Etude de la longueur des télomères des leucocytes

Pol Labarbarie [1] and Mehdy Hounkonnou [1]

[1]Master 2 Modélisation Statistique et Stochastique, Université de Bordeaux

8 janvier 2023

Table des matières

I. Introduction

Les télomères sont des séquences répétitives de nucléotides situées aux extrémités des chromosomes linéaires de la plupart des organismes eucaryotes. Il ne code pas pour une information précise mais intervient dans la stabilité du chromosome et dans les processus de vieillissement cellulaire. Les télomères servent à protéger les chromosomes et participent à l'intégrité du patrimoine génétique. Des recherches ont mis en évidence les associations entre la longueur du télomère et la longévité, ainsi que les troubles liés à l'âge et la tumorigenèse (Eitan, Hutchison et Mattson 2014 ; Haycock et al. 2017 ; De Meyer et al. 2018). Par conséquent, il semble être très important d'explorer les déterminants environnementaux et génétiques de la longueur individuelle des télomères. D'autant plus que la longueur des télomères est sensible aux facteurs environnementaux de stress. Dans ce rapport, en se basant sur l'étude de (Bai et al. 2020), nous allons analyser les effets de la co-exposition à plusieurs métaux ainsi que leurs effets conjoints avec les variantes de TERT-CLPTM1L sur la longueur des télomères des leucocytes (LTL). Nous allons comparer diverses méthodes de selection de variables adaptées à la grande dimension sur un critère de prédiction et sur un critère d'identification, à partir des données simulées et en tenant en compte de la possible linéarité des prédicteurs. Ici, notre variable réponse est LTL (le logarithme népérien de la longueur des télomères des leucocytes en nmol/L). Cette variable est générée à partir de l'âge des individus (AGE), du nombre d'années de travail (TRAVAIL), et des métaux Manganèse, Aluminium et Vanadium. Les valeurs des autres 20 métaux listés dans l'article, l'indice de masse corporelle des individus (IMC) (kg/m2), le sexe des individus (H=0/F=1), le statut tabagique des individus (Jamais=0/Fumeur=1), et la consommation d'alcool des individus (Jamais=0/Buveur=1) sont également disponibles. Nous allons réaliser deux études. La première avec uniquement les métaux comme prédicteurs, et la seconde en regroupant les métaux et les facteurs de confusion. Pour ces deux études, nous appliquerons une transformation polynomiale sur nos variables métaux.

II. Méthodes

Tout d'abord, nous allons utiliser la base de fonctions des polynômes d'ordre 3 afin de créer de nouvelles variables pour les métaux. Chacun des métaux sera donc représenter avec 3 variables polynomiales. Dans la suite, si une des 3 composantes d'un métal est sélectionnée la variable sera jugée pertinente et si aucune de ses composantes n'est sélectionnée la variable sera jugée comme inintéressante. Après avoir réalisé une partition aléatoire des données i.e deux tiers pour l'apprentissage et un tiers pour le test nous allons utiliser et comparer les différentes méthodes présentées ci-dessous. Chacune de ces méthodes sera comparée en comparant leur erreur de prédiction et les variables qui ont été retenues pour faire cette prédiction. Malgré que notre travail ne se trouve pas dans un cadre de grande dimension, il faut rappeler néanmoins que les méthodes que nous avons utilisé trouve leur intérêt dans ce cadre la. Même si l'ensemble de nos méthodes se basent sur un cadre de régression elles se distinguent en trois catégories (même si l'idée derrière chacune est commune). La première catégorie se base sur une sélection de variables à partir d'un critère, nous retrouvons la méthode se basant sur l'AIC et les méthodes de test (Bonferonni, et Benjamini et Hochberg). Une fois cette sélection de variables effectuée à partir de ce critère, nous effectuons une régression. La seconde catégorie se base sur une pénalisation de notre régression. Nous retrouvons les méthodes Lasso et Ridge. Et finalement, la troisième catégorie se base sur une réduction de dimensionalité, généralement par une ACP (Analyse en Composantes Principales). Une fois cette réduction de dimensionalité faite, modulo les particularités de certaines méthodes, nous effectuons notre régression sur nos nouvelles variables. Nous retrouvons les méthodes PCR, PLS et sPLS. Dans un objectif d'interprétabilité de nos résultats, nous avons standardisé nos données. Présentons maintenant de façon succincte chacune de ces méthodes.

I. Méthode de sélection de variables sur critère

Dans cette sous-partie, nous allons présenter les méthodes de sélection de variables qui se base sur un critère précis. Tout d'abord, présentons la méthode se basant sur le critère AIC (Akaïke's Information criterion). Ce critère est défini par

$$\text{AIC} = -2\log(L) + 2k \tag{1}$$

où L est le maximum de la fonction de vraisemblance du modèle et k est le nombre de paramètres de notre modèle. L'hypothèse forte de ce critère est que notre modèle est paramétrique étant donné que nous avons utilisé la fonction de vraisemblance. Afin de diminuer le temps de calcul, nous utilisons une procédure de recherche ascendante pas à pas du meilleur modèle. Le modèle retenu est celui dont l'AIC est la plus faible, c'est à dire le modèle qui minimise l'information perdue en représentant nos données par ce modèle. Nous utilisons une procédure ascendante et non descendante car la procédure descendante conduit à des valeurs de l'AIC non définies (infini). La procédure ascendante consiste à partir du modèle plus simple et au fur et à mesure à ajouter les variables qui vont minimiser l'AIC du modèle. Finalement, nous retiendrons les variables sélectionnées par ce critère pour construire notre modèle linéaire. Présentons maintenant les méthodes se basant sur des tests d'hypothèses. Les procédures de Bonferroni et de Benjamini-Hochberg sont des méthodes de correction du risque lors de tests statistiques multiples. Ces méthodes trouvent tout leur intérêt lorsque nous sommes confrontés à des situations où un grand nombre de paramètres sont testés. Dans notre problème nous testons la nullité des coefficients de notre régression en utilisant un test de Student. Lors de ces tests multiples, il apparaît un phénomène d'inflation du risque d'erreur de première espèce. Afin de remédier à ce problème il existe deux généralisations du risque d'erreur de première espèce. Ce sont le FWER (FamilyWise Error Rate) et le FDR (False Discovery Rate). Le FWER est la probabilité de commettre au moins une erreur de première espèce, tandis que le FDR est défini comme l'espérance de la proportion d'erreurs de première espèce parmi les hypothèses rejetées. La méthode de Bonferroni est une procédure en une étape qui compare les p-valeurs non ajustées à un seuil commun $\frac{\alpha}{l}$ où α est le risque d'erreur de première espèce et l est le nombre d'hypothèses testées (ici $l = k$ car nous avons autant d'hypothèse que de variable). Cette procédure contrôle le FWER. La méthode de Benjamini-Hochberg est une procédure séquentielle ascendante, dont l'objectif est de contrôler le FDR. Cette méthode consiste à ordonner les p-valeurs par ordre croissant, puis nous calculons $m = \max\{1 \leq j \leq l; p_j \leq \frac{j\alpha}{l}\}$, si m existe, alors nous rejettons les hypothèses nuelles $H_{0_1}, ..., H_{0_m}$, sinon, nous n'en rejettons aucune. Toutes les méthodes que nous avons vu dans cette sous-partie permettent de faire de la sélection de variables.

II. Méthode de pénalisation

Dans cette sous-partie, nous allons expliquer les méthodes de pénalisation de notre régression. Les méthodes Ridge et Lasso, sont des méthodes visant à trouver (β, λ) où β est un vecteur de coefficients associé à nos variables et λ un réel positif tels que

$$\min_{\beta,\lambda} \sum_{i=1}^{n} \left(y_i - \sum_{k=0}^{m} \beta_k X_{ik} \right)^2 + \lambda ||\beta||_q^q \tag{2}$$

Tout d'abord, nous remarquons facilement que pour $\lambda = 0$, nous retrouvons l'estimateur des moindres carrés. La méthode Lasso est quand le paramètre $q = 1$ et la méthode Ridge est quand le paramètre $q = 2$. Au vue de la valeur de q, la pénalisation appliquée permet à la méthode Lasso contrairement à la méthode Ridge d'effectuer de la sélection de variables. En effet, la régression Ridge conserve toutes les variables mais, contraignant la norme des paramètres β_k, elle les empêche de prendre de trop grandes valeurs et limite ainsi la variance des prévisions. Seule la méthode Lasso permet donc d'obtenir un modèle avec moins de variables et donc plus parcimonieux. Pour ces deux méthodes, nous ne pouvons comparer les coefficients β_k qui les données ont été préalablement standardisé. Pour nos deux méthodes, le paramètre λ sera sélectionné par validation croisée, et nous conserverons le λ qui minimise l'erreur quadratique de prédiction.

III. Méthode de réduction de dimensionnalité

Dans cette sous-partie, nous allons présenter les méthodes de réduction de dimensionnalité. Ces méthodes se basent sur une ACP normée ou sur une variation d'une ACP. La régression sur composantes principales appelée PCR consiste à tout d'abord, construire des nouvelles variables

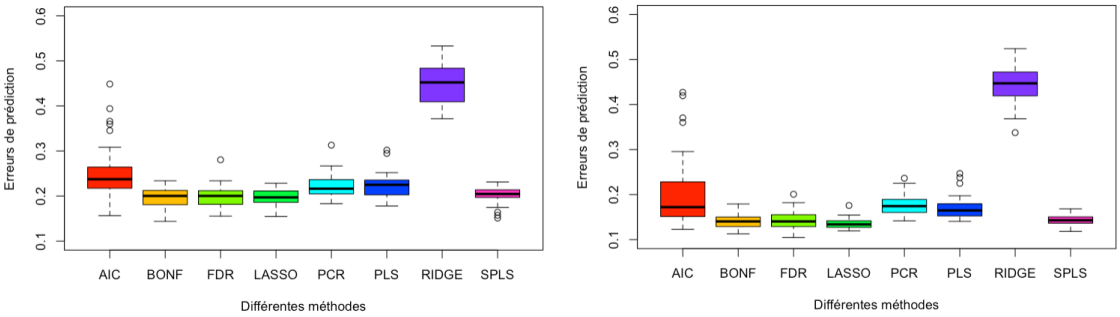
$Z_1, ..., Z_r$ qui sont les composantes principales de la matrice X de nos variables de départ. Une fois ces nouvelles variables obtenues, nous construisons le prédicteur PCR défini comme suit par

$$\hat{Y} = \sum_{m=1}^{r} \hat{\theta}_m Z^m \qquad \hat{\theta}_m = \frac{\langle Z^m, Y \rangle}{|Z^m|^2} \tag{3}$$

Le principal problème posé par la PCR est que les premières composantes, associées aux plus grandes valeurs propres, ne sont pas nécessairement corrélées avec Y et ne sont donc pas nécessairement les meilleures candidates pour résumer ou modéliser Y. Afin de palier à ce problème nous pouvons utiliser la régression PLS. Cette méthode se base également sur la même construction de nouvelles variables par une ACP normée, mais une contrainte est rajoutée afin que nos nouvelles variables synthétisées soient le plus liées à Y au sens de la covariance empirique. Le défaut de la régression PLS est qu'il est difficile d'effectuer une analyse statistique qui nous permettrait d'obtenir les lois de nos estimateurs. Ensuite nous avons une dernière méthode qui est la sparsePLS (sPLS), c'est une approche parcimonieuse qui va permettre d'intégrer une sélection de variables dans la méthodologie de la PLS. On utilise des pénalités sur la norme des coefficients pour réaliser cette sélection. La sPLS consiste à intégrer la pénalisation l_1 sur le calcul des coefficients estimés. A cet effet, chaque composante définie sera donc composée d'une combinaison linéaires des prédicteurs qui auront été sélectionnés, et dont l'effet ne sera plus masqué par les variables non influentes qui ne seront pas sélectionnées. Ces méthodes ne permettent pas implicitement d'effectuer de la sélection de variables.

III. Résultats

Dans cette partie, nous allons exposer les résultats de la comparaison des différentes méthodes utilisées pour notre étude. Pour pouvoir évaluer et comparer les méthodes, nous les avons appliqué sur les mêmes 30-découpages apprentissage/test de notre jeu de données avec et sans les facteurs de confusion afin de mettre en confrontation les différentes erreurs de prédictions. Traçons les diagrammes en moustaches des erreurs de prédiction pour l'ensemble des méthodes que nous avons appliqué.



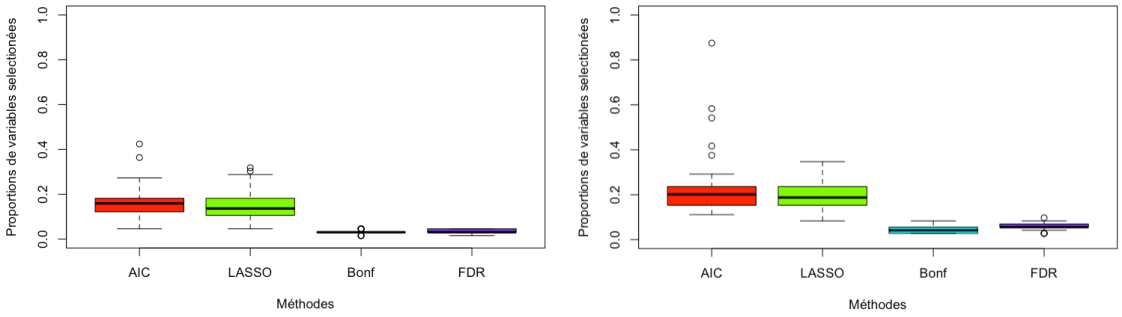
(a) DEM des erreurs de prédiction sans facteurs de confusion

(b) DEM des erreurs de prédiction avec facteurs de confusion

Figure 1 – Diagrammes en moustaches des erreurs de prédiction des méthodes en fonction de la présence des facteurs de confusion

La figure 1a est la figure représentant les diagrammes en moustaches des erreurs de prédiction de nos méthodes sans les facteurs de confusion et la figure 1b avec les facteurs de confusion. Tout d'abord, nous constatons que les méthodes ont une légère amélioration de l'erreur de prédiction avec les facteurs de confusion, mais les tendances sont les mêmes d'un graphique à l'autre. Nous interprétons donc les graphiques dans leur ensemble. Globalement, nous remarquons que la régression linéaire Ridge est la méthode qui se trompe le plus sur les données tests. D'autre part, nous remarquons que les autres méthodes ont des erreurs de prédiction à peu de chose près semblables, il semblerait que la méthode Lasso et sPLS soient les plus performantes sur les données tests. Traçons maintenant

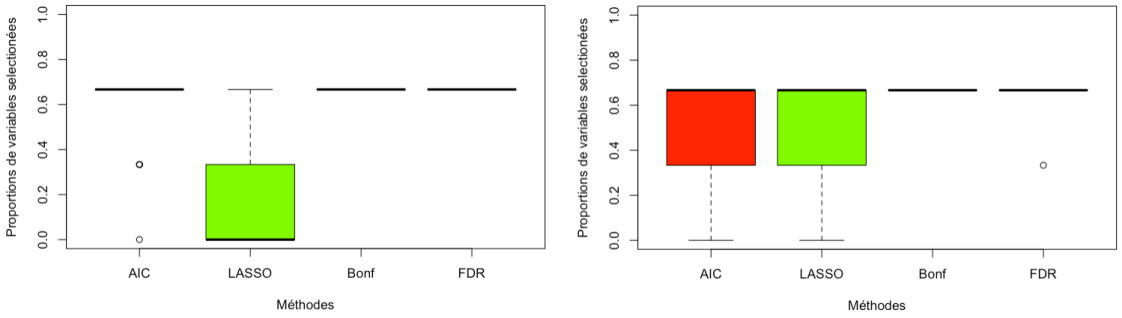
les diagrammes en moustaches des ratios des variables sélectionnées autres que Manganèse par les méthodes en fonction de la présence des facteurs de confusion. Nous rappelons que les méthodes Ridge, PCR, PLS et sPLS ne permettent pas de faire de la sélection de variables. Ainsi, nous ne traçerons pas les diagrammes pour ces méthodes car ils conduisent à des résultats triviaux.



(a) DEM sans facteurs de confusion (b) DEM avec facteurs de confusion

FIGURE 2 – Ratio de variables sélectionnées autres que Mg en fonction de la présence des facteurs de confusion

D'après la figure 2, il semblerait que l'absence des facteurs de confusion permet de réduire la variabilité dans la sélection de variables de nos modèles. De manière générale, nous remarquons que les méthodes AIC et Lasso sélectionnent en moyenne 20% de variable autre que Manganèse. La sélection des méthodes Bonferonni et Benjamini-Hochberg offre de meilleurs résultats en ne sélectionnant presque aucune autre variable que la variable Manganèse. Traçons maintenant les diagrammes en moustaches des ratios des variables sélectionnées autres que Manganèse et de ses transformations polynomiales par les méthodes en fonction de la présence des facteurs de confusion.



(a) DEM sans facteurs de confusion (b) DEM avec facteurs de confusion

FIGURE 3 – Ratio de variables oubliées parmi Mg, Mg2, Mg3 en fonction de la présence des facteurs de confusion

D'après la figure 3, nous remarquons que toutes les proportions de variables oubliées parmi Mg, Mg2 et Mg3 pour chacune de nos méthodes ne dépassent jamais deux tiers. Ce résultat est cohérent, car chacune de nos méthodes n'oublient jamais de sélectionner Mg. Nous remarquons que les méthodes Bonferonni et Benjamini-Hochberg conservent presque tout le temps qu'une seule variable parmis Mg, Mg2 et Mg3. En regardant plus précisemment, nous pouvons remarquer que la variable que ces méthodes n'oublient jamais est la méthode Mg. Contrairement aux méthodes Bonferonni et Benjamini-Hochberg, il arrive que les méthodes AIC et Lasso sélectionnent plus que une seule variable, et en regardant plus précissement, sélectionnent les variables Mg2 et Mg3. De plus, nous remarquons qu'il semblerait que les facteurs de confusions conduisent la méthode Lasso à être plus parcimonieuse quant à l'oublie des variables Mg2 et Mg3.

IV. Conclusion

Nous avons vu dans ce rapport un court rappel des différentes méthodes utilisées dans un cadre de grande dimension. Dans un objectif de minimisation de l'erreur de prédiction, il semblerait que la présence des facteurs de confusion permette d'obtenir une légère diminution de l'erreur de prédiction. De plus, nous remarquons que les meilleures méthodes pour minimiser l'erreur de prédiction sont la méthode Lasso et la méthode sPLS. La méthode Ridge conduit quant à elle à un fort pourcentage d'erreur de prédiction. Dans un objectif de sélection de variables, il semblerait que les méthodes basées sur les tests d'hypothèses et donc les méthodes de Bonferroni et de Benjamini-Hochberg soient les plus cohérentes au vu de comment ont été générées les données. Néanmoins, il semblerait que la méthode Lasso soit la plus parcimonieuse dans sa sélection de variables. Et donc finalement, dans un objectif de minimisation de l'erreur de prédiction et de sélection de variables les plus pertinentes, la méthode Lasso semble être la méthode la plus performante et la plus parcimonieuse de toutes les méthodes testées. Au vu de nos résultats, il semblerait que la méthodologie appliquée dans l'article (Bai et al. 2020) est cohérente. Il est donc difficile au vu de nos résultats de proposer une piste d'amélioration de l'étude menée dans cet article.

Bibliographie

[Bai+20] Yansen Bai et al. “Co-exposure to multiple metals, TERT-CLPTM1L variants, and their joint influence on leukocyte telomere length”. In : *Environment International* 140 (2020), p. 105762. ISSN : 0160-4120. DOI : https://doi.org/10.1016/j.envint.2020.105762. URL : http://www.sciencedirect.com/science/article/pii/S0160412019347531.

[De +18] Tim De Meyer et al. “Telomere length as cardiovascular aging biomarker : JACC review topic of the week”. In : *Journal of the American College of Cardiology* 72.7 (2018), p. 805-813.

[EHM14] Erez Eitan, Emmette R Hutchison et Mark P Mattson. “Telomere shortening in neurological disorders : an abundance of unanswered questions”. In : *Trends in neurosciences* 37.5 (2014), p. 256-263.

[Hay+17] Philip C Haycock et al. “Association between telomere length and risk of cancer and non-neoplastic diseases : a Mendelian randomization study”. In : *JAMA oncology* 3.5 (2017), p. 636-651.