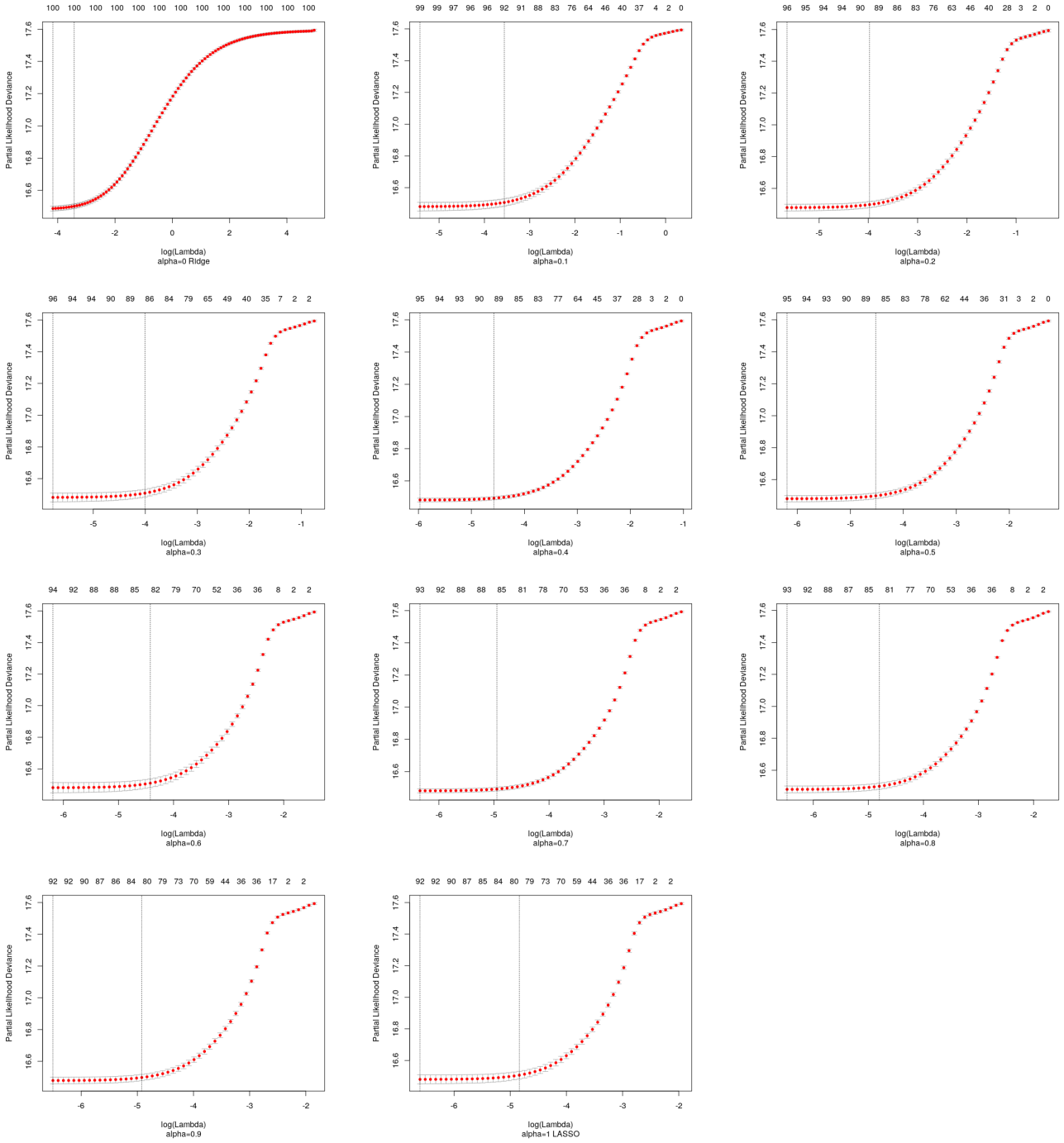
V1	-0.6582163888	V21	-0.1244010959	V41	0.0090362846	V61	0.1220650723	V81	0.6260749262
V2	-0.3266915319	V22	-0.0746031809	V42	-0.0181980050	V62	0.0503348508	V82	0.2972494793
V3	-0.3222441430	V23	-0.0504848326	V43	0.0063995773	V63	0.0686149571	V83	0.3186420648
V4	-0.4216993627	V24	-0.0815002800	V44	-0.0106529131	V64	0.0968097568	V84	0.4135293903
V5	-0.4212264346	V25	-0.1010795328	V45	0.0428134046	V65	0.1160069151	V85	0.4093959112
V6	-0.4144583932	V26	-0.0556662069	V46	0.0167191064	V66	0.0933447411	V86	0.4174078247
V7	-0.4341511843	V27	-0.0903814865	V47	0.0044418779	V67	0.0745482955	V87	0.4057733596
V8	-0.4053649688	V28	-0.1081646829	V48	0.0118687043	V68	0.0741625824	V88	0.4435013879
V9	-0.4253071046	V29	-0.0990933801	V49	-0.0132869619	V69	0.0751768540	V89	0.4380033175
V10	-0.4163728442	V30	-0.0584074449	V50	0.0127798716	V70	0.0497192876	V90	0.4177527561
V11	-0.3945121559	V31	-0.0980786462	V51	0.0317221111	V71	0.0783523706	V91	0.4052157554
V12	-0.4539955484	V32	-0.0601403382	V52	-0.0034501858	V72	0.1032227230	V92	0.4164414726
V13	-0.4040381935	V33	-0.0497213989	V53	0.0085993360	V73	0.0567749006	V93	0.4391160941
V14	-0.4107712364	V34	-0.1134322987	V54	0.0004206921	V74	0.0563513723	V94	0.4327800777
V15	-0.4287939494	V35	-0.1081341642	V55	0.0456295239	V75	0.0851046692	V95	0.4138401897
V16	-0.4213149202	V36	-0.0911029323	V56	-0.0392815150	V76	0.0480018711	V96	0.4304662830
V17	-0.4208368585	V37	-0.0814461356	V57	-0.0377999120	V77	0.0542938794	V97	0.3920449136
V18	-0.4295297518	V38	-0.0788912265	V58	0.0159798562	V78	0.0921696334	V98	0.4337030143
V19	-0.4345709943	V39	-0.0765745196	V59	0.0186061141	V79	0.0441683903	V99	0.4421183721
V20	-0.4248593040	V40	-0.0986155101	V60	-0.0078280441	V80	0.0969000989	V100	0.4310365920

cible -0.5 cible -0.1 cible 0 cible 0.1 cible 0.5

FIGURE 49 – Coefficients obtenus pour la simulation Ridge

Annexe 6 : Simulation Elastic Net - Univers corrélé





On remarque que pour une valeur de α proche de 0, on conserve l'avantage de la régression Ridge pour les variables corrélées, tout en permettant de fixer des coefficients à 0 et pouvoir ainsi opérer une sélection de variables comme avec la pénalité LASSO.

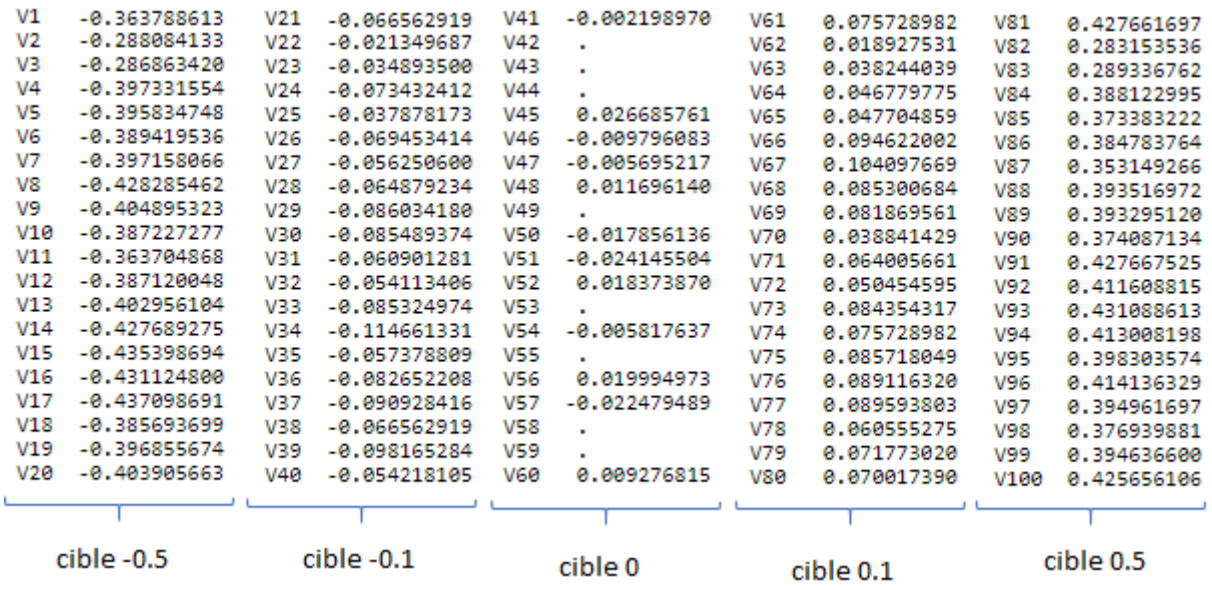


FIGURE 50 – Coefficients obtenus pour la simulation Elastic-net

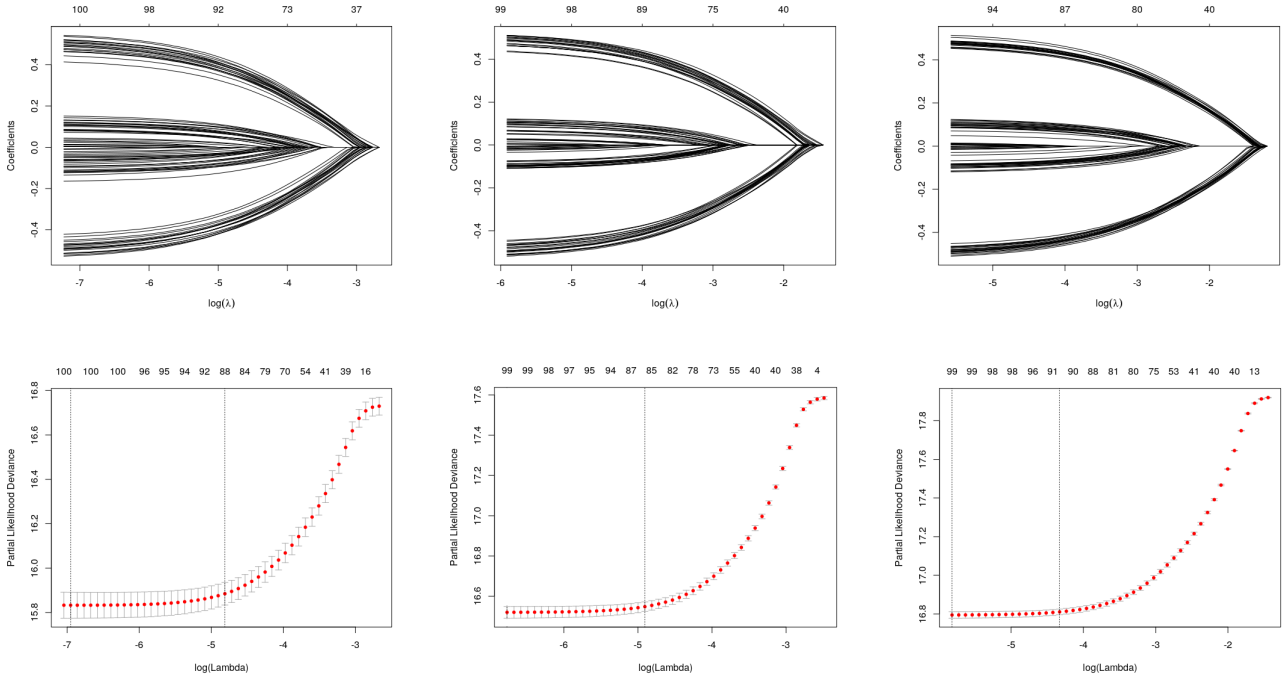
Annexe 7 : Univers censuré

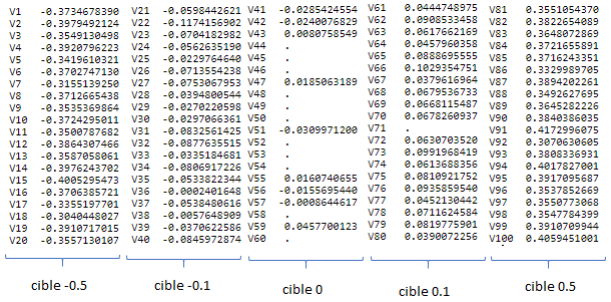
Univers censuré

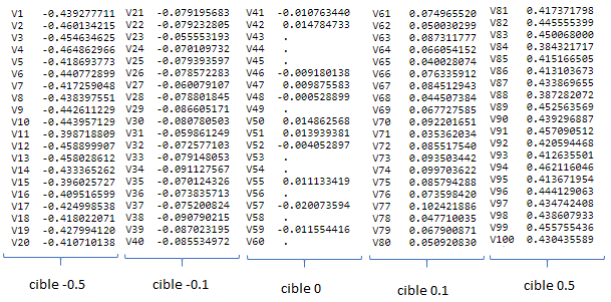
Dans cette section nous faisons varier le taux de censure.

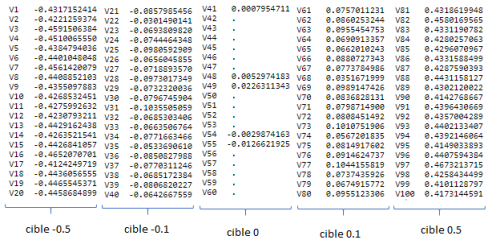
- Le nombre de covariables p est fixé à 100 ;
- Le nombre d'observations n est fixé à n = 10 000. ;
- Les taux de censure testés sont τ=70%, τ=30% et τ=0% ;
- Les covariables sont identiquement distribuées ($Y_i \in \mathcal{U}[-1;1]$) et sont indépendantes deux à deux ;
- La mesure de prédiction prise pour optimiser le paramètre de régression est la déviance. Une mesure AUC dépendante du temps sera introduite pour mettre en évidence l'influence du taux de censure sur la qualité de l'estimation ;
- Nous utilisons une 10-folds validation croisée comme méthode d'échantillonnage ;

Les graphes suivant montrent le chemin pris pour chaque coefficient et l'évolution de la déviance en fonction de la valeur du paramètre de régularisation λ pour la pénalisation LASSO avec des taux de censure de τ=70%,τ=30% et τ=0% de gauche à droite :



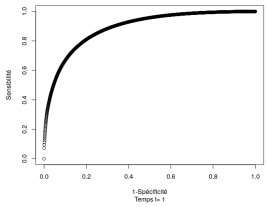


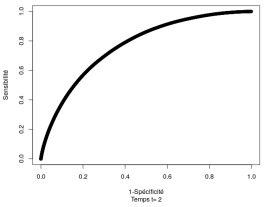


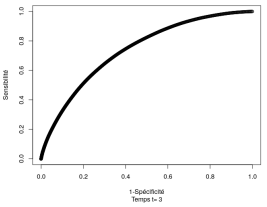


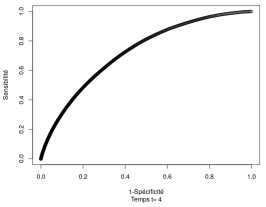
Les graphes et les valeurs des coefficients précédents indiquent que plus le taux de censure est important moins la qualité de la prédiction est bonne. Pour comparer la qualité de plusieurs modèles dans le cadre de données dépendantes du temps censurées, il a été introduit au chapitre 3 la courbe ROC dépendante du temps. Celle-ci avait été définie comme la probabilité que le score d'un individu ayant connu la survenance à t soit supérieur à celui d'un individu dont l'événement ne s'est pas encore produit à l'instant t.

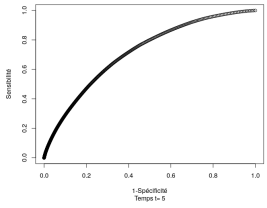
Nous traçons ci-après les courbes AUC dépendantes du temps entre $t = 0$ et $t = 30$ pour la simulation de référence.

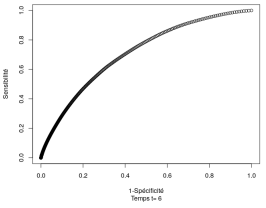

Sensibilité
1-Spécificité
Temps t= 1

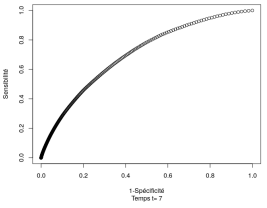

Sensibilité
1-Spécificité
Temps t= 2

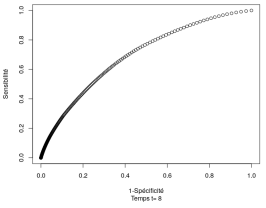

Sensibilité
1-Spécificité
Temps t= 3

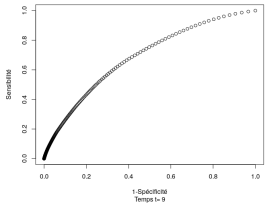

Sensibilité
1-Spécificité
Temps t= 4

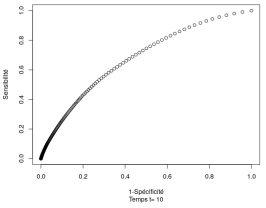

Sensibilité
1-Spécificité
Temps t= 5

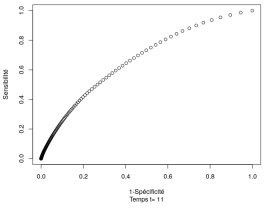

Sensibilité
1-Spécificité
Temps t= 6

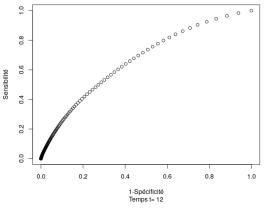

Sensibilité
1-Spécificité
Temps t= 7

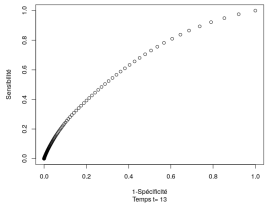

Sensibilité
1-Spécificité
Temps t= 8

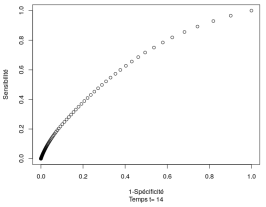

Sensibilité
1-Spécificité
Temps t= 9

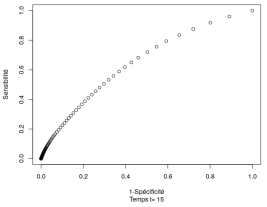

Sensibilité
1-Spécificité
Temps t= 10

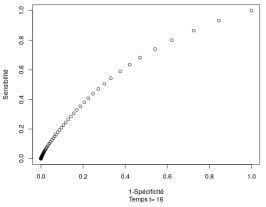

Sensibilité
1-Spécificité
Temps t= 11

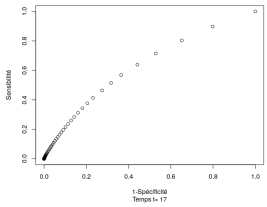

Sensibilité
1-Spécificité
Temps t= 12

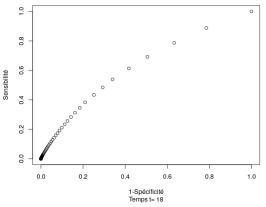

Sensibilité
1-Spécificité
Temps t= 13

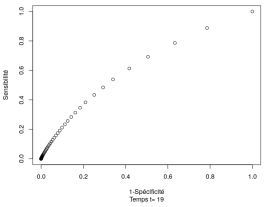

Sensibilité
1-Spécificité
Temps t= 14

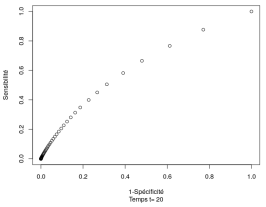

Sensibilité
1-Spécificité
Temps t= 15

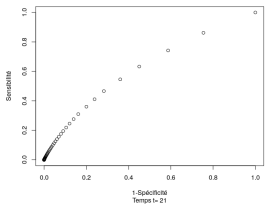

Sensibilité
1-Spécificité
Temps t= 16

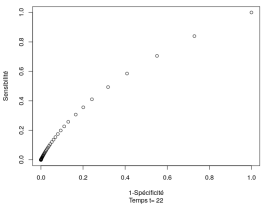

Sensibilité
1-Spécificité
Temps t= 17

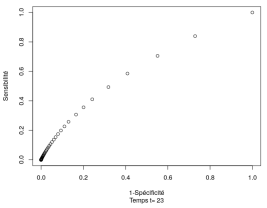

Sensibilité
1-Spécificité
Temps t= 18

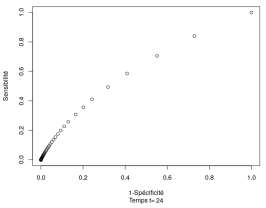

Sensibilité
1-Spécificité
Temps t= 19

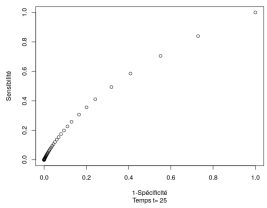

Sensibilité
1-Spécificité
Temps t= 20

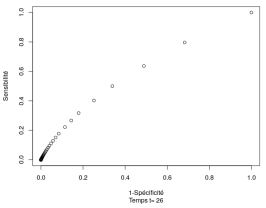

Sensibilité
1-Spécificité
Temps t= 21

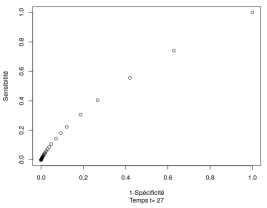

Sensibilité
1-Spécificité
Temps t= 22

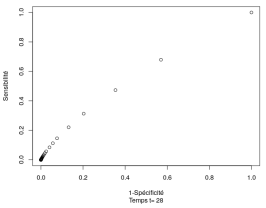

Sensibilité
1-Spécificité
Temps t= 23

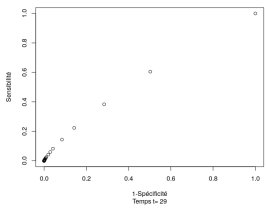

Sensibilité
1-Spécificité
Temps t= 24

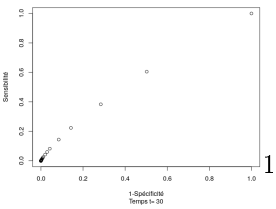

Sensibilité
1-Spécificité
Temps t= 25


Sensibilité
1-Spécificité
Temps t= 26


Sensibilité
1-Spécificité
Temps t= 27


Sensibilité
1-Spécificité
Temps t= 28


Sensibilité
1-Spécificité
Temps t= 29


Sensibilité
1-Spécificité
Temps t= 30

Par construction, les données censurées sont celles dont le temps simulé avant survenance est supérieur au temps $t_{censure}$ simulé par le biais d'une loi exponentielle dont le paramètre est choisi de manière à faire correspondre le taux de censure global avec le taux de censure souhaité. Par conséquent, plus le temps d'observation t est grand plus la probabilité que les données soient censurées est élevé.
Les courbes indiquent, que plus le temps est grand, plus la courbe ROC tend vers la première bissectrice (y=x) dont l'aire sous la courbe est 0.5. Autrement dit, vers l'estimateur dont la probabilité que le score d'un individu ayant subi l'événement est égale à celui d'un individu ne l'ayant pas subi, c'est à dire un estimateur purement aléatoire.
Ceci indique bien que plus les données sont censurées moins le modèle est performant.
Cependant, en observant la qualité du modèle (au sens de la mesure AUC) en chaque temps t et en construisant la mesure consistant à intégrer entre $[0; t_{max}]$ l'ensemble des valeurs des aires sous la courbe ROC associée à chaque instant dt, nous pourrions donner une mesure de comparaison entre différents modèles censurés. Néanmoins au vu de la complexité de mise en œuvre et étant donnée que le calcul de la vraisemblance partielle (et donc de la déviance associée) permet de prendre en compte le caractère censuré des données, la mesure de déviance a été utilisé comme mesure de prédiction.

Annexe 8 : Univers de petite, moyenne et grande dimension

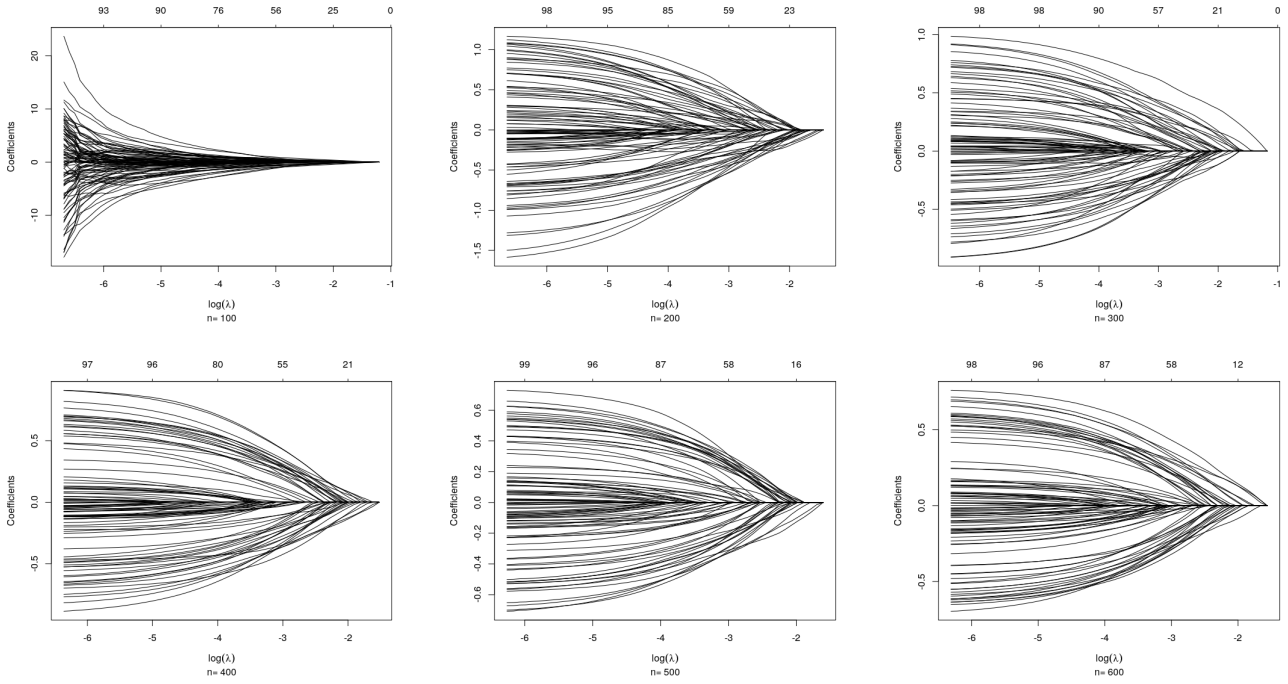
Une étude a également été menée sur l'influence du nombre d'observations afin d'étudier l'impact de la dimension de la base sur la qualité des estimateurs. Les résultats théoriques introduits dans la partie précédente sur l'évolution de l'erreur de prédiction en fonction de la taille des échantillons seront confrontés, l'intérêt de la validation croisée sera mise en pratique.

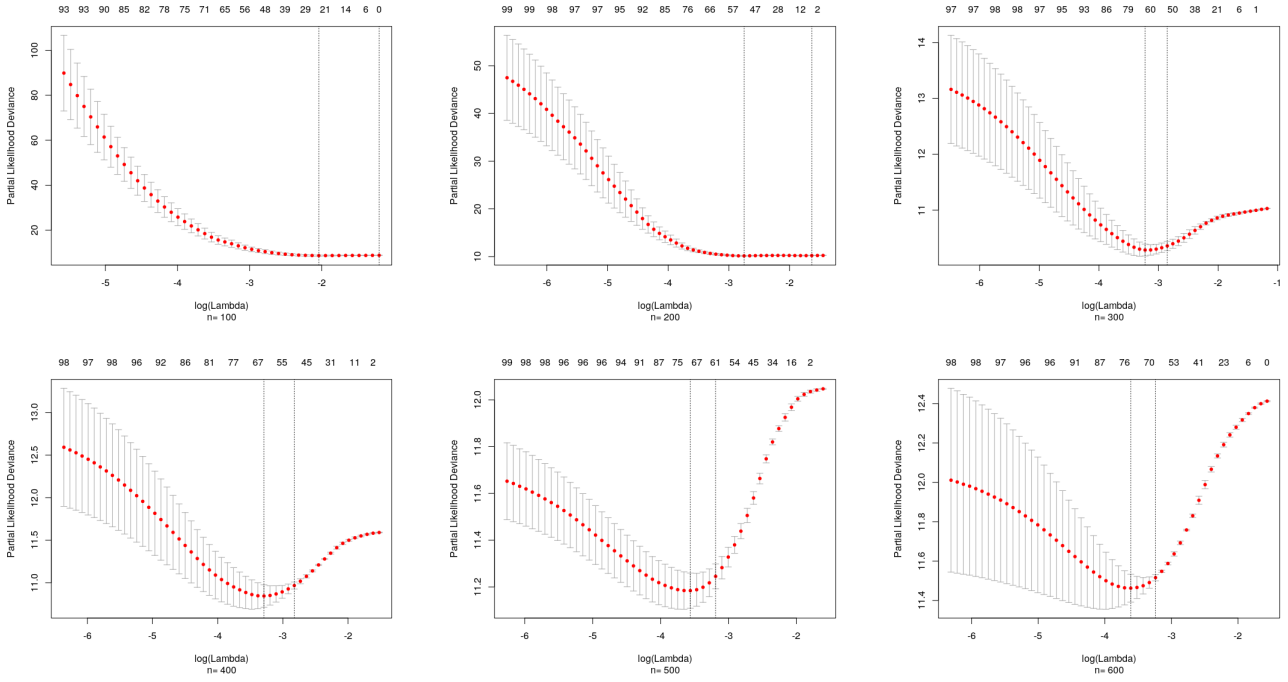
- Le nombre de covariables p est fixé à 100 ;
- Les jeux de données comprendront des tailles comprises entre n = 100 ($p \sim n$) et n=1 000 000 ($p \ll n$) ;
- Le taux de censure est fixé à τ=0% ;
- Les covariables sont identiquement distribuées ($Y_i \in \mathcal{U}[-1;1]$) et indépendantes deux à deux ;
- La mesure de prédiction prise pour optimiser le paramètre de régression est la déviance ;
- Pour le choix des échantillons d'apprentissage et de test, différentes méthodes d'échantillonnage seront comparées : de la simple resubstitution (ou l'échantillon d'apprentissage et de validation sont confondus)à la méthode de la k-folds validation croisée ;

Comme dans les parties précédentes, nous fixons le vecteur β de la façon suivante :

- les 20 premières covariables ont un poids de -0.5,
- les 20 suivantes un poids de -0.1,
- les 20 suivantes un poids nul,
- les 20 suivantes un poids de 0.1,
- les 20 dernières un poids de 0.5.

Les graphes suivants montrent le chemin pris pour chaque coefficient et l'évolution de la déviance en fonction de la valeur du paramètre de régularisation λ pour la pénalisation LASSO pour les nombres de variables $n = 100$,$n = 200$,$n = 300$,$n = 400$,$n = 500$ et $n = 600$




93 93 90 85 82 78 75 71 65 56 48 39 29 21 14 6 0
Partial Likelihood Deviance
log(Lambda)
n= 100
99 99 98 97 97 95 92 85 76 66 57 47 28 12 2
Partial Likelihood Deviance
log(Lambda)
n= 200
97 97 98 98 97 95 93 86 79 60 50 38 21 6 1
Partial Likelihood Deviance
log(Lambda)
n= 300
98 97 98 96 92 86 81 77 67 55 45 31 11 2
Partial Likelihood Deviance
log(Lambda)
n= 400
99 98 98 96 96 96 94 91 87 75 67 61 54 45 34 16 2
Partial Likelihood Deviance
log(Lambda)
n= 500
98 98 97 96 96 91 87 76 70 53 41 23 6 0
Partial Likelihood Deviance
log(Lambda)
n= 600

n =100

	cible -0.5	cible -0.1	cible 0	cible 0.1	cible 0.5
V1	.	.	.	.	.
V2	.	.	.	.	.
V3	.	.	.	.	.
V4	.	.	.	.	.
V5	.	.	.	.	.
V6	.	.	.	.	.
V7	.	.	.	.	.
V8	.	.	.	.	.
V9	.	.	.	.	.
V10	.	.	.	.	.
V11	.	.	.	.	.
V12	.	.	.	.	.
V13	.	.	.	.	.
V14	.	.	.	.	.
V15	.	.	.	.	.
V16	.	.	.	.	.
V17	.	.	.	.	.
V18	.	.	.	.	.
V19	-0.11579340	.	.	.	.
V20	.	.	.	.	.
V21		.	.	.	.
V22		.	.	.	.
V23		.	.	.	.
V24		.	.	.	.
V25		.	.	.	.
V26		.	.	.	.
V27		.	.	.	.
V28		.	.	.	.
V29		.	.	.	.
V30		.	.	.	.
V31		.	.	.	.
V32		.	.	.	.
V33		.	.	.	.
V34		.	.	.	.
V35		.	.	.	.
V36		.	.	.	.
V37		.	.	.	.
V38		.	.	.	.
V39		.	.	.	.
V40		.	.	.	.
V41			.	.	.
V42			.	.	.
V43			.	.	.
V44			.	.	.
V45			.	.	.
V46			.	.	.
V47			.	.	.
V48			.	.	.
V49			.	.	.
V50			.	.	.
V51			.	.	.
V52			.	.	.
V53			.	.	.
V54			.	.	.
V55			.	.	.
V56			.	.	.
V57			.	.	.
V58			.	.	.
V59			.	.	.
V60			.	.	.
V61				.	.
V62				.	.
V63				.	.
V64				.	.
V65				.	.
V66				.	.
V67				.	.
V68				.	.
V69				.	.
V70				.	.
V71				.	.
V72				.	.
V73				.	.
V74				.	.
V75				.	.
V76				.	.
V77				.	.
V78				.	.
V79				.	.
V80				.	.
V81					.
V82					.
V83					0.02645846
V84					.
V85					.
V86					.
V87					.
V88					.
V89					.
V90					.
V91					.
V92					.
V93					.
V94					.
V95					.
V96					.
V97					.
V98					.
V99					.
V100					.

n=200

	cible -0.5	cible -0.1	cible 0	cible 0.1	cible 0.5
V1	.	.	.	.	.
V2	.	.	.	.	.
V3	.	.	.	.	.
V4	.	.	.	.	.
V5	.	.	.	.	.
V6	.	.	.	.	.
V7	.	.	.	.	.
V8	.	.	.	.	.
V9	.	.	.	.	.
V10	-0.07709847	.	.	.	.
V11	.	.	.	.	.
V12	.	.	.	.	.
V13	-0.05123545	.	.	.	.
V14	.	.	.	.	.
V15	.	.	.	.	.
V16	.	.	.	.	.
V17	-0.03471858	.	.	.	.
V18	.	.	.	.	.
V19	.	.	.	.	.
V20	.	.	.	.	.
V21		.	.	.	.
V22		.	.	.	.
V23		.	.	.	.
V24		.	.	.	.
V25		.	.	.	.
V26		.	.	.	.
V27		.	.	.	.
V28		.	.	.	.
V29		.	.	.	.
V30		.	.	.	.
V31		.	.	.	.
V32		.	.	.	.
V33		.	.	.	.
V34		.	.	.	.
V35		.	.	.	.
V36		.	.	.	.
V37		.	.	.	.
V38		.	.	.	.
V39		.	.	.	.
V40		.	.	.	.
V41			.	.	.
V42			.	.	.
V43			.	.	.
V44			.	.	.
V45			.	.	.
V46			.	.	.
V47			.	.	.
V48			.	.	.
V49			.	.	.
V50			.	.	.
V51			.	.	.
V52			.	.	.
V53			.	.	.
V54			.	.	.
V55			.	.	.
V56			.	.	.
V57			.	.	.
V58			.	.	.
V59			.	.	.
V60			.	.	.
V61				.	.
V62				.	.
V63				.	.
V64				.	.
V65				.	.
V66				.	.
V67				.	.
V68				.	.
V69				.	.
V70				.	.
V71				.	.
V72				.	.
V73				.	.
V74				.	.
V75				.	.
V76				.	.
V77				.	.
V78				.	.
V79				.	.
V80				.	.
V81					.
V82					.
V83					.
V84					.
V85					.
V86					.
V87					.
V88					.
V89					.
V90					.
V91					.
V92					.
V93					.
V94					.
V95					.
V96					.
V97					.
V98					.
V99					.
V100					.

n=300

	cible -0.5	cible -0.1	cible 0	cible 0.1	cible 0.5
V1	-0.00464333	.	.	.	.
V2	-0.16485650	.	.	.	.
V3	-0.15556865	.	.	.	.
V4	-0.08725505	.	.	.	.
V5	-0.10876432	.	.	.	.
V6	-0.32981072	.	.	.	.
V7	-0.11306316	.	.	.	.
V8	-0.19657937	.	.	.	.
V9	-0.27019250	.	.	.	.
V10	-0.21555826	.	.	.	.
V11	-0.13978876	.	.	.	.
V12	-0.13852364	.	.	.	.
V13	-0.16003547	.	.	.	.
V14	-0.19761608	.	.	.	.
V15	-0.28437229	.	.	.	.
V16	-0.30138133	.	.	.	.
V17	-0.24835421	.	.	.	.
V18	-0.20663945	.	.	.	.
V19	-0.25947711	.	.	.	.
V20	-0.17527316	.	.	.	.
V21		.	.	.	.
V22		.	.	.	.
V23		.	.	.	.
V24		.	.	.	.
V25		.	.	.	.
V26		-0.01645436	.	.	.
V27		-0.05926005	.	.	.
V28		.	.	.	.
V29		0.03981478	.	.	.
V30		.	.	.	.
V31		.	.	.	.
V32		.	.	.	.
V33		.	.	.	.
V34		.	.	.	.
V35		.	.	.	.
V36		.	.	.	.
V37		.	.	.	.
V38		.	.	.	.
V39		.	.	.	.
V40		.	.	.	.
V41			.	.	.
V42			-0.04480226	.	.
V43			.	.	.
V44			0.06206434	.	.
V45			.	.	.
V46			.	.	.
V47			.	.	.
V48			.	.	.
V49			.	.	.
V50			.	.	.
V51			.	.	.
V52			.	.	.
V53			-0.04593531	.	.
V54			.	.	.
V55			0.05975192	.	.
V56			.	.	.
V57			-0.17312121	.	.
V58			.	.	.
V59			.	.	.
V60			.	.	.
V61				.	.
V62				.	.
V63				.	.
V64				.	.
V65				.	.
V66				.	.
V67				.	.
V68				.	.
V69				.	.
V70				.	.
V71				.	.
V72				.	.
V73				.	.
V74				.	.
V75				.	.
V76				.	.
V77				.	.
V78				.	.
V79				.	.
V80				.	.
V81					.
V82					.
V83					.
V84					.
V85					.
V86					.
V87					.
V88					.
V89					.
V90					.
V91					.
V92					.
V93					.
V94					.
V95					.
V96					.
V97					.
V98					.
V99					.
V100					.

n=400

	cible -0.5	cible -0.1	cible 0	cible 0.1	cible 0.5
V1	-0.174324018	.	.	.	0.217093081
V2	-0.326796816	.	.	.	0.196792351
V3	-0.092816705	.	.	.	0.259443231
V4	-0.192120271	.	.	.	0.259545418
V5	-0.253287931	.	.	.	0.368053701
V6	-0.279990494	.	.	.	0.222741407
V7	-0.188736149	.	.	.	0.211008199
V8	-0.232701512	.	.	.	0.139916298
V9	-0.377976471	.	.	.	0.068559412
V10	-0.256531168	.	.	.	0.186571990
V11	-0.189474825	.	.	.	0.277603778
V12	-0.297178240	.	.	.	0.161310918
V13	-0.313070477	.	.	.	0.238549072
V14	-0.167655860	.	.	.	0.300630380
V15	-0.319253353	.	.	.	0.238033000
V16	-0.190175667	.	.	.	0.158264314
V17	-0.137197896	.	.	.	0.289306392
V18	-0.109410335	.	.	.	0.240334423
V19	-0.157029978	.	.	.	0.252294567
V20	-0.264077837	.	.	.	0.368661948
V21		.	.	.	.
V22		.	.	.	.
V23		.	.	.	.
V24		.	.	.	.
V25		.	.	.	.
V26		.	.	.	.
V27		.	.	.	.
V28		.	.	.	.
V29		.	.	.	.
V30		.	.	.	.
V31		-0.016050734	.	.	.
V32		-0.086926074	.	.	.
V33		.	.	.	.
V34		.	.	.	.
V35		.	.	.	.
V36		.	.	.	.
V37		.	.	.	.
V38		-0.090463562	.	.	.
V39		-0.087262809	.	.	.
V40		.	.	.	.
V41			.	.	.
V42			.	.	.
V43			.	.	.
V44			.	.	.
V45			-0.009467674	.	.
V46			.	.	.
V47			.	.	.
V48			.	.	.
V49			.	.	.
V50			.	.	.
V51			.	.	.
V52			0.043381722	.	.
V53			.	.	.
V54			-0.076003379	.	.
V55			.	.	.
V56			.	.	.
V57			.	.	.
V58			.	.	.
V59			.	.	.
V60			.	.	.
V61				.	.
V62				.	.
V63				.	.
V64				0.084718610	.
V65				.	.
V66				.	.
V67				.	.
V68				0.006369849	.
V69				0.012213715	.
V70				.	.
V71				.	.
V72				.	.
V73				.	.
V74				.	.
V75				.	.
V76				.	.
V77				.	.
V78				.	.
V79				.	.
V80				.	.
V81					0.217093081
V82					0.196792351
V83					0.259443231
V84					0.259545418
V85					0.368053701
V86					0.222741407
V87					0.211008199
V88					0.139916298
V89					0.068559412
V90					0.186571990
V91					0.277603778
V92					0.161310918
V93					0.238549072
V94					0.300630380
V95					0.238033000
V96					0.158264314
V97					0.289306392
V98					0.240334423
V99					0.252294567
V100					0.368661948

n=500

	cible -0.5	cible -0.1	cible 0	cible 0.1	cible 0.5
V1	-0.368321788	.	.	0.105680162	0.291016496
V2	-0.290579296	-0.030601590	.	.	0.247590811
V3	-0.275633209	.	.	.	0.302533837
V4	-0.281193921	.	.	.	0.377529364
V5	-0.203581265	-0.013888744	.	0.159025705	0.339305573
V6	-0.147110774	.	.	0.027095865	0.329464503
V7	-0.183920926	.	.	0.143822023	0.350790266
V8	-0.364632736	-0.009340816	.	0.111792438	0.286357255
V9	-0.367749701	.	.	.	0.257178919
V10	-0.248862904	-0.081991545	.	.	0.288829371
V11	-0.339367051	.	.	0.049932227	0.300663101
V12	-0.241907433	-0.094551375	.	.	0.246424281
V13	-0.335632945	-0.074763208	.	0.002539901	0.274472093
V14	-0.281419700	-0.049778934	.	.	0.114034421
V15	-0.158673995	.	.	.	0.303780587
V16	-0.163681651	-0.085417554	.	.	0.316939709
V17	-0.246836335	.	.	.	0.218691821
V18	-0.226216344	.	.	0.045804920	0.093348859
V19	-0.337548624	-0.080997057	.	0.081872264	0.234453098
V20	-0.271492317	-0.049991985	0.026233528	.	0.319903069
V21		.	.	.	.
V22		-0.030601590	.	.	.
V23		.	.	.	.
V24		.	.	.	.
V25		-0.013888744	.	.	.
V26		.	.	.	.
V27		.	.	.	.
V28		-0.009340816	.	.	.
V29		.	.	.	.
V30		-0.081991545	.	.	.
V31		.	.	.	.
V32		-0.094551375	.	.	.
V33		-0.074763208	.	.	.
V34		-0.049778934	.	.	.
V35		.	.	.	.
V36		-0.085417554	.	.	.
V37		.	.	.	.
V38		.	.	.	.
V39		-0.080997057	.	.	.
V40		-0.049991985	.	.	.
V41			.	.	.
V42			-0.031750166	.	.
V43			.	.	.
V44			.	.	.
V45			.	.	.
V46			.	.	.
V47			.	.	.
V48			.	.	.
V49			.	.	.
V50			.	.	.
V51			.	.	.
V52			.	.	.
V53			.	.	.
V54			.	.	.
V55			.	.	.
V56			.	.	.
V57			.	.	.
V58			.	.	.
V59			.	.	.
V60			0.026233528	.	.
V61				0.105680162	.
V62				.	.
V63				.	.
V64				.	.
V65				0.159025705	.
V66				0.027095865	.
V67				0.143822023	.
V68				0.111792438	.
V69				.	.
V70				.	.
V71				0.049932227	.
V72				.	.
V73				0.002539901	.
V74				.	.
V75				.	.
V76				.	.
V77				.	.
V78				0.045804920	.
V79				0.081872264	.
V80				.	.
V81					0.291016496
V82					0.247590811
V83					0.302533837
V84					0.377529364
V85					0.339305573
V86					0.329464503
V87					0.350790266
V88					0.286357255
V89					0.257178919
V90					0.288829371
V91					0.300663101
V92					0.246424281
V93					0.274472093
V94					0.114034421
V95					0.303780587
V96					0.316939709
V97					0.218691821
V98					0.093348859
V99					0.234453098
V100					0.319903069

n=600

	cible -0.5	cible -0.1	cible 0	cible 0.1	cible 0.5
V1	-0.368321788	.	.	0.105680162	0.291016496
V2	-0.290579296	-0.030601590	-0.031750166	.	0.247590811
V3	-0.275633209	.	.	.	0.302533837
V4	-0.281193921	.	.	.	0.377529364
V5	-0.203581265	-0.013888744	.	0.159025705	0.339305573
V6	-0.147110774	.	.	0.027095865	0.329464503
V7	-0.183920926	.	.	0.143822023	0.350790266
V8	-0.364632736	-0.009340816	.	0.111792438	0.286357255
V9	-0.367749701	.	.	.	0.257178919
V10	-0.248862904	-0.081991545	.	.	0.288829371
V11	-0.339367051	.	.	0.049932227	0.300663101
V12	-0.241907433	-0.094551375	.	.	0.246424281
V13	-0.335632945	-0.074763208	.	0.002539901	0.274472093
V14	-0.281419700	-0.049778934	.	.	0.114034421
V15	-0.158673995	.	.	.	0.303780587
V16	-0.163681651	-0.085417554	.	.	0.316939709
V17	-0.246836335	.	.	.	0.218691821
V18	-0.226216344	.	.	0.045804920	0.093348859
V19	-0.337548624	-0.080997057	.	0.081872264	0.234453098
V20	-0.271492317	-0.049991985	0.026233528	.	0.319903069
V21		.	.	.	.
V22		-0.030601590	.	.	.
V23		.	.	.	.
V24		.	.	.	.
V25		-0.013888744	.	.	.
V26		.	.	.	.
V27		.	.	.	.
V28		-0.009340816	.	.	.
V29		.	.	.	.
V30		-0.081991545	.	.	.
V31		.	.	.	.
V32		-0.094551375	.	.	.
V33		-0.074763208	.	.	.
V34		-0.049778934	.	.	.
V35		.	.	.	.
V36		-0.085417554	.	.	.
V37		.	.	.	.
V38		.	.	.	.
V39		-0.080997057	.	.	.
V40		-0.049991985	.	.	.
V41			.	.	.
V42			-0.031750166	.	.
V43			.	.	.
V44			.	.	.
V45			.	.	.
V46			.	.	.
V47			.	.	.
V48			.	.	.
V49			.	.	.
V50			.	.	.
V51			.	.	.
V52			.	.	.
V53			.	.	.
V54			.	.	.
V55			.	.	.
V56			.	.	.
V57			.	.	.
V58			.	.	.
V59			.	.	.
V60			0.026233528	.	.
V61				0.105680162	.
V62				.	.
V63				.	.
V64				.	.
V65				0.159025705	.
V66				0.027095865	.
V67				0.143822023	.
V68				0.111792438	.
V69				.	.
V70				.	.
V71				0.049932227	.
V72				.	.
V73				0.002539901	.
V74				.	.
V75				.	.
V76				.	.
V77				.	.
V78				0.045804920	.
V79				0.081872264	.
V80				.	.
V81					0.291016496
V82					0.247590811
V83					0.302533837
V84					0.377529364
V85					0.339305573
V86					0.329464503
V87					0.350790266
V88					0.286357255
V89					0.257178919
V90					0.288829371
V91					0.300663101
V92					0.246424281
V93					0.274472093
V94					0.114034421
V95					0.303780587
V96					0.316939709
V97					0.218691821
V98					0.093348859
V99					0.234453098
V100					0.319903069

Lorsque le nombre de variables est égal au nombre d'observations, le chemin des coefficients est très mauvais, la procédure de sélection du paramètre par apprentissage est chaotique, le λ choisi par la méthode rejette quasiment tous les coefficients. Les résultats deviennent acceptables quand le nombre de variables est supérieur à $n = 500$.

En appliquant la méthode AIC au même jeu de données de taille $n = 500$, on obtient les coefficients suivants :

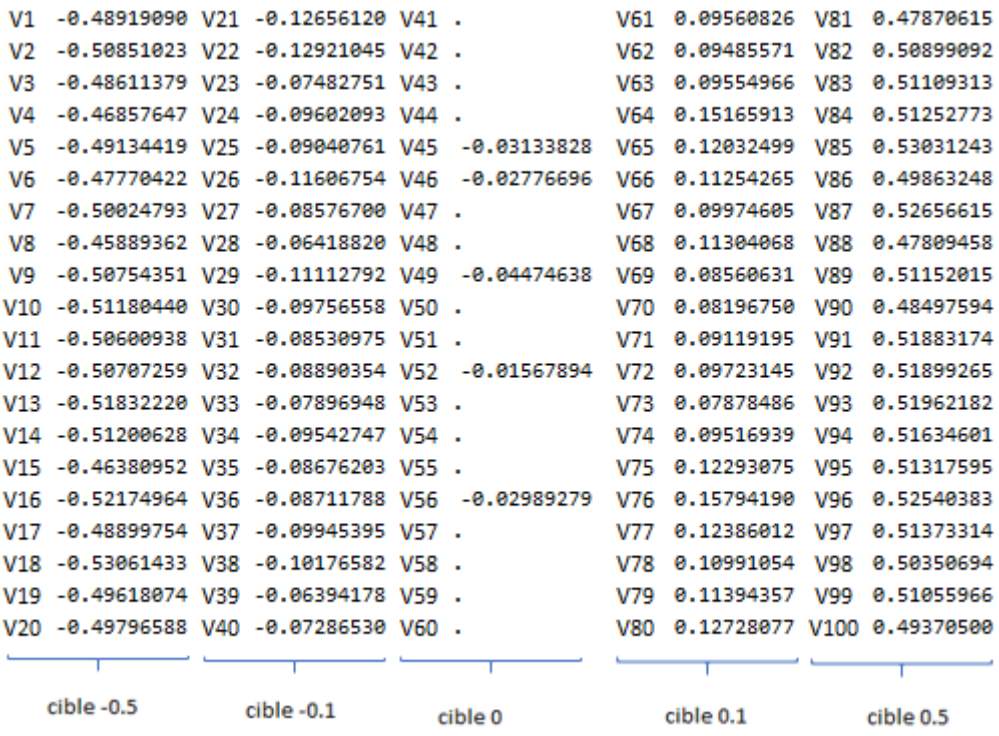


FIGURE 51 – Coefficients obtenus pour la simulation initiale de référence avec la méthode AIC

En termes de sélection, les deux méthodes sont équivalentes puisqu'elles sélectionnent 85 variables dont 5 variables non-souhaitées.
En termes de justesse des coefficients la méthode AIC est plus performante. En effet, en calculant la moyenne des erreurs quadratiques des deux ensembles de coefficients, on obtient respectivement $3.5 * 10^{-5}$ et $4.1 * 10^{-3}$.
En terme de temps de calcul, la méthode AIC met $4min\ 40sec$ à s'exécuter alors que la méthode LASSO prend $9,1sec$.

La méthode classique donne une fois de plus de meilleurs résultats que la régression pénalisée lorsque la dimension de la base augmente mais impose des temps de calculs plus long ($10sec$ pour la régression pénalisée contre $2min$ pour la méthode AIC).

Pour mesurer l'influence de la méthode d'échantillonnage, 100 simulations sont effectuées et l'erreur quadratique moyenne est calculée pour différentes valeurs du nombre de subdivisions k de la validation croisée et pour différentes tailles d'échantillons. Les temps de calcul moyens sont également renseignés pour chaque valeur de k.

n=100 ; p=100		
k-folds	Erreur standard moyenne	Temps de calcul moyen (secondes)
3	3,194021	2,852397
5	3,183584	3,485241
10	3,173434	5,174535
20	3,174621	8,031446
30	3,168838	10,756462

n=200 ; p=100		
k-folds	Erreur standard moyenne	Temps de calcul moyen (secondes)
3	3,031405	0,4088982
5	2,913024	0,4569946
10	2,736206	0,7182674
20	2,673622	1,3311388
30	2,603157	1,8998653

n=500 ; p=100		
k-folds	Erreur standard moyenne	Temps de calcul moyen (secondes)
3	1,548425	0,3152818
5	1,475991	0,4598259
10	1,419287	0,8382729
20	1,413432	1,5813957
30	1,438743	2,3232574

n=1 000 ; p=100		
k-folds	Erreur standard moyenne	Temps de calcul moyen (secondes)
3	1,10521	0,4699412
5	1,041018	0,6993494
10	1,01678	1,2858571
20	1,009006	2,4493223
30	0,99943	3,6356007

n=10 000 ; p=100		
k-folds	Erreur standard moyenne	Temps de calcul moyen (secondes)
3	0.4061328	3.221832
5	0.4231844	4.937101
10	0.4145708	9.169105
20	0.4201639	17.690297
30	0.4206514	26.309024

FIGURE 52 – Evolution de l'erreur en prédiction en fonction de la taille et de la technique d'échantillonnage

Comme attendu, plus le nombre de subdivisions est élevé, plus l'estimation est juste. En effet, à tailles égales, plus le paramètre k est élevé, plus l'erreur est faible. De même, plus la base de donnée est volumineuse, plus l'estimation est juste. On constate en effet qu'à paramètres k égaux, l'erreur de prédiction diminue à mesure que la taille de l'échantillon augmente.
On remarque par ailleurs que le choix de subdiviser la base totale en 10 sous échantillons semble pertinent au regard du compromis entre les temps de calcul et la qualité d'ajustement. On remarque également que plus la base est volumineuse, moins le nombre de subdivisions participe à améliorer la justesse de la prédiction. Dans le cas ou $n = 10000$ on constate qu'au delà de 10 subdivisions, la justesse d'ajustement semble constante. Le choix classique de subdiviser la validation croisée en 10 subdivisions semble adéquat.

Bibliographie

[1] ANDERSEN, P.K. & GILL R.D [1982]. *Cox's regression model for counting processes :a large sample study.* Ann. Statist. 10.

[2] BACCHETTA Jean-Pierre [2005]. *Critères diagnostics de la démence vasculaire : étude de validité dans une population de nonagénaires et centenaires.* Thèse pour la Faculté de Médecine De l'Université de Genève.

[3] BICKEL P.J. & *al.* [1993] *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Series in the Mathematical Sciences. Baltimore, MD : Johns Hopkins University Press.

[4] BONNETIER Delphine[2010]. *Analyses de survie sur données transcriptomiques.*

[5] BRESLOW, N. [1974]. *Covariance Analysis of Censored Survival Data.* Biometrics 30

[6] EL ANBARI M. [2012]. *Régularisation et sélection de variables par le biais de la vraisemblance pénalisée.*HAL archives-ouvertes.fr

[7] FAN Jianging & LI Runze [2002]. *Variable selection for Cox's proportional hazards model and frailty model.* The Annals of Statistics, volume 30, numéro 1, pages 74-99]

[8] FANG H.-B.& *al.* [2005]. *Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model.*Scand. J. Statist.32

[9] GERMAIN J.F [2010]. *Sélection de modèles à l'aide des chemins de régularisation pour l'objectivation mono et multi-prestations. Application à l'agrément de conduite.*HAL archives-ouvertes.fr

[10] HASTIE Trevor & *al.* [2016]. *Statistical Learning with Sparsity The Lasso and Generalizations.* CRC Press.

[11] HASTIE T. & *al.* [2001] *The elements of statistical learning. Data mining, inference, and prediction.* Springer Series in Statistics. New York : Springer-Verlag.

[12] HEAGERTY Patrick & ZHENG Yingye [2003]. *Survival Model Predictive Accuracy and ROC Curves.* Dalloz.

[13] KLEIN J. P. & MOESCHBERGER M.L [1997]. *Survival Analysis : Methods for Censored and Truncated Data. Statistics for Biology and Health.* New York : Springer.

[14] KNIGHT Keith & FU Wenjiang [2002]. *Asymptotics for Lasso-type estimators.* The Annals of Statistics, volume 28, numéro 5, pages 1356-1378].

[15] LOPEZ Olivier & Milhaud Xavier [2016]. *Tree-based censored regression with applications in insurance.*

[16] LIBAULT Dominique [2019]. *175 propositions pour une politique nouvelle et forte du grand âge en France.*

[17] PLANCHET Frédéric & THEROND Pierre[2000]. *Modélisation statistique des phénomènes de durée.* Economica.

[18] PLANCHET Frédéric & GUIBERT Quentin & SCHWARZINGER Michaël [2019]. *Mesure du risque de perte d'autonomie totale en France métropolitaine.* Laboratoire SAF.

[19] PLANCHET Frédéric & GUIBERT Quentin & SCHWARZINGER Michaël [2019]. *Lois biométriques pour le risque de perte d'autonomie en France.* Groupe de travail QalyDays

[20] SCHWARZINGER Michaël [2018]. *Contribution of alcohol use disorders to the burden of dementia in France 2008–13 : a nationwide retrospective cohort study.* Lancet Public Health.

[21] SCHWARZINGER Michaël [2018]. *Données source et retraitements pour l'étude du risque de perte d'autonomie*. Laboratoire SAF.

[22] VERWEIJ PJ [1996]. *Penalized likhood in Cox regression*.

[23] VILLANI Cedric [2018]. *Donner un sens à l'intelligence artificielle (IA)*.
[1996]