Projet 2 Apprentissage Statistique : Régression

Léo Pichery, Dejoie Adèle

Mars 2025

Table des matières

Introduction

Nous nous intéressons ici à un jeu de données portant sur 41 auteurs de langue française. Celui-ci recense les occurrences de 15 518 mots dans une partie de l'œuvre de ces auteurs, répartis en quatre catégories grammaticales : noms, adverbes, adjectifs et verbes. L'objectif est de déterminer si ces données permettent de prédire la date de naissance des auteurs.

Une autre question essentielle concerne la pertinence de ces 15 518 mots : peut-on réduire le nombre de variables sans perdre d'informations significatives ? Cette réduction est particulièrement importante pour optimiser les temps de calcul et ne conserver que les variables réellement utiles.

Dans un premier temps, nous appliquerons la méthode LASSO afin de sélectionner les mots les plus pertinents directement à partir des données brutes. Nous évaluerons ensuite si ces variables permettent effectivement de prédire la date de naissance des auteurs. Ce processus sera répété après normalisation des données, en tenant compte non plus des occurrences brutes des mots, mais de leur fréquence d'utilisation par auteur.

1 Compréhension des données

Cette étude vise à analyser les données disponibles afin d'orienter notre travail. L'objectif principal est d'examiner si l'on peut retrouver la date de naissance de chaque auteur à partir du nombre d'occurrences des mots dans leurs œuvres. En particulier, nous cherchons à approcher au mieux la courbe présentée en figure 1, où les auteurs ont été classés par ordre croissant selon leur date de naissance.

Une première question consiste à déterminer si la fréquence d'utilisation des mots, en

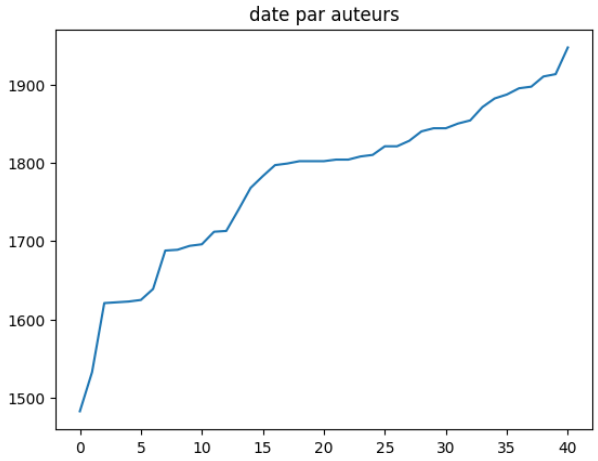

FIGURE 1 – Graphique des dates de naissances des auteurs

fonction de leur catégorie grammaticale, suffit à prédire efficacement la date de naissance des auteurs. Pour explorer cette hypothèse, nous nous appuyons sur les figures 2, 3, 4 et 5, qui ont été construites à partir des données normalisées, en considérant les fréquences plutôt que les occurrences brutes.

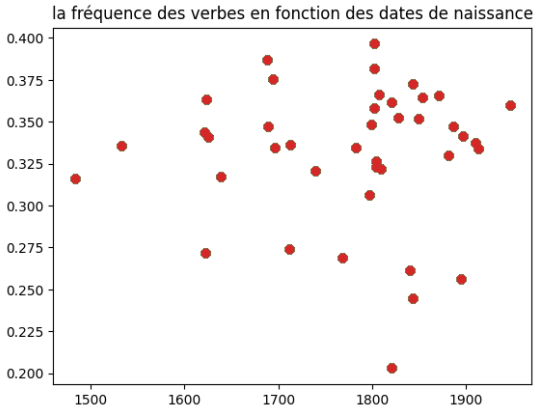

FIGURE 2 – Fréquence des verbes en fonction des dates de naissance

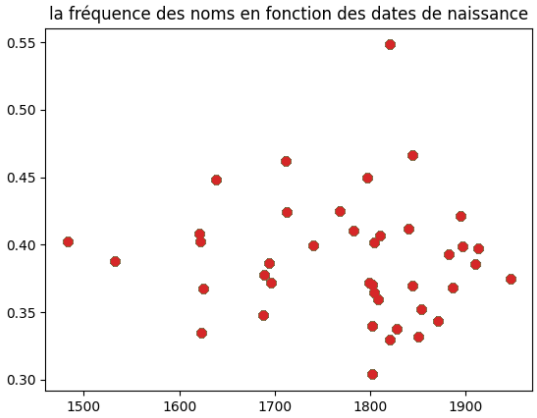

FIGURE 3 – Fréquence des noms en fonction des dates de naissance

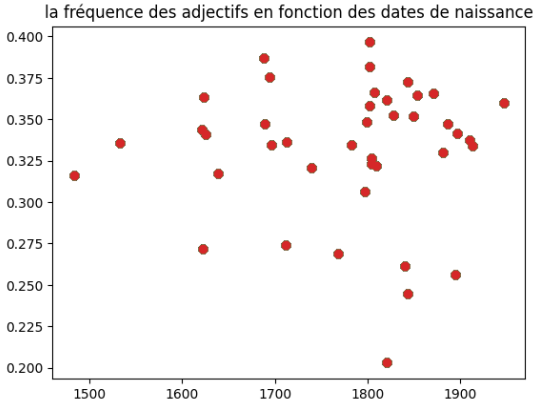

FIGURE 4 – Fréquence des adjectifs en fonction des dates de naissance

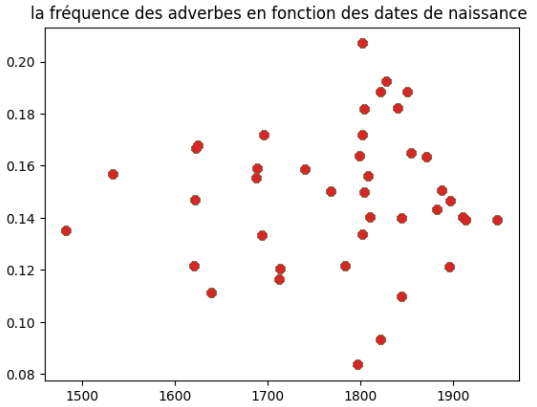

FIGURE 5 – Fréquence des adverbes en fonction des dates de naissance

Cependant, comme le montrent ces figures, il ne semble pas exister de lien linéaire évident entre la fréquence d'utilisation d'une classe grammaticale et la date de naissance des auteurs. Une régression linéaire menée sur ces données confirme cette observation :

- Coefficient de détermination R2 = 0.18
- Score en validation croisée = 12194

Ces résultats indiquent clairement une absence de corrélation linéaire entre ces variables et la date de naissance des auteurs. Même en appliquant une transformation logarithmique, l'ajustement reste insuffisant, avec un R2 de 0.17, ce qui confirme la faiblesse du modèle.

Par conséquent, nous n'approfondirons pas l'analyse des classes grammaticales et nous concentrerons plutôt sur l'étude des mots eux-mêmes. La question devient alors : tous les mots sont-ils nécessaires pour déterminer la date de naissance des auteurs ?

Pour répondre à cette interrogation, nous examinons les figures 6, 7 et 8, qui illustrent la distribution de trois mots spécifiques :

• Mot 10 : "même", utilisé par l'ensemble des auteurs, ce qui en fait un bon candidat.

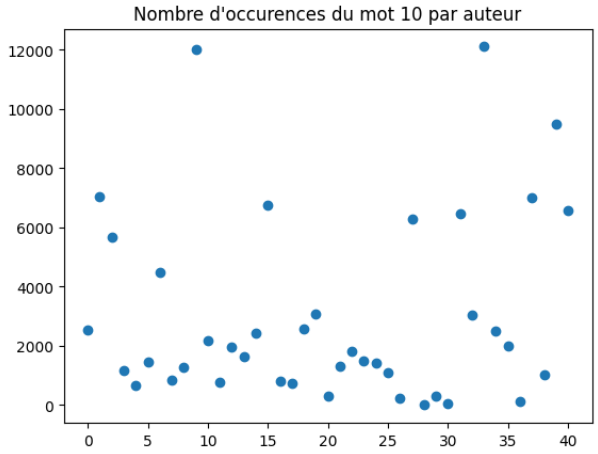

FIGURE 6 – Nombre d'occurence du mot "même"

• Mot 4 237 : "infinité", employé par certains auteurs et potentiellement discriminant.

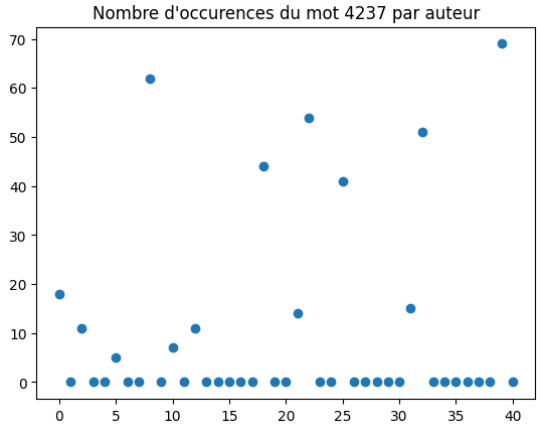

FIGURE 7 – Nombre d'occurence du mot "infinité"

• Mot 14 666 : "fougeraie", utilisé par un seul auteur, ce qui le rend peu pertinent pour la prédiction.

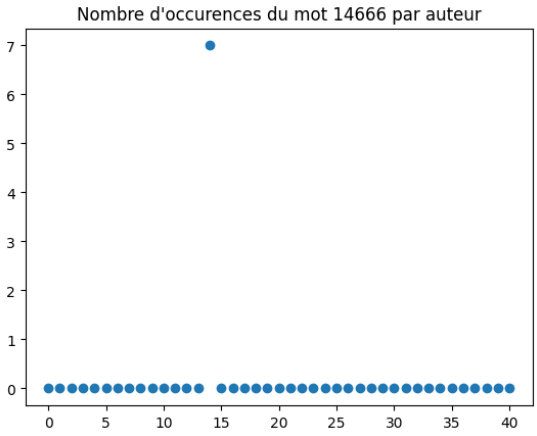

FIGURE 8 – Nombre d'occurence du mot "fougeraie"

Ainsi, un premier travail consiste à éliminer les variables inutiles, c'est-à-dire les mots dont la distribution n'apporte pas d'information significative pour la prédiction de la date de naissance des auteurs.

2 Travail sur les données brutes

Nous travaillons ici avec les données brutes, sans normalisation. Observons les résultats obtenus en effectuant une régression linéaire sur ces données :

- R2 = 1
- RMSE = $1.7 * 10^{-2}$
- Score en validation croisée : 67 597

Ces résultats révèlent un problème majeur de sur-apprentissage. En effet, avec l'ensemble des données disponibles, le modèle parvient parfaitement à prédire la date de naissance des 41 auteurs de l'échantillon. Cependant, le score en validation croisée est excessivement élevé, ce qui indique que le modèle est incapable de généraliser : sa précision chute drastiquement lorsqu'il est appliqué à des données hors échantillon.

Autrement dit, bien que le modèle s'ajuste parfaitement aux auteurs connus, il présente une erreur de prédiction non négligeable dès qu'il s'agit d'estimer la date de naissance d'un nouvel auteur. Il est donc impératif de réduire le nombre de variables afin d'améliorer la robustesse du modèle et d'éviter l'overfitting.

2.1 LASSO

Comme mentionné précédemment, il est essentiel d'éliminer un maximum de variables inutiles. En plus de faciliter la manipulation des données, cette réduction permet de diminuer le terme de variance du risque empirique, qui est proportionnel à $\sigma^2 * \frac{d}{n}$ avec d = 15 000 et n = 41. Il est donc crucial de réduire d.

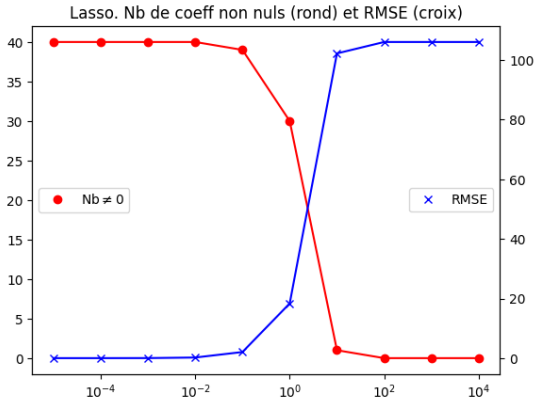

FIGURE 9 – RMSE et nombre de coefficients non nuls en fonction de α

Pour ce faire, nous appliquons une régression LASSO, qui force certaines variables inutiles à s'annuler. Un élément clé de cette approche est le choix du paramètre α, qui contrôle la régularisation. La figure 9 permet d'identifier une plage de valeurs pertinentes pour α allant de 10^{-2}, où le nombre de variables non nulles commence à décroître et où le RMSE augmente, jusqu'à 10^2, où ces deux courbes se stabilisent.

Afin de sélectionner la valeur optimale de α, nous effectuons une validation croisée : pour différentes valeurs de α dans notre plage, nous évaluons le score en validation croisée d'un modèle LASSO entraîné sur nos données. Le paramètre optimal est celui qui minimise ce score, ce qui nous conduit à choisir $\alpha = 2.51$.

L'exécution de cette étape est particulièrement longue, car chaque itération de LASSO est réalisée sur 15 518 variables, ce qui souligne l'importance de réduire ce nombre pour améliorer l'efficacité du calcul.

En appliquant LASSO avec le paramètre α sélectionné, nous constatons que la grande majorité des variables sont éliminées. À l'aide de la commande active, nous obtenons la liste des mots restants après la régularisation :

accommoder, aine, appas, commodité, déloger, dérèglement, die, finalement, incommoder, infinité, jusques, motocyclette, omnium, oiseau, ouïr, parfois, phoenix, pollen, rythme.

Certains de ces mots méritent d'être commentés. Par exemple, "phoenix" est utilisé par la plupart des auteurs (cf. figure 10), ce qui pourrait limiter son pouvoir discriminant. En revanche, un mot comme "motocyclette" peut sembler, à première vue, peu pertinent. Toutefois, en observant sa répartition parmi nos 41 auteurs (cf. figure 11), on remarque qu'il n'est utilisé que par trois auteurs, probablement nés après l'invention de la motocyclette (vers 1870). Ce mot permet donc potentiellement de discriminer les auteurs en fonction de leur époque de naissance, bien qu'il ne concerne qu'un nombre restreint d'entre eux, ce qui peut limiter son efficacité.

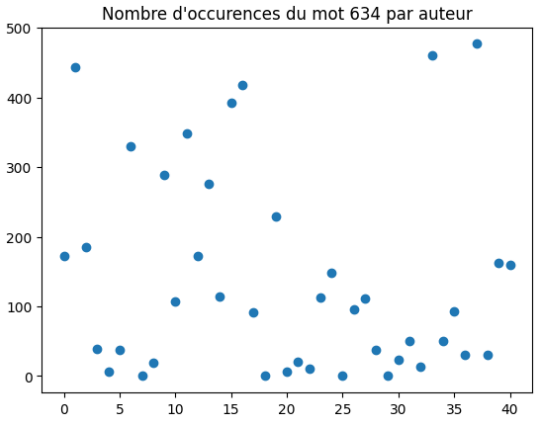

FIGURE 10 – Nombre d'occurences du mot "Phoenix"

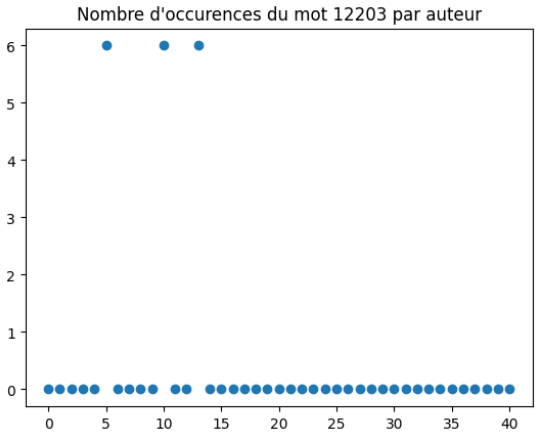

FIGURE 11 – Nombre d'occurences du mot "motocyclette"

Ainsi, notre prochaine étape consiste à affiner davantage la sélection des variables en supprimant celles qui ne sont pas optimales. Cette réduction nous permettra de diminuer le temps de calcul et d'améliorer le score en validation croisée.

2.2 Régression linéaire et RIDGE

Nous pouvons maintenant refaire une régression linéaire et analyser les résultats obtenus :

- R2 = 0.98
- RMSE = 220
- Score en validation croisée = 2216

Cette approche a considérablement réduit le score de validation croisée, tout en conservant un bon ajustement du modèle avec un coefficient de détermination élevé. Toutefois, l'augmentation du RMSE indique une légère perte de précision dans la prédiction des auteurs du jeu d'entraînement. En contrepartie, cela améliore la prédiction des auteurs hors échantillon, ce qui témoigne d'une réduction progressive du sur-apprentissage.

Cependant, le score en validation croisée reste encore trop élevé. Nous décidons donc d'éliminer manuellement les variables qui contribuent le moins à l'amélioration de ce score. En supprimant successivement les trois mots les moins pertinents selon ce critère (déloger, dérèglement et accommoder), nous obtenons :

- R2 = 0.98
- RMSE = 239
- Score en validation croisée = 838

Ainsi, la régression reste bien ajustée et l'erreur moyenne dans la prédiction de la date de naissance d'un nouvel auteur est d'environ 30 ans, ce qui reste acceptable. Il serait possible d'améliorer encore les résultats en affinant davantage la sélection des mots.

En nous basant sur les mêmes données (avec un nombre réduit de mots), nous testons maintenant une régression RIDGE, bien que nos données ne soient pas normalisées. Nous devons d'abord choisir le paramètre β, ce que nous faisons en validation croisée en explorant une plage de valeurs entre 10^{-1} et 10^{5} (cf. figure 12). Nous sélectionnons le β optimal, qui minimise le score en validation croisée, et trouvons $\beta = 1.58$

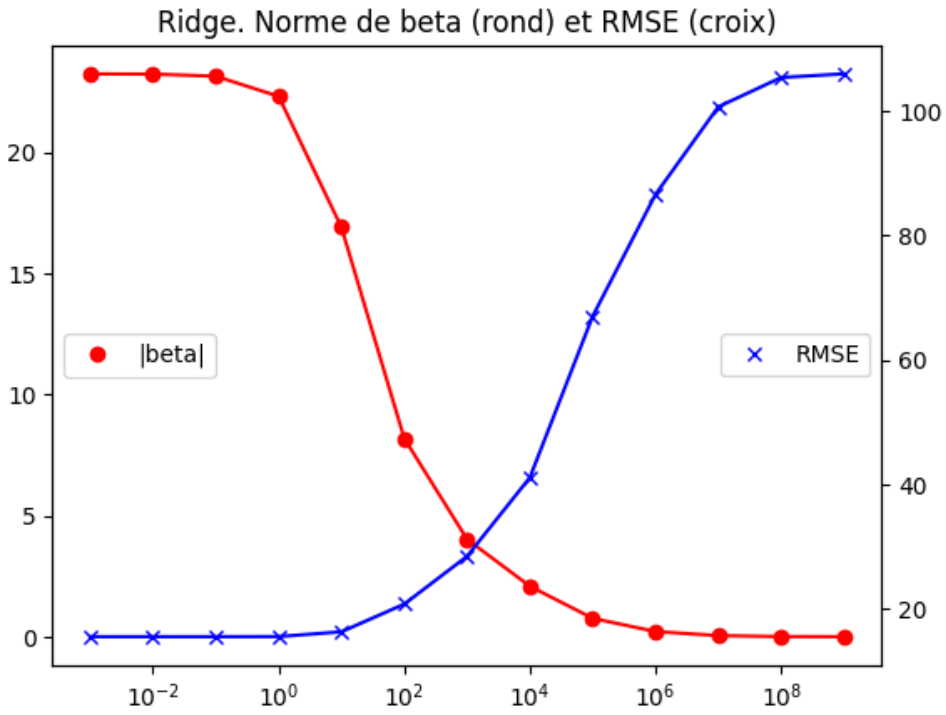

FIGURE 12 – Courbes déterminant le choix de β

Les résultats sont alors :

- R2 = 0.98
- RMSE = 240

• Score en validation croisée = 812

Légèrement meilleure que la précédente, cette approche confirme l'intérêt de réduire le nombre de variables et d'utiliser une régularisation adaptée. On observe sur la figure 13 que la prédiction suit bien la tendance attendue.

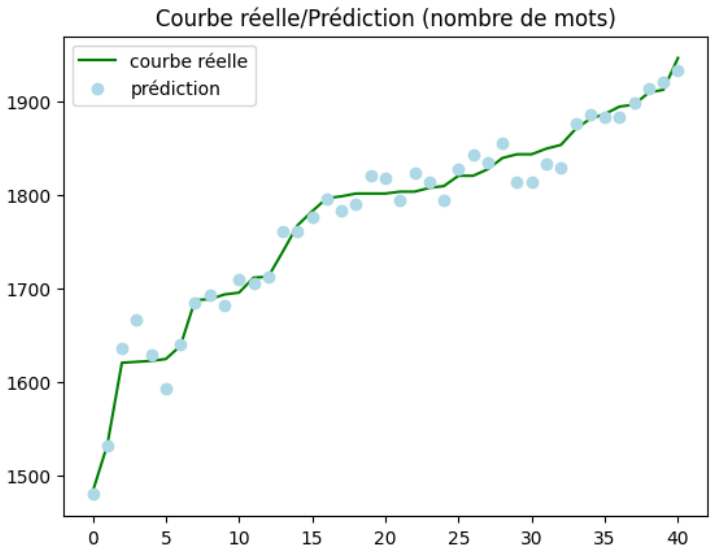

FIGURE 13 – Valeur prédites / Valeurs réelles

Le travail réalisé sur les données brutes permet donc de prédire de manière raisonnable l'année de naissance d'un auteur à partir d'un sous-ensemble de variables sélectionnées. Le score final en validation croisée de 812 correspond toujours à une erreur d'environ 30 ans en moyenne, ce qui, bien que nettement amélioré, reste significatif.

Nous allons donc explorer une nouvelle approche afin de réduire davantage cette erreur.

3 Travail sur les données en fréquence

Nous allons maintenant appliquer la même méthodologie que précédemment, mais cette fois en travaillant sur les fréquences d'apparition des mots pour chaque auteur. L'objectif est d'améliorer les résultats en validation croisée, en espérant que cette représentation des données soit plus pertinente pour la discrimination entre auteurs.

Travailler sur les fréquences présente plusieurs avantages :

• Meilleure normalisation des données, ce qui est essentiel pour des modèles comme la régression RIDGE.

• Comparaison plus cohérente entre auteurs, en évitant les biais liés à la longueur des textes.

Nous commençons par effectuer une régression linéaire sur l'ensemble des variables en fréquence. Les résultats obtenus sont :

- R2 = 1
- RMSE = $9.6 * 10^{-25}$
- Score en validation croisée = 1303

Ces résultats sont meilleurs que ceux de la toute première régression linéaire effectuée sur les données brutes, confirmant l'intuition que travailler avec des fréquences permet une meilleure modélisation. Cependant, la valeur très élevée du score en validation croisée indique que le modèle souffre encore de sur-apprentissage.

Nous avons donc la confirmation que cette approche est prometteuse, mais qu'il est toujours nécessaire de réduire le nombre de variables pour améliorer la généralisation du modèle.

3.1 LASSO

Comme pour les données brutes, la première étape consiste à appliquer une régression LASSO afin de réduire le nombre de variables. Pour cela, il est nécessaire de choisir un paramètre α.

Nous commençons donc par tracer les courbes du RMSE et du nombre de coefficients non nuls en fonction de α, ce qui donne la figure 14. Afin de trouver un α optimal, nous explorons une plage de valeurs comprises entre 10^{-2} et 10^{2}. Nous effectuons ensuite plusieurs régressions LASSO avec différentes valeurs de α, évaluons le score de validation croisée, et sélectionnons la valeur qui minimise ce score. On trouve ainsi $\alpha = 2.51$.

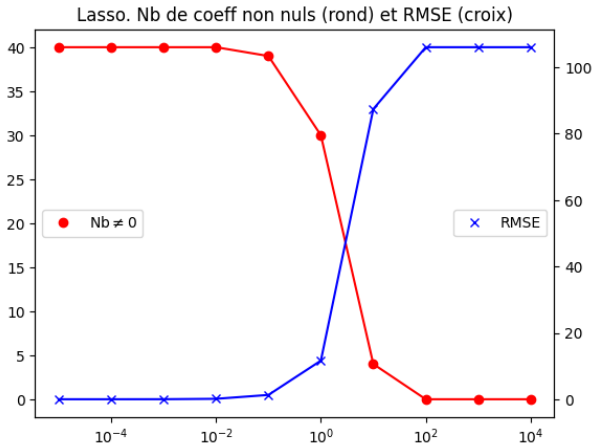

FIGURE 14 – RMSE et nombre de coefficients non nuls en fonction de α

Cette étape est une fois encore coûteuse en temps de calcul, mais elle permet de réduire considérablement le nombre de variables. Après application du LASSO avec α optimal, nous obtenons 21 mots conservés :

appel, arbitre, cadre, carte, chance, entretenir, évolution, hardiment, lèvre, mur, ouïr, parfois, pourvoir, quai, retrouver, rideau, secourir, servir, silencieux, sourire, tant.

Certains mots comme "parfois" et "ouïr" étaient déjà présents dans la sélection réalisée sur les données brutes, tandis que d'autres, comme "sourire", sont nouveaux. En observant la fréquence d'apparition du mot "sourire" pour chaque auteur (figure 15), on remarque qu'il est utilisé par presque tous, mais avec des différences significatives de fréquence. Toutefois, aucune tendance claire en fonction de la date de naissance ne semble apparaître.

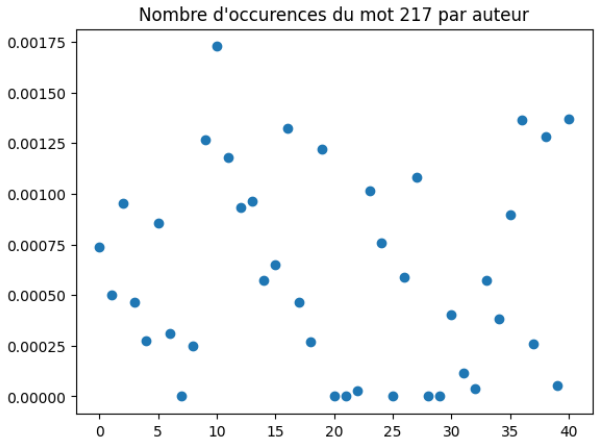

FIGURE 15 – Nombre d'occurences du mot "sourire"

Nous allons maintenant réaliser une régression linéaire en utilisant uniquement les variables sélectionnées après cette régression LASSO, dans l'espoir d'obtenir une meilleure performance en validation croisée qu'avec les données brutes.

3.2 Régression linéaire et RIDGE

Après avoir effectué la régression linéaire sur les mots sélectionnés par le LASSO, nous obtenons les résultats suivants :

- R2 = 0.99
- RMSE = 94
- Score en validation croisée : 464

Le coefficient de détermination R2 reste très élevé, ce qui indique que le modèle explique toujours bien la variance des données. Cependant, le RMSE a légèrement augmenté par rapport aux données brutes. Cette augmentation est inévitable : en réduisant le nombre de variables, nous perdons en précision sur le prédicteur entraîné sur les auteurs, mais améliorons la généralisation pour la prédiction des dates de naissance, limitant ainsi le sur-apprentissage.

Le score en validation croisée a quant à lui diminué, confirmant ainsi que la réduction du sur-apprentissage améliore la robustesse du modèle. Il est nettement meilleur que celui obtenu sur les données brutes, renforçant l'idée que les fréquences des mots sont les variables les plus pertinentes.

Afin d'affiner encore notre sélection, nous allons cette fois procéder à une réduction manuelle des variables. Pour chaque variable restante, nous la retirons individuellement et évaluons l'impact sur le score de validation croisée. Si le score s'améliore ou reste stable après le retrait d'une variable, celle-ci peut être éliminée.

Nous répétons ce processus trois fois, ce qui nous conduit à supprimer les mots "mur", "rideau" et "carte". Après cette nouvelle régression linéaire sur l'ensemble réduit, nous obtenons :

- R2 = 0.99
- RMSE = 94
- Score en validation croisée : 263

Puisque R2 et le RMSE restent inchangés, nous pouvons conclure que ces trois variables n'étaient pas essentielles et pouvaient être supprimées sans perte de performance. De plus, le score en validation croisée est presque divisé par deux, ce qui constitue un très bon résultat. L'erreur moyenne d'estimation est désormais d'environ 15 ans, soit une amélioration significative par rapport aux données brutes.

Nous allons maintenant tester une régression RIDGE sur cet ensemble de variables optimisé. Étant donné que nous travaillons sur des fréquences normalisées, nos données sont idéalement adaptées à cette méthode.

Comme pour le LASSO, nous devons sélectionner un paramètre β optimal. Pour cela, nous traçons la figure 16 afin d'identifier une plage de valeurs pertinente. Nous choisissons d'explorer β entre 10^{-10} et 10^{-2}. Ensuite, nous raffinons notre choix par validation croisée, en testant plusieurs valeurs dans cette plage. Nous sélectionnons finalement le paramètre $\beta = 5 * 10^{-9}$, qui offre le meilleur score.

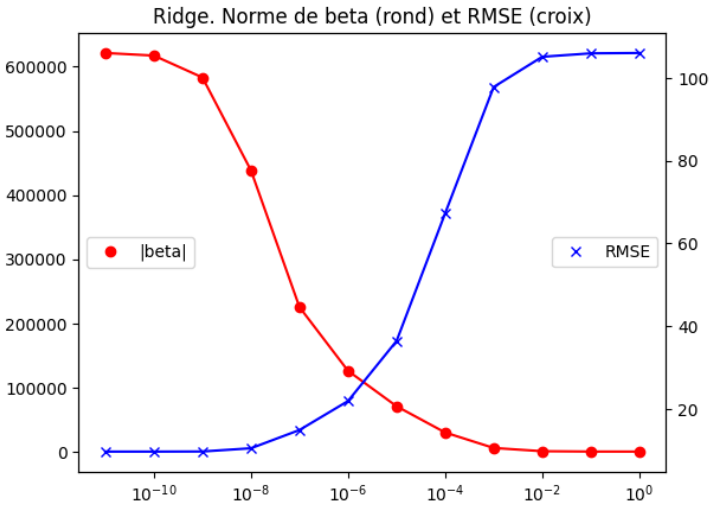

FIGURE 16 – Courbes déterminant le choix de β

Les résultats obtenus avec cette régression RIDGE sont :

- R2 = 0.99
- RMSE = 101
- Score en validation croisée : 237

Ces scores sont très proches de ceux obtenus précédemment, indiquant que la régularisation RIDGE ne dégrade pas significativement la performance du modèle.

Nous comparons maintenant les prédictions du RIDGE aux dates de naissance réelles (figure 17). Les points prédits sont très proches de la courbe réelle, indiquant que les résultats sont très satisfaisants et meilleurs que ceux obtenus avec les données brutes.

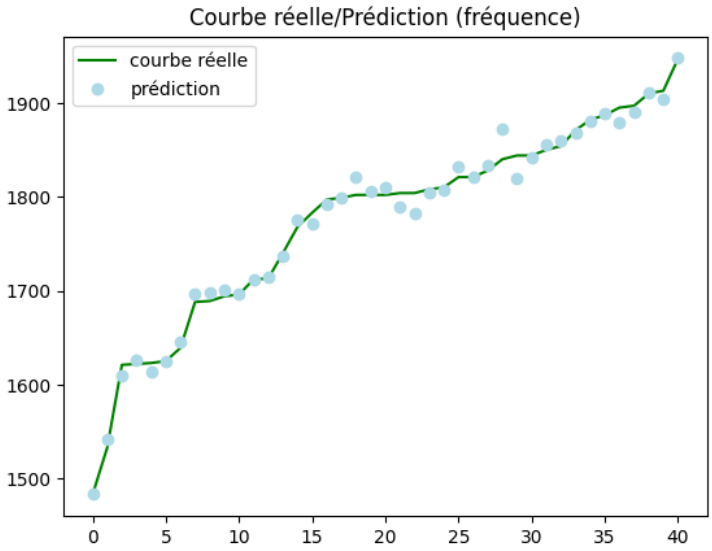

FIGURE 17 – Valeurs réelles / Valeurs prédites

Enfin, nous évaluons l'écart moyen entre l'année prédite et l'année réelle pour les 41 auteurs étudiés :

- Données brutes : écart moyen de 12 ans
- Fréquences : écart moyen de 7 ans

Ces résultats confirment que le modèle basé sur les fréquences est plus performant que celui utilisant les données brutes.

3.3 ANOVA pour éliminer une variable ?

En reprenant les données normalisées avant la suppression du dernier mot, nous testons une autre approche pour améliorer la performance du prédicteur : l'ANOVA. Cette méthode nous permet d'identifier les variables ayant le moins d'influence significative. D'après la figure 18, le mot évolution semble avoir un impact négligeable et peut être supprimé. Après son retrait, les résultats obtenus sont les suivants :

- R2 = 0.99
- RMSE = 94
- Score en validation croisée : 262

On constate une forte augmentation du score en validation croisée, ce qui indique une dégradation des performances du modèle. L'ANOVA ne s'est donc pas révélée efficace dans ce contexte. Ce résultat suggère que la relation entre la date de naissance des auteurs et les variables restantes n'est pas strictement linéaire, ce qui n'était d'ailleurs pas garanti a priori.

	sum_sq	df	F	PR(>F)
ouïr	57571.940588	1.0	7.881566	0.010553
secourir	9892.487329	1.0	1.354276	0.257584
entretenir	1806.589830	1.0	0.247321	0.624135
retrouver	3874.911066	1.0	0.530473	0.474452
quai	14654.122482	1.0	2.006141	0.171327
hardiment	2335.390042	1.0	0.319714	0.577772
sourire	7259.258157	1.0	0.993788	0.330167
chance	937.118200	1.0	0.128291	0.723786
appel	1043.304852	1.0	0.142828	0.709279
pourvoir	1419.519255	1.0	0.194331	0.663841
servir	879.465008	1.0	0.120398	0.732056
silencieux	2051.321191	1.0	0.280825	0.601720
carte	3398.369726	1.0	0.465235	0.502642
lèvre	1894.144909	1.0	0.259307	0.615906
parfois	6887.867424	1.0	0.942945	0.342574
arbitre	354.024876	1.0	0.048466	0.827881
évolution	0.020160	1.0	0.000003	0.998690
tant	25511.621584	1.0	3.492526	0.075660
cadre	10.336946	1.0	0.001415	0.970347

FIGURE 18 – Table ANOVA

4 Conclusion

Après avoir appliqué différentes méthodes sur plusieurs types de données, nous obtenons des résultats de prédiction satisfaisants. Comme l'intuition le suggérait, l'utilisation des fréquences d'apparition des mots est la meilleure approche. La réduction du nombre de variables à l'aide d'un LASSO, suivie d'une analyse par validation croisée, s'avère essentielle pour limiter le sur-apprentissage.

En effet, lorsque toutes les variables sont prises en compte, les indicateurs de performance sur l'ensemble des 41 auteurs, comme R2 et le RMSE, sont très bons, mais le score en validation croisée reste médiocre. Cela signifie que le modèle s'adapte trop aux données d'apprentissage sans généraliser efficacement à de nouveaux auteurs. La stratégie fondée sur les fréquences permet donc d'améliorer cette généralisation et d'obtenir un prédicteur plus fiable.

En revanche, la méthode ANOVA ne s'est pas montrée pertinente dans ce contexte. Comme mentionné précédemment, rien ne garantit une relation linéaire entre la date de naissance des auteurs et la fréquence des mots sélectionnés, ce qui limite l'efficacité de cette approche.