École nationale de la statistique et de l'analyse de l'information

Mastère Data Science - Connaissance Client

Projet Machine Learning

Thème

Régression Pénalisée

Réalisé par

DIAKITE GAOUSSOU

Destiné au professeur

Denys POMMERETD

2022/2023

Table des matières

Résumé

Ce travail consiste en une analyse quantitative approfondie des données issues de 37 entreprises de transport maritime, centrée sur la relation entre les émissions de CO2, le chiffre d'affaires, et les incidents conduisant à des retards dans les ports. Diverses méthodes de régression pénalisée telles que Ridge, LASSO et ElasticNet ont été employées pour construire des modèles explicatifs. En outre, des techniques avancées telles que la régression sur composantes principales et la régression PLS ont été explorées pour optimiser la sélection des variables et la compréhension des facteurs déterminants.

Chapitre 1

Introduction

Dans le domaine complexe de l'apprentissage automatique, la régression pénalisée se révèle être une méthode incontournable pour équilibrer la précision et la simplicité des modèles prédictifs. À travers l'analyse des données de 37 entreprises de transport maritime, ce projet individualisé a mis en œuvre des techniques avancées telles que Ridge, LASSO, et ElasticNet. Ces méthodes, en introduisant des pénalités sur les coefficients de régression, s'avèrent particulièrement efficaces face à des défis tels que la multicollinéarité et le risque de surajustement, permettant de minimiser l'influence de variables superflues. L'expertise requise pour choisir adéquatement les hyperparamètres et comprendre l'effet de ces pénalités est mise en avant, soulignant l'importance d'une utilisation réfléchie de la régression pénalisée pour obtenir des modèles robustes et généralisables.

Cet effort méticuleux pour sélectionner et interpréter les variables significatives a conduit à l'émergence de modèles prédictifs à la fois précis et faciles à comprendre, révélant les interactions subtiles entre les émissions de CO2, les retards de transport et les revenus des entreprises. L'exploration supplémentaire de la régression sur composantes principales et de la régression PLS a enrichi cette étude, permettant une comparaison approfondie des approches et une sélection rigoureuse des coefficients les plus pertinents. L'objectif ultime de ce travail a été de dévoiler des insights précieux à partir de données complexes, reflétant une démarche analytique rigoureuse et une interprétation claire, pour répondre aux exigences actuelles de l'analyse de données en transport maritime.

Chapitre 2

Revue de Littérature

La régression pénalisée est une branche de l'apprentissage automatique qui gagne en popularité grâce à sa capacité à améliorer la prédiction et l'interprétation des modèles statistiques. Cette section passe en revue les méthodes de régression pénalisée couramment utilisées, en explorant leur formulation mathématique, leurs domaines d'application, ainsi que leurs forces et faiblesses inhérentes.

2.1 Régression Ridge

La régression Ridge s'attaque au problème de multicollinéarité en ajoutant une pénalité proportionnelle au carré des coefficients de régression. Elle est particulièrement utile lorsque les variables explicatives sont corrélées entre elles.

Formule : $Minimiser(Y - X\beta)^2 + \lambda\|\beta\|^2$.

Utilisation : Recommandée dans des situations où la réduction du surajustement est préférable à la sélection de variables.

Avantages et Inconvénients : La méthode Ridge rétrécit les coefficients, ce qui aide à réduire l'overfitting et améliore la stabilité du modèle. Cependant, elle ne fait pas de sélection de variables, ce qui signifie que tous les prédicteurs sont conservés dans le modèle final.

2.2 LASSO

LASSO est une technique de sélection de variables qui permet de réduire certains coefficients à zéro, facilitant ainsi l'interprétation du modèle.

Formule : $Minimiser(Y - X\beta)^2 + \lambda\|\beta\|_1$.

Utilisation : Idéale pour la simplification des modèles en réduisant leur dimensionnalité.

Avantages et Inconvénients : LASSO est efficace pour la sélection de variables et la réduction de la dimensionnalité. Toutefois, cette méthode peut être instable dans certaines situations, notamment lorsque le nombre de variables est supérieur au nombre d'observations ou lorsque plusieurs variables sont fortement corrélées.

2.3 ElasticNet

ElasticNet combine les caractéristiques de la régression Ridge et LASSO pour gérer à la fois la sélection de variables et la multicollinéarité.

Formule : $Minimiser(Y - X\beta)^2 + \lambda_1\|\beta\|^2 + \lambda_2\|\beta\|_1$.

Utilisation : Particulièrement adaptée à des données où plusieurs variables sont corrélées.

Avantages et Inconvénients : ElasticNet peut capturer les avantages de la Ridge et LASSO, la rendant efficace pour des données complexes. Toutefois, le réglage de deux paramètres de pénalisation peut rendre la calibration du modèle plus exigeante.

2.4 Régression sur Composantes Principales (PCR)

PCR est une technique de réduction de dimension qui combine l'Analyse en Composantes Principales (ACP) avec la régression linéaire.

Formule : Les composantes principales $PC = XW$ sont extraites, puis utilisées pour la régression $Y \approx PC \times \beta$.

Utilisation : Pertinente lorsque l'on est confronté à un grand nombre de prédicteurs.

Avantages et Inconvénients : La PCR peut simplifier les modèles avec de nombreuses variables explicatives, mais il existe un risque de perte d'information pertinente due à la

réduction de dimension.

2.5 Régression PLS (Partial Least Squares)

PLS est une méthode de régression qui cherche à modéliser la relation entre les matrices de prédicteurs X et la réponse Y en maximisant la covariance entre eux.

Formule : Identification des vecteurs w et c pour que $t = Xw$ et $u = Yc$ maximisent la covariance entre t et u.

Utilisation : Recommandée pour les données avec des prédicteurs fortement corrélés.

Avantages et Inconvénients : PLS est efficace pour gérer la multicollinéarité et peut être utile lorsque le nombre de prédicteurs est grand par rapport au nombre d'observations. Cependant, la méthode peut être moins interprétable par rapport aux modèles basés sur des composantes principales.

Chapitre 3

Application des Méthodes de Régression Pénalisée

Dans cette section, nous appliquons les méthodes de régression pénalisée discutées précédemment sur les données des compagnies maritimes disponibles dans les fichiers « transport1.txt » et « transportmod3.txt ».

3.1 Présentation des Données

Les données dans « transport1.txt » comprennent 37 lignes et 63 colonnes, y compris trois variables à prédire : les émissions de CO2 (CO2), les incidents (INCID) et le chiffre d'affaires (CA). Le fichier « transportmod3.txt » contient également 37 lignes mais avec 64 colonnes, incluant une variable supplémentaire, le CA3. Ces ensembles de données fournissent une base solide pour l'analyse des relations entre les différentes variables et l'impact des incidents portuaires sur les émissions de CO2 et les performances financières des compagnies maritimes.

Dans les sous-sections suivantes, chaque méthode de régression pénalisée sera appliquée pour modéliser ces relations, en tenant compte des particularités de chaque jeu de données.

3.2 Modélisation du CO2

3.2.1 Analyse de la Variable CO2

Dans cette sous-section, nous débutons par une analyse détaillée de la variable CO2. Cela comprend l'exploration des caractéristiques de distribution de CO2, comme la moyenne, la médiane, et l'écart type. L'objectif est de comprendre la nature de cette variable et son influence potentielle sur les résultats de modélisation. Cette étape préliminaire est cruciale pour orienter le choix de la méthode de régression pénalisée la plus appropriée pour modéliser l'impact des différents facteurs sur les émissions de CO2 des compagnies maritimes.

Nous présentons ci-dessous un tableau récapitulatif des statistiques descriptives de la variable CO2 :

Statistique	**Valeur**
Nombre d'observations	37
Moyenne	7.46
Écart-type	1.31
Minimum	4.97
1er Quartile (25%)	6.71
Médiane (50%)	7.41
3ème Quartile (75%)	8.29
Maximum	10.55

TABLE 3.1 – Statistiques descriptives de la variable CO2

Interprétation : La moyenne des émissions de CO2 est de 7.46, avec un écart significatif entre le minimum et le maximum, indiquant une variabilité dans les émissions entre les différentes compagnies. L'écart-type de 1.31 souligne cette variabilité. La distribution semble être centrée autour de la médiane de 7.41, suggérant une répartition relativement symétrique des émissions de CO2.

Les histogrammes et boxplots sont des outils complémentaires pour visualiser la distribution des données. L'histogramme révèle la forme de la distribution et le boxplot montre la médiane, les quartiles et les valeurs extrêmes. Ensemble, ils fournissent une vue d'ensemble de la tendance centrale, de la dispersion et de la forme de la distribution des émissions de CO2, soulignant l'existence de quelques valeurs extrêmes à la haute gamme

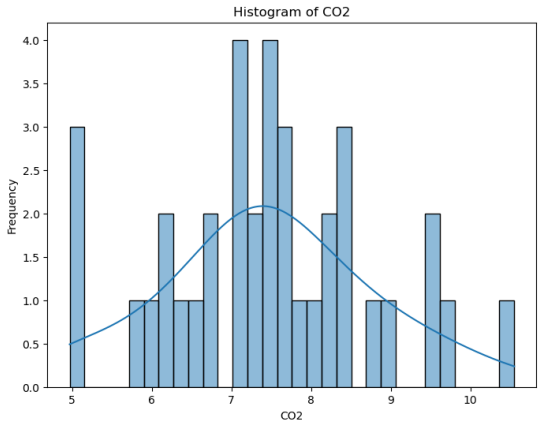

FIGURE 3.1 – Histogramme des émissions de CO2

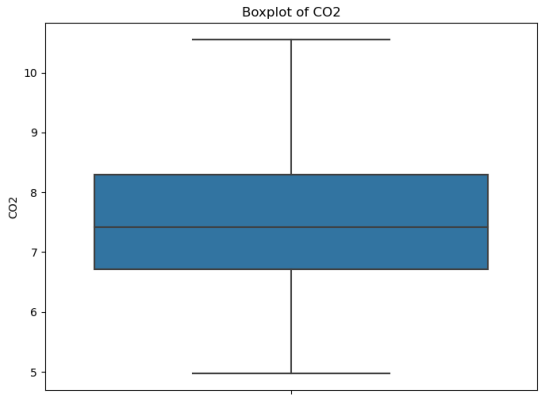

FIGURE 3.2 – Boxplot des émissions de CO2

de l'échelle. L'histogramme montre une distribution de la variable CO2 avec une légère asymétrie vers la droite, ce qui suggère une tendance des valeurs de CO2 à se regrouper vers la gauche de la moyenne, avec quelques valeurs extrêmes plus élevées. Cela complète l'analyse des statistiques descriptives en montrant visuellement la répartition des émissions de CO2 parmi les compagnies maritimes.

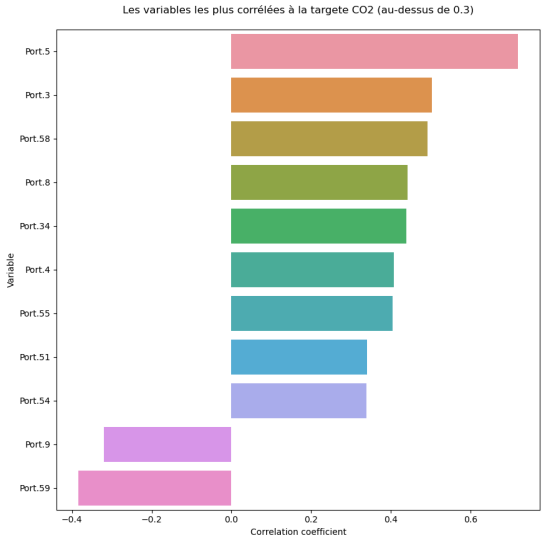

FIGURE 3.3 – Variables corrélées à la variable CO2

Le graphique Figure 3.3 ci-dessus présente les variables qui ont une corrélation significative avec les émissions de CO2, avec une corrélation seuil de 0.3. Cette visualisation est essentielle dans le processus de sélection de caractéristiques pour la modélisation prédic-

tive, car elle aide à identifier les prédicteurs potentiels qui ont le plus grand impact sur la variable cible, CO2. L'identification de ces variables est une étape préliminaire clé avant d'appliquer les méthodes de régression pénalisée.

FIGURE 3.4 – Réseau de fortes corrélations entre les variables explicatives

Le graphique Figure 3.4 illustre le réseau des corrélations fortes entre les variables explicatives, soulignant l'importance de l'interdépendance dans l'ensemble de données. Cette visualisation est cruciale pour la détection de la multicollinéarité, un phénomène

qui peut sérieusement affecter la performance des modèles de régression. En identifiant les groupes de variables fortement corrélées, on peut mieux comprendre la structure des données et prendre des décisions éclairées sur la sélection des variables ou la nécessité d'appliquer des techniques de réduction de dimension.

Le réseau de corrélations illustre la forte interdépendance entre les variables explicatives, révélant les défis posés par la multicollinéarité dans l'ensemble de données. Cette interdépendance est une raison principale de l'utilisation de la régression pénalisée. Cette méthode est avantageuse car elle permet de pénaliser et de réduire l'influence des variables corrélées, ce qui améliore la performance et la fiabilité des modèles prédictifs en atténuant l'impact de la multicollinéarité. Identifier des clusters de variables corrélées aide à affiner la sélection de caractéristiques pour des modèles plus robustes et interprétables.

3.2.2 Régression Ridge

Cette sous-section se consacre à la mise en œuvre de la régression Ridge pour modéliser l'impact des variables explicatives sur les émissions de CO2. Les variables explicatives seront d'abord normalisées pour garantir que leurs échelles variées ne faussent pas les résultats du modèle. La normalisation est une étape préparatoire qui améliore la performance du modèle en rendant les variables comparables, ce qui est particulièrement bénéfique lors du choix du paramètre de régularisation λ. La procédure de sélection de ce λ optimal se fera via une méthode de validation croisée, en ajustant la grille de recherche au besoin. Nous entamerons cette analyse en utilisant Python pour sa capacité à gérer efficacement les données et à exécuter des analyses préliminaires, puis nous transitionnerons vers R pour exploiter ses fonctionnalités statistiques avancées dans les étapes de modélisation ultérieures.

La courbe de validation montre l'effet de différents choix de λ sur l'erreur quadratique moyenne (MSE) du modèle Ridge. Le λ optimal minimise le MSE, mais ici, le choix se porte sur le λ moyen, car il permet de retenir un nombre plus important de variables explicatives, ce qui est souhaitable pour l'analyse. Le choix de ce λ est crucial et Python a été préféré pour cette tâche en raison de ses outils de visualisation avancés, qui peuvent différer des résultats obtenus par validation croisée en R.

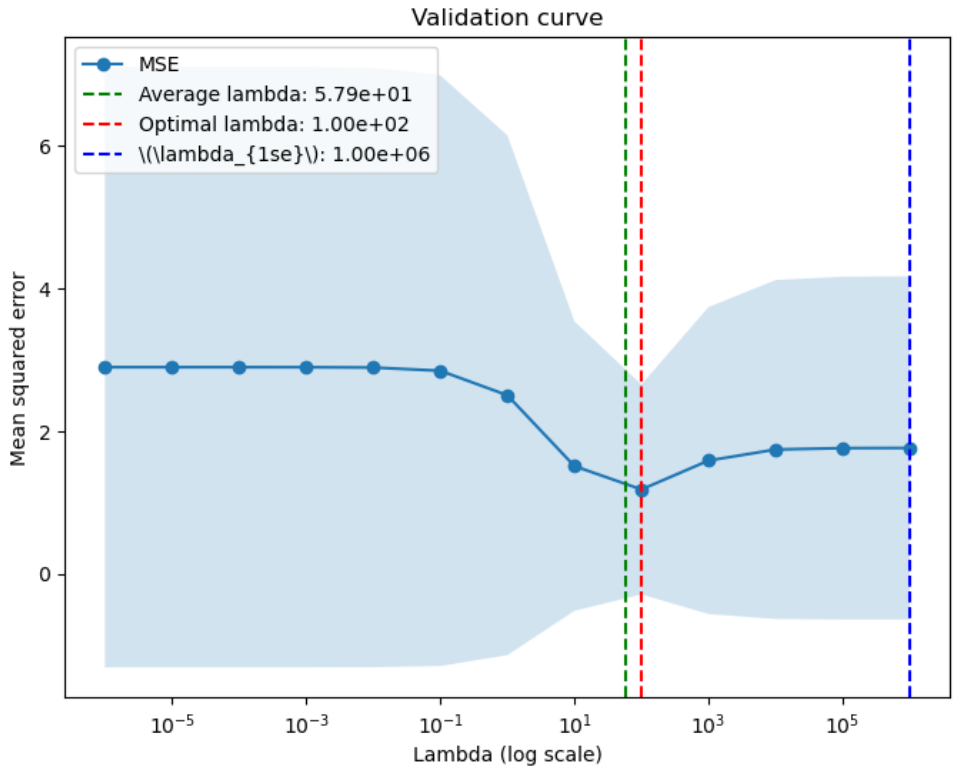

FIGURE 3.5 – Courbe de validation pour la sélection de λ dans la régression Ridge

Les deux graphiques illustrent les coefficients de la régression Ridge triés par valeur absolue. Le premier graphique indique un seuil prédéfini qui aide à identifier les variables les plus significatives. Les variables dont les coefficients dépassent ce seuil sont considérées comme ayant un impact plus fort sur la variable cible. Le deuxième graphique, un boxplot, montre la distribution des coefficients et met en évidence les valeurs extrêmes. Ensemble, ces graphiques fournissent une base pour sélectionner les variables qui contribueront le plus au modèle prédictif.

Les coefficients représentés dans le graphique et listés dans le tableau illustrent les variables finales retenues par la régression Ridge. La méthode a réduit les coefficients de la plupart des variables en-dessous du seuil de 0.1, seules les variables significatives le dépassant sont conservées. Cette sélection indique leur importance dans la prédiction des émissions de CO2, ce qui permet de construire un modèle prédictif à la fois robuste et interprétable.

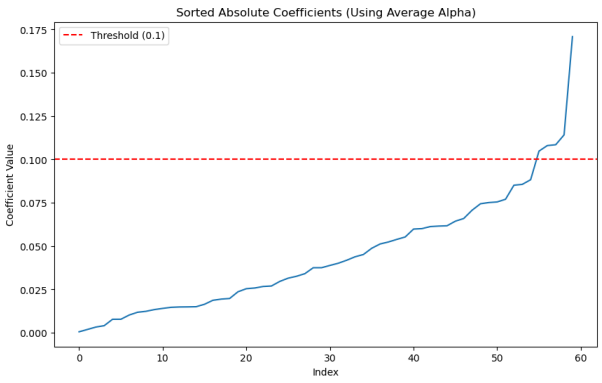

FIGURE 3.6 – Coefficients triés par valeur absolue

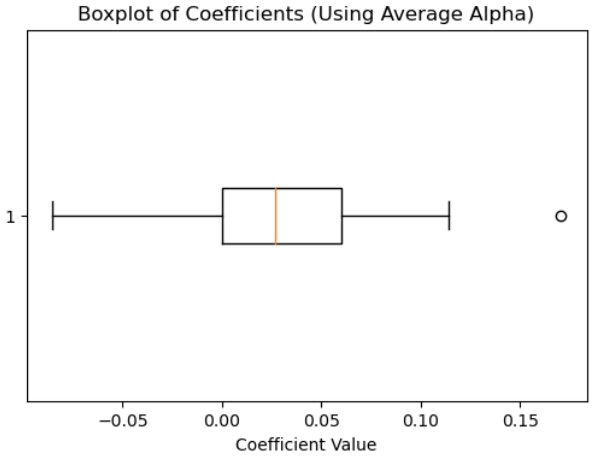

FIGURE 3.7 – Boxplot des coefficients

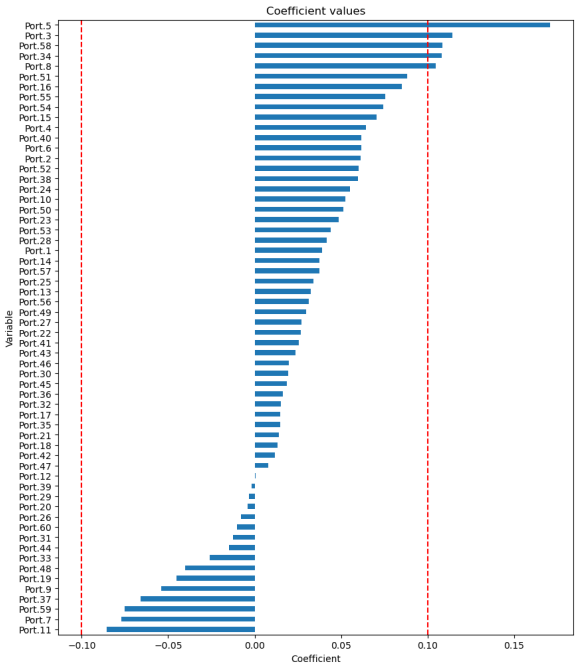

FIGURE 3.8 – Coefficients après Ridge

Variable	**Coefficient**
Port.5	0.170815
Port.58	0.108461
Port.34	0.107997
Port.3	0.114165
Port.8	0.104785

FIGURE 3.9 – Variables finales sélectionnées par Ridge (triées par valeur absolue)

La régression Ridge a démontré sa capacité à identifier et à conserver les variables les plus pertinentes pour prédire les émissions de CO2, tout en minimisant l'impact de la multicollinéarité. Le modèle final, qui inclut les variables sélectionnées avec des coefficients supérieurs au seuil établi, s'attend à offrir une balance optimale entre la précision et la simplicité, conduisant à une interprétabilité accrue tout en maintenant la robustesse nécessaire pour des prédictions fiables.

3.2.3 LASSO

La régression LASSO sera employée dans cette section pour affiner notre modèle prédictif des émissions de CO2. La méthode LASSO est choisie pour sa capacité à effectuer une sélection de variables intrinsèque, réduisant potentiellement la complexité du modèle en annulant les coefficients de certaines variables. Nous déterminerons un seuil approprié pour la sélection des variables en considérant le lambda qui minimise l'erreur de validation croisée (lambda.min) ou celui qui est le plus robuste dans le contexte de la validation croisée (lambda.1se). Le résultat final identifiera les ports ayant les influences les plus significatives sur les émissions de CO2, fournissant ainsi un modèle épuré et efficace.

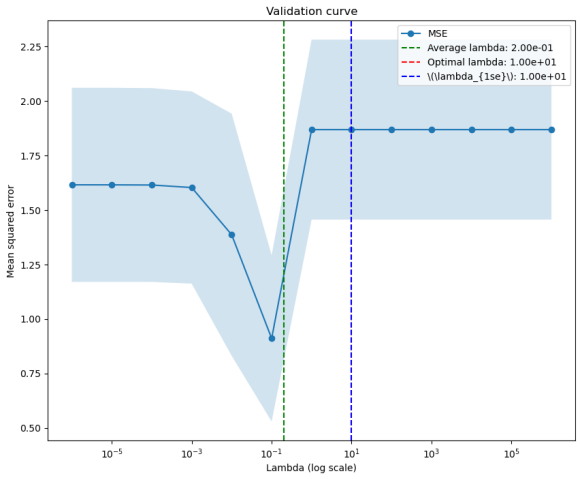

FIGURE 3.10 – Courbe de validation pour la sélection de λ avec LASSO

La courbe de validation illustre la recherche du paramètre de régularisation λ pour la régression LASSO. Le choix de $\lambda.1se$ comme seuil pour la sélection des variables est stratégique car il fournit un modèle plus généralisable en privilégiant des coefficients légèrement plus grands tout en conservant une erreur de validation croisée compétitive. Ce

choix vise à équilibrer la complexité du modèle et la précision des prédictions.

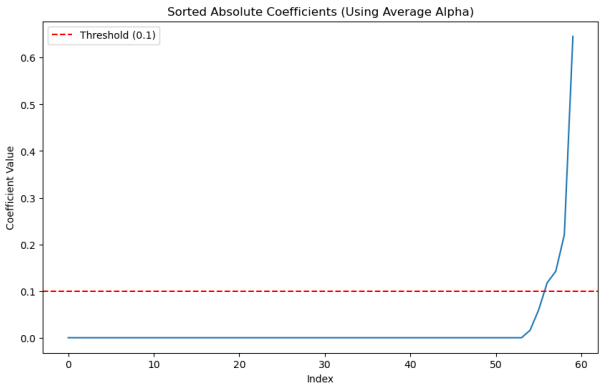

FIGURE 3.11 – Coefficients triés du LASSO

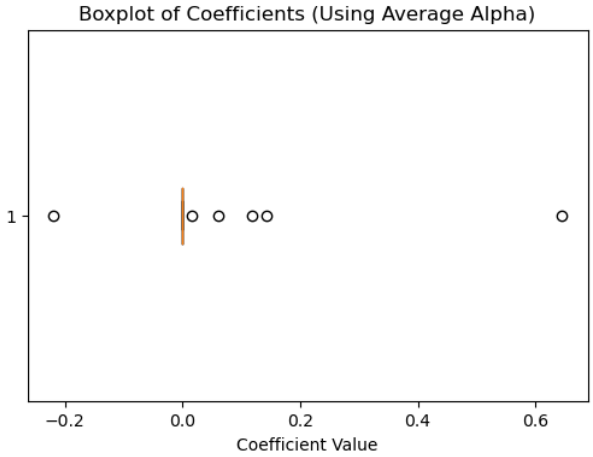

FIGURE 3.12 – Boxplot des coefficients du LASSO

Ces graphiques illustrent les coefficients résultant du modèle LASSO. Le premier montre les coefficients triés par valeur absolue avec un seuil de 0.1, indiquant quels coefficients sont considérés suffisamment importants pour être inclus dans le modèle. Le second graphique, le boxplot, visualise la dispersion des coefficients et met en exergue les valeurs atypiques. Ces deux visualisations sont essentielles pour comprendre quels prédicteurs le modèle LASSO a sélectionnés comme ayant un impact significatif sur les émissions de CO2.

Le graphique à gauche démontre les coefficients triés issus du modèle LASSO, soulignant la capacité de LASSO à réduire les coefficients de certaines variables à zéro. Les variables retenues, celles avec des coefficients non nuls, sont celles qui influencent réellement la variable réponse. Le tableau à droite liste ces variables significatives, indiquant leur importance dans le modèle. Ensemble, ces éléments reflètent l'efficacité de LASSO pour la sélection de variables dans un contexte de grande dimensionnalité.

La régression LASSO a affiné le modèle en éliminant les variables moins pertinentes, ce qui a permis de mettre en évidence les facteurs clés influençant les émissions de CO2. Par rapport à la régression Ridge, qui tend à conserver davantage de variables mais avec des coefficients réduits, LASSO offre un modèle plus épuré en sélectionnant un ensemble plus restreint de prédicteurs. Les variables communes aux deux modèles sont d'un intérêt

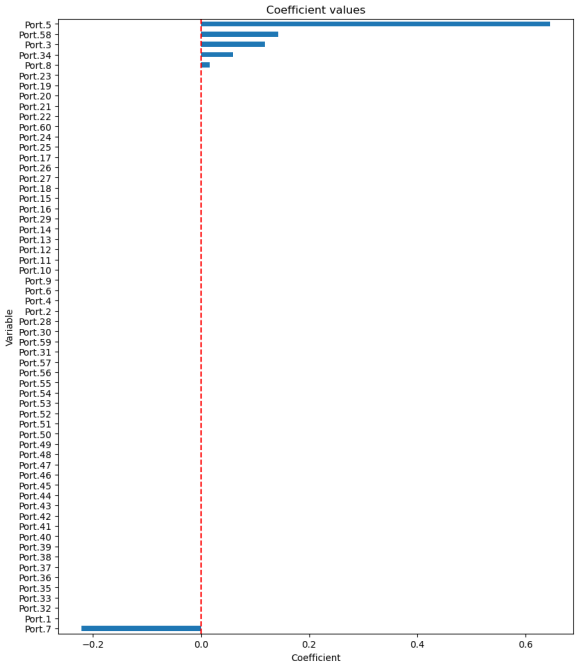

FIGURE 3.13 – Coefficients LASSO

Variable	Coefficient
Port.5	0.645565
Port.7	-0.220656
Port.58	0.142283
Port.3	0.117940
Port.34	0.059782
Port.8	0.016093

FIGURE 3.14 – Variables finales sélectionnées par LASSO (triées par valeur absolue)

particulier car elles sont robustes aux méthodes de régularisation, renforçant ainsi leur pertinence présumée.

3.2.4 ElasticNet

Dans cette sous-section, nous appliquerons la régression ElasticNet, qui intègre les propriétés de Ridge et de LASSO pour un ajustement de modèle optimal. ElasticNet est particulièrement utile quand les prédicteurs sont nombreux et corrélés. Nous chercherons à déterminer la meilleure combinaison des paramètres de régularisation, α et λ, pour équilibrer entre la capacité de sélection de variables de LASSO et la contraction des coefficients de Ridge. Cette approche vise à produire un modèle robuste, précis et bien adapté à la complexité des données à notre disposition.

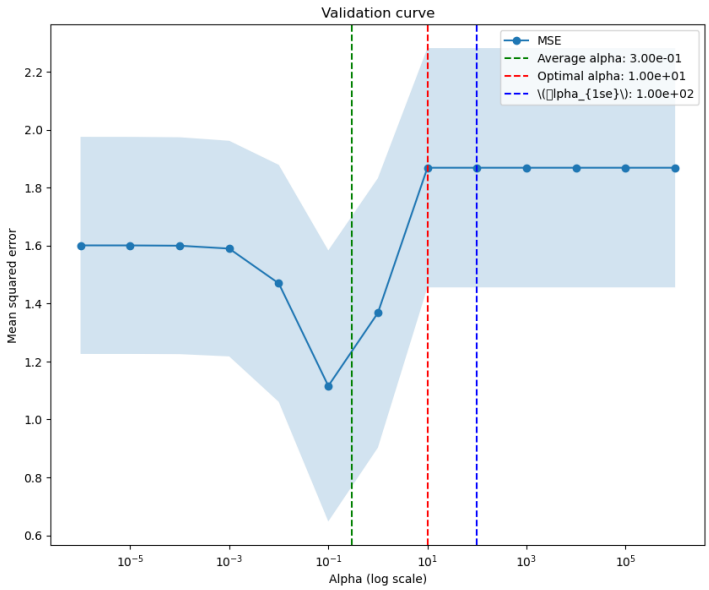

FIGURE 3.15 – Courbe de validation pour le choix d'alpha dans ElasticNet

La courbe de validation illustre l'effet de différents niveaux d'alpha sur l'erreur quadratique moyenne (MSE) pour la régression ElasticNet. Un alpha moyen est choisi pour équilibrer la sélection des variables et la réduction de la multicollinéarité. Cet alpha est préférable pour un modèle équilibré qui n'est ni trop biaisé ni trop variable.

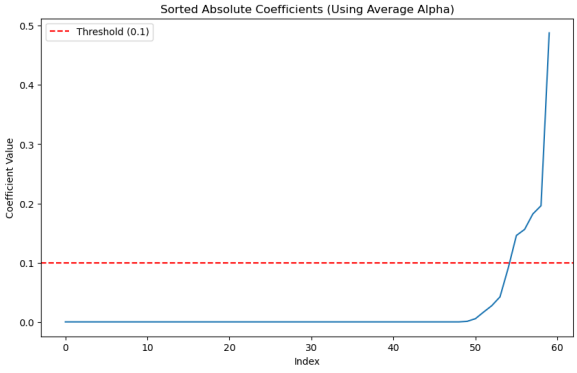

FIGURE 3.16 – Coefficients triés ElasticNet

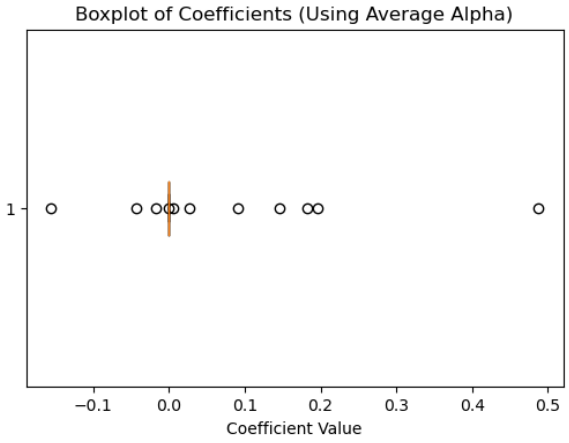

FIGURE 3.17 – Boxplot des coefficients ElasticNet

Le premier graphique montre les coefficients absolus triés de la régression ElasticNet, avec un seuil qui identifie les variables les plus influentes. Les variables au-dessus de ce seuil sont jugées importantes pour le modèle. Le boxplot révèle la distribution des coef-

ficients et met en évidence les valeurs extrêmes, fournissant une autre perspective sur la pertinence des variables.

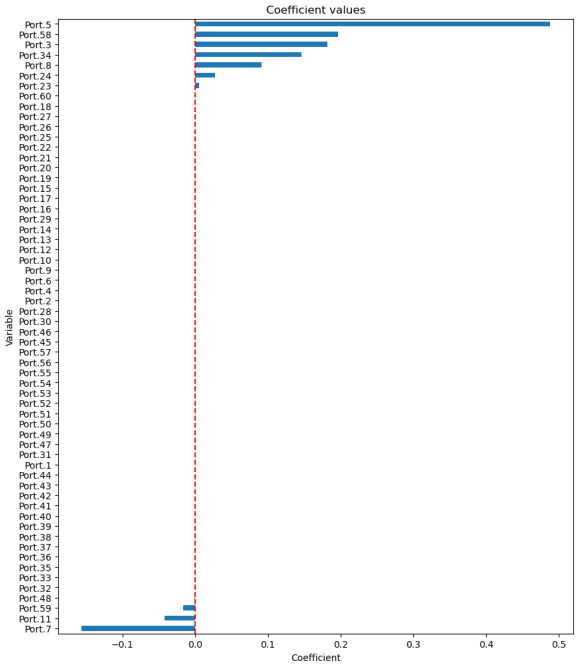

FIGURE 3.18 – Coefficients ElasticNet

Variable	Coefficient
Port.5	0.487834
Port.58	0.196097
Port.3	0.182120
Port.34	0.145900
Port.7	-0.156124
Port.8	0.090853
Port.24	0.027395
Port.11	-0.042297
Port.59	-0.016598
Port.48	-0.000944
Port.23	0.005404

FIGURE 3.19 – Variables sélectionnées par ElasticNet

Le graphique illustre les coefficients d'ElasticNet triés, indiquant les variables avec un impact significatif sur la réponse. Seules les variables avec des coefficients non nuls, c'est-à-dire différents de zéro, sont retenues dans le modèle final. Le graphique illustre les coefficients d'ElasticNet triés, indiquant les variables avec un impact significatif sur la réponse. Seules les variables avec des coefficients non nuls, c'est-à-dire différents de zéro, sont retenues dans le modèle final.

ElasticNet a sélectionné un ensemble de variables qui combine les caractéristiques des modèles Ridge et LASSO, prenant des variables communes aux deux tout en ajustant les coefficients de manière unique. Il tend à favoriser un groupe de prédicteurs qui pourrait être omis si on utilisait seulement Ridge ou LASSO, illustrant sa capacité à capter l'importance des prédicteurs dans un contexte où la multicollinéarité est présente. Cela résulte en un modèle qui pourrait offrir un meilleur compromis entre la variance et le

biais, potentiellement conduisant à une meilleure

3.2.5 Modèle Linéaire

Dans cette sous-section, notre objectif est de construire et d'évaluer un modèle linéaire classique, en commençant par les variables sélectionnées par les méthodes ElasticNet et LASSO. Nous intégrerons d'abord les prédicteurs identifiés par ElasticNet, puis nous évaluerons la significativité de ces variables dans le modèle. Si nécessaire, nous inclurons également les variables sélectionnées par LASSO. L'analyse nous permettra de déterminer quelles variables retiennent leur importance et contribuent significativement à la prédiction des émissions de CO2 dans notre modèle linéaire.

OLS Regression Results

Dep. Variable:	CO2	R-squared:	0.812
Model:	OLS	Adj. R-squared:	0.730
Method:	Least Squares	F-statistic:	9.831
Date:	Thu, 14 Dec 2023	Prob (F-statistic):	1.40e-06
Time:	06:05:03	Log-Likelihood:	-31.032
No. Observations:	37	AIC:	86.06
Df Residuals:	25	BIC:	105.4
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.4646	0.112	66.675	0.000	7.234	7.695
Port.3	0.2790	0.154	1.812	0.082	-0.038	0.596
Port.5	0.6129	0.200	3.064	0.005	0.201	1.025
Port.7	-0.2554	0.159	-1.606	0.121	-0.583	0.072
Port.8	0.2133	0.171	1.247	0.224	-0.139	0.566
Port.11	-0.2899	0.163	-1.780	0.087	-0.625	0.046
Port.23	0.1747	0.133	1.313	0.201	-0.099	0.449
Port.24	0.1061	0.121	0.880	0.387	-0.142	0.354
Port.34	0.2111	0.177	1.191	0.245	-0.154	0.576
Port.48	-0.1768	0.138	-1.283	0.211	-0.461	0.107
Port.58	0.2230	0.154	1.448	0.160	-0.094	0.540
Port.59	0.0568	0.144	0.395	0.696	-0.239	0.352

Omnibus:	0.627	Durbin-Watson:	2.184
Prob(Omnibus):	0.731	Jarque-Bera (JB):	0.676
Skew:	0.277	Prob(JB):	0.713
Kurtosis:	2.639	Cond. No.	3.88

FIGURE 3.20 – Résultats de régression ElasticNet

OLS Regression Results

Dep. Variable:	CO2	R-squared:	0.754
Model:	OLS	Adj. R-squared:	0.705
Method:	Least Squares	F-statistic:	15.33
Date:	Thu, 14 Dec 2023	Prob (F-statistic):	5.87e-08
Time:	06:07:49	Log-Likelihood:	-36.025
No. Observations:	37	AIC:	86.05
Df Residuals:	30	BIC:	97.33
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.4646	0.117	63.820	0.000	7.226	7.703
Port.3	0.2551	0.156	1.640	0.112	-0.063	0.573
Port.5	0.6490	0.208	3.125	0.004	0.225	1.073
Port.7	-0.3600	0.152	-2.375	0.024	-0.670	-0.050
Port.8	0.0509	0.144	0.354	0.726	-0.243	0.344
Port.34	0.2699	0.157	1.715	0.097	-0.051	0.591
Port.58	0.3007	0.146	2.062	0.048	0.003	0.599

Omnibus:	6.150	Durbin-Watson:	2.248
Prob(Omnibus):	0.046	Jarque-Bera (JB):	2.098
Skew:	-0.107	Prob(JB):	0.350
Kurtosis:	1.853	Cond. No.	3.54

FIGURE 3.21 – Résultats de régression LASSO

Les résultats indiquent une préférence pour le modèle LASSO en raison de son équilibre entre un R^2 ajusté satisfaisant et un nombre supérieur de variables significatives. Malgré un R^2 ajusté légèrement supérieur pour ElasticNet, signifiant une meilleure explication de la variance, LASSO présente des variables avec des coefficients plus significatifs et pertinents. Notamment, Port.5 montre une forte positivité dans les deux modèles, tandis que Port.7 et Port.8 sont significatifs uniquement dans le modèle LASSO. Cela justifie le choix de LASSO pour une utilisation future, privilégiant la pertinence des prédicteurs sur la seule explication de la variance.

Dans l'analyse, le modèle LASSO démontre une robustesse avec un R^2 ajusté de 0.754, mettant en évidence plusieurs prédicteurs significatifs : Port.5 avec un coefficient positif de 0.6490, indiquant une forte association positive avec les émissions de CO2, Port.7 avec un coefficient négatif de -0.3600, suggérant une influence inverse sur les émissions, et Port.8 avec un coefficient de 0.3007. En comparaison, ElasticNet, malgré un R^2 ajusté supérieur de 0.812, identifie principalement Port.5 comme variable significative avec un coefficient de 0.6129. L'importance donnée à Port.5 dans les deux modèles confirme son rôle clé, mais la présence d'autres variables significatives dans LASSO renforce sa sélection pour des prévisions plus détaillées et diversifiées.

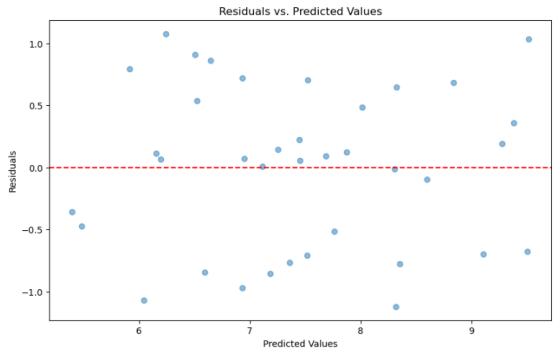

FIGURE 3.22 – Résidus vs. Valeurs Prédites par LASSO

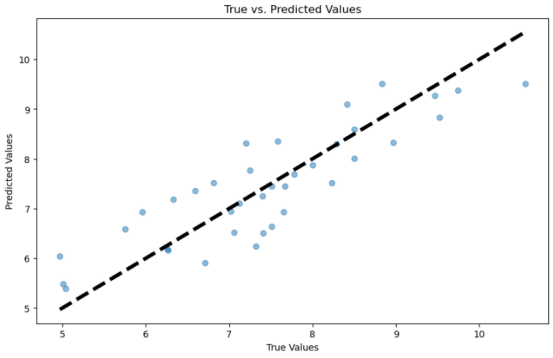

FIGURE 3.23 – Valeurs Réelles vs. Valeurs Prédites par LASSO

Les graphiques issus de la régression linéaire avec les variables sélectionnées par LASSO donnent un aperçu de la performance du modèle. Le premier graphique, montrant les résidus par rapport aux valeurs prédites, semble indiquer une distribution aléatoire des résidus, ce qui est un bon signe d'ajustement du modèle. Le second graphique illustre la relation entre les valeurs prédites et les valeurs réelles ; plus les points se rapprochent de la ligne pointillée, meilleur est le modèle. Cependant, il y a une légère tendance des prédictions à sous-estimer ou surestimer dans certains cas, indiquée par les points éloignés de la ligne.

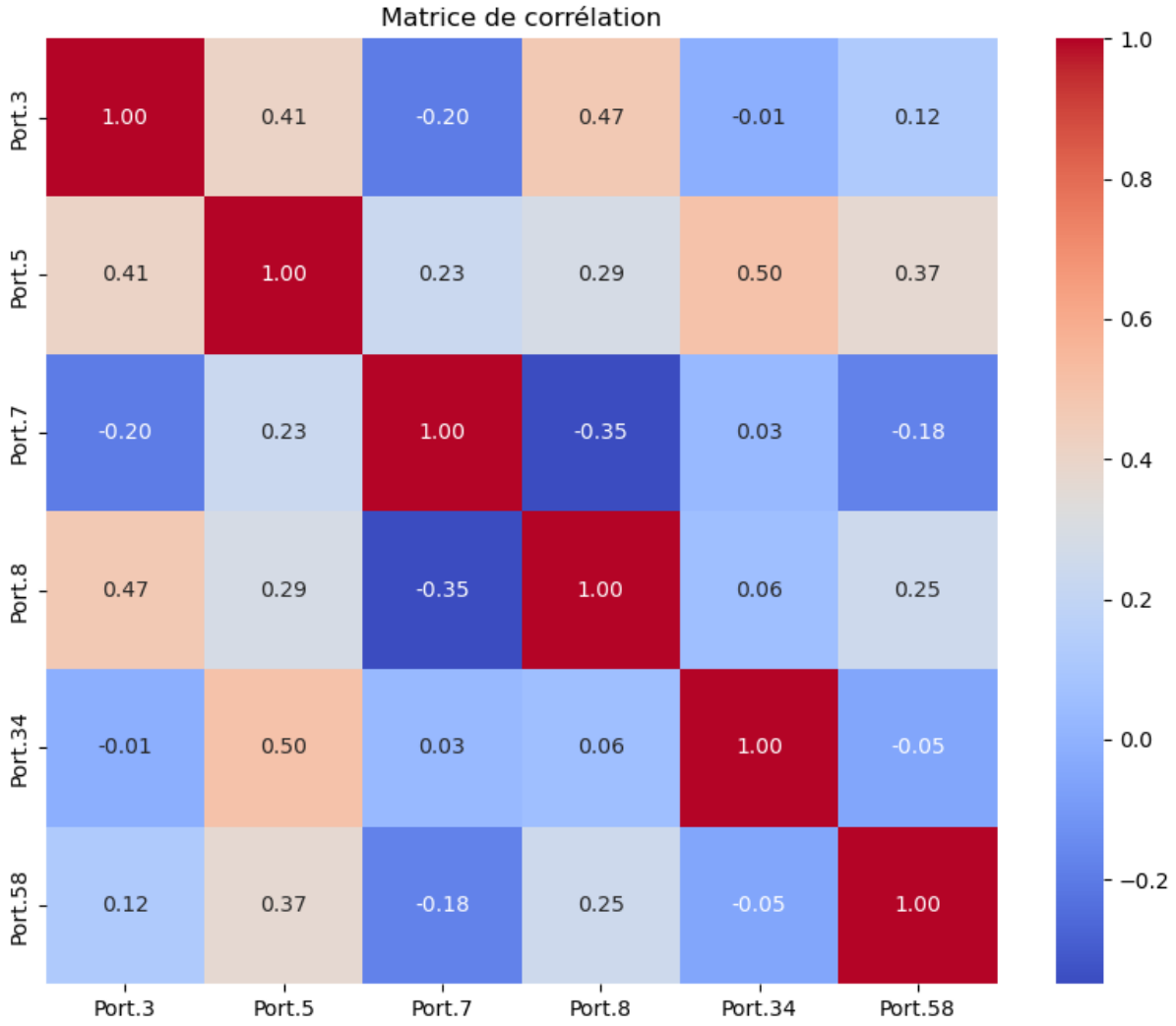

FIGURE 3.24 – Matrice de corrélation des variables sélectionnées par LASSO

Et enfin matrice de corrélation pour les variables sélectionnées par LASSO révèle l'absence de corrélation élevée entre les prédicteurs, ce qui est favorable car cela suggère peu de redondance entre les variables et réduit le risque de multicollinéarité dans le modèle linéaire. Cela valide la qualité de la sélection faite par LASSO pour la régression linéaire. La compréhension de ces relations est cruciale pour l'interprétabilité et la fiabilité des prédictions du modèle.

L'analyse a montré que les variables sélectionnées par LASSO contribuent de manière significative et indépendante à la prédiction des émissions de CO2, justifiant leur inclusion dans le modèle linéaire final.

3.2.6 Régression sur Composantes Principales

La régression sur composantes principales (PCR) sera utilisée pour analyser les données. Cette méthode combine la réduction de dimensionnalité, grâce à l'analyse en composantes principales (ACP), avec la régression linéaire. Nous déterminerons d'abord le nombre optimal de composantes principales à utiliser pour capturer la majeure partie de la variance tout en évitant le surajustement. Ensuite, nous appliquerons une régression linéaire sur ces composantes. Un seuillage sera utilisé pour identifier les coefficients les plus importants, permettant de mettre en évidence les relations les plus significatives dans le modèle réduit.

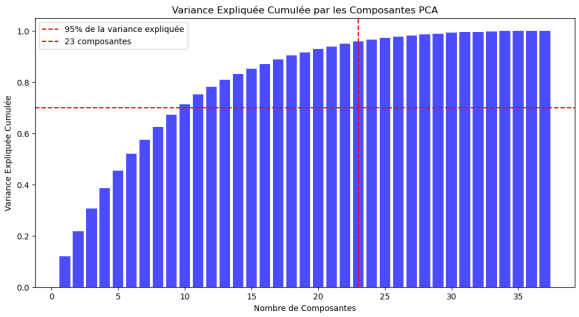

FIGURE 3.25 – Variance expliquée par les composantes

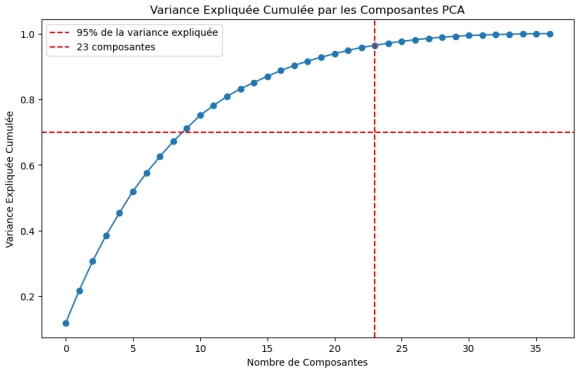

FIGURE 3.26 – Variance cumulée expliquée

Les deux graphiques illustrent la variance expliquée cumulée par le nombre de composantes principales dans une ACP. Le premier, en barres, montre clairement qu'il faut 23 composantes pour atteindre 95% de la variance expliquée. Le second, en courbe, confirme cette observation. Comprendre ce point est crucial pour la régression sur composantes principales, car il guide la quantité d'information à conserver pour une modélisation efficace sans introduire de bruit superflu.

Ce graphique montre les résultats de la validation croisée pour différents nombres de composantes PCA, où le score moyen de validation croisée s'améliore jusqu'à 8 composantes avant de chuter, suggérant que 8 composantes sont suffisantes pour un bon équilibre entre performance du modèle et complexité. Avec 63% de la variance expliquée, ces 8 composantes constituent un compromis entre la simplicité du modèle et sa capacité à capturer les informations essentielles des données.

Le graphique illustre les coefficients de la régression sur composantes principales (PCR)

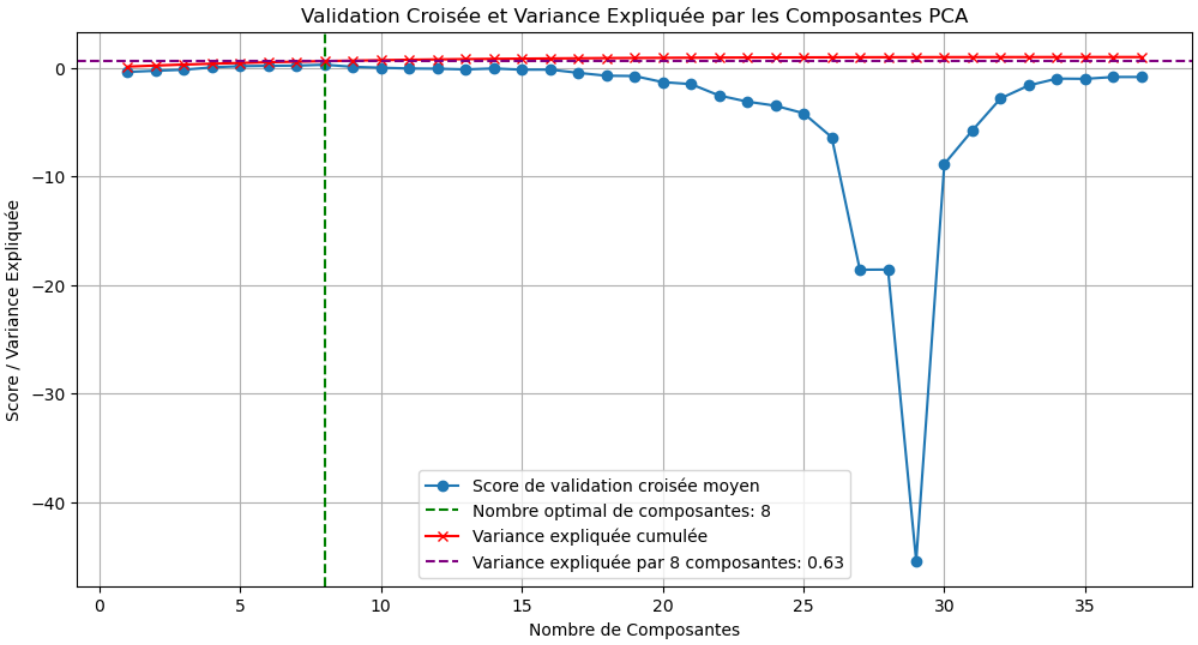

FIGURE 3.27 – Validation croisée et variance expliquée par PCA

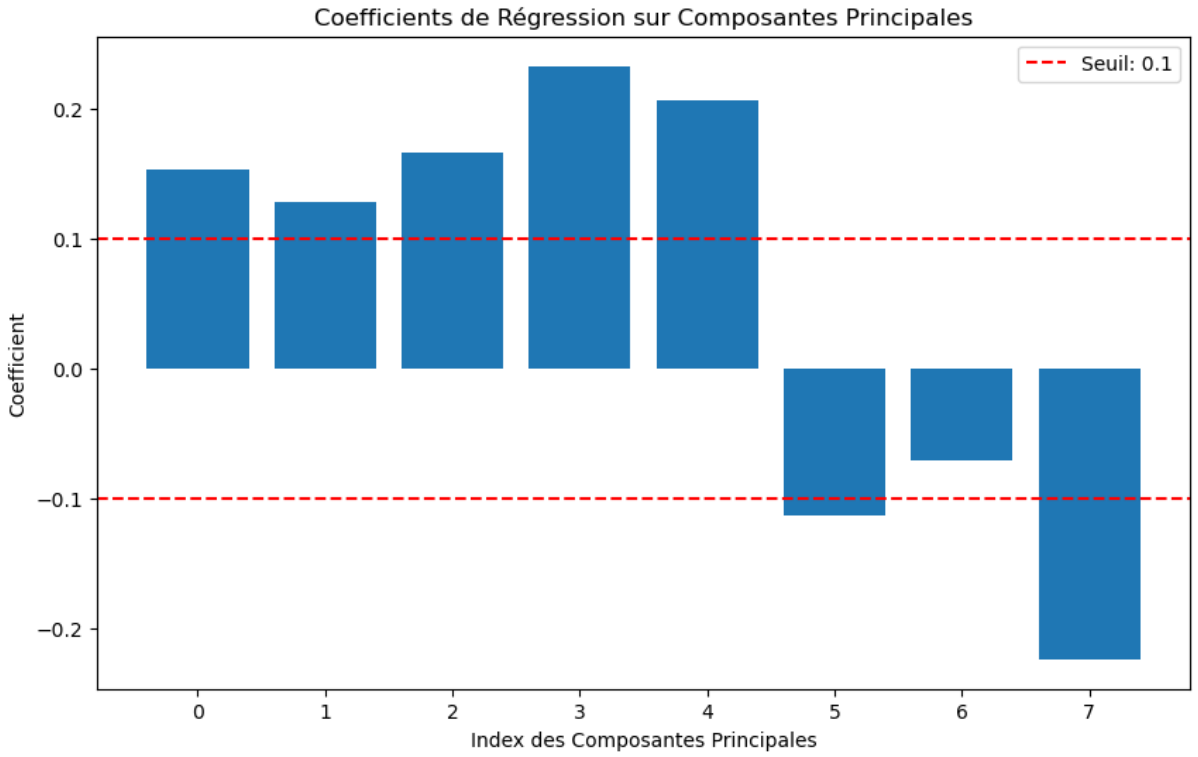

FIGURE 3.28 – Coefficients de Régression sur Composantes Principales avec seuillage à 0.1

en se basant sur un seuil de 0.1. Les barres représentent l'influence de chaque composante principale sur le modèle de régression. Les composantes qui dépassent le seuil, en valeur absolue, sont considérées comme ayant un impact significatif sur la variable dépendante. Dans ce cas, les composantes 0, 1, 2, 3, 4, 5 et 7 sont au-dessus du seuil et donc retenues comme importantes pour le modèle.

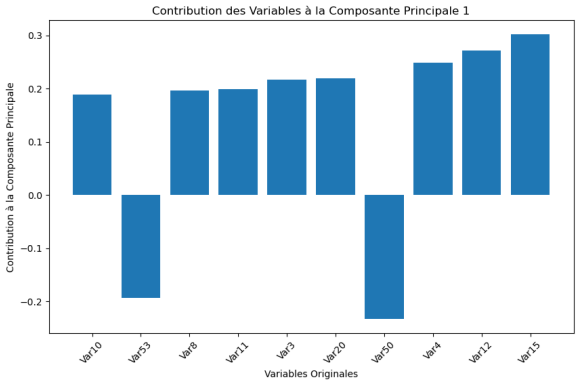

FIGURE 3.29 – Contribution à la composante principale 1

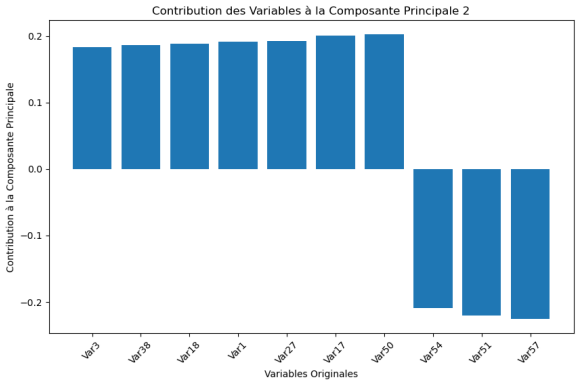

FIGURE 3.30 – Contribution à la composante principale 2

Les graphiques dépeignent les coefficients significatifs de la régression sur composantes principales avec un seuil de 0.1 et les contributions des variables originales aux deux premières composantes principales. Ces visualisations aident à comprendre les éléments clés qui influencent le modèle, mettant en évidence les variables les plus influentes pour chaque composante principale. Les graphiques dépeignent les coefficients significatifs de la régression sur composantes principales avec un seuil de 0.1 et les contributions des variables originales aux deux premières composantes principales. Ces visualisations aident à comprendre les éléments clés qui influencent le modèle, mettant en évidence les variables les plus influentes pour chaque composante principale.

La régression sur composantes principales (PCR) a permis de réduire la dimensionnalité de notre ensemble de données tout en conservant les informations essentielles. En sélectionnant les composantes principales qui expliquent une proportion substantielle de la variance, nous avons simplifié le modèle sans sacrifier significativement la précision. Cependant, un inconvénient de la PCR est qu'elle peut rendre l'interprétation des composantes difficiles, car elles sont des combinaisons linéaires des variables originales. Cette

méthode est avantageuse lorsque la multicollinéarité est présente, mais elle peut également masquer les contributions individuelles des variables originales. La section suivante examinera l'approche de la régression PLS, qui peut offrir une meilleure interprétabilité dans certains cas.

3.2.7 Régression PLS (Partial Least Squares)

La régression PLS a démontré son efficacité en identifiant les variables les plus pertinentes avec seulement deux composantes principales, contrairement à la PCR qui en nécessitait 23 pour une variance expliquée similaire. Cela indique que la PLS est plus efficace pour condenser l'information utile dans moins de composantes, offrant ainsi un modèle potentiellement plus simple et plus interprétable sans sacrifier la performance. Comparativement à la PCR, la PLS semble être un choix plus judicieux pour ce jeu de données particulier.

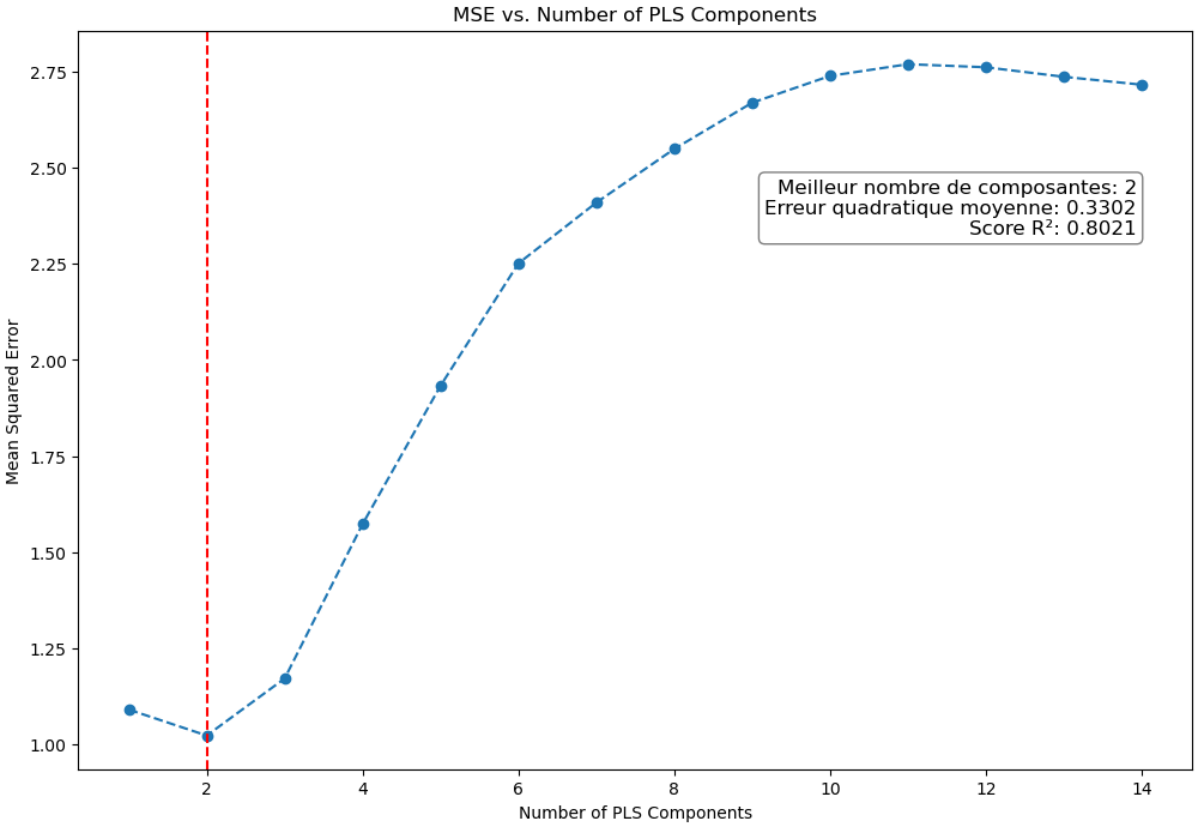

FIGURE 3.31 – Validation croisée pour la sélection des composantes PLS

Le graphique présente une analyse de validation croisée pour déterminer le nombre

optimal de composantes dans un modèle de régression PLS. Avec seulement 2 composantes, le modèle atteint un score R^2 de 0.8021, indiquant une forte capacité à prédire la variable dépendante, tout en conservant une erreur quadratique moyenne (MSE) relativement faible de 0.3302. Cela suggère que ces deux composantes captent une information significative avec une perte minimale de détails pertinents.

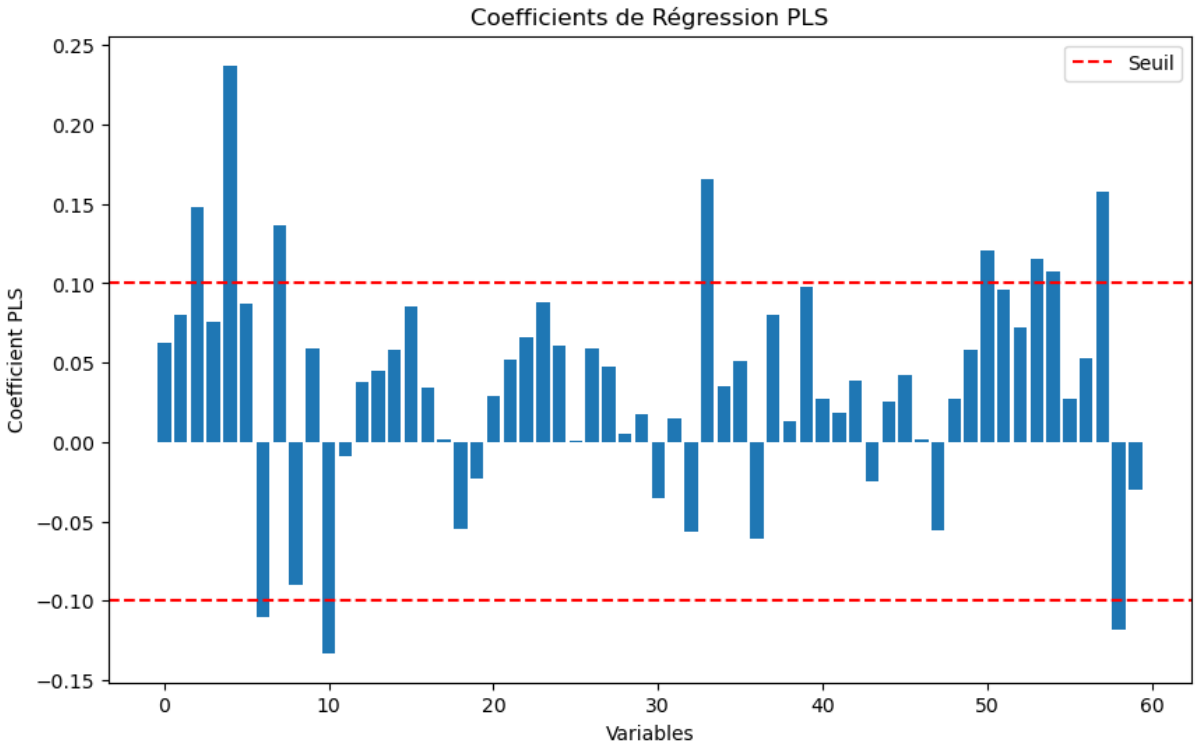

FIGURE 3.32 – Coefficients significatifs obtenus par la régression PLS avec seuillage à 0.1

Ce graphique présente les coefficients obtenus par la régression PLS, avec un seuil fixé à 0.1 pour déterminer les variables significatives. Les barres dépassant la ligne rouge horizontale représentent les variables ayant un impact notable sur la réponse prédite. Les variables importantes identifiées sont celles associées aux indices [2, 4, 6, 7, 10, 33, 50, 53, 54, 57, 58]. Cette information est précieuse pour cibler les variables les plus influentes dans le modèle PLS et pour l'interprétation des résultats dans le contexte de l'ensemble de données.

Les 15 coefficients les plus importants issus de la régression PLS, sélectionnés sur la base d'un seuil supérieur à 0.1, reflètent les variables ayant le plus d'influence sur la variable dépendante. Les valeurs positives indiquent une relation directe avec la réponse, tandis que les valeurs négatives suggèrent une relation inverse. Ces variables devraient

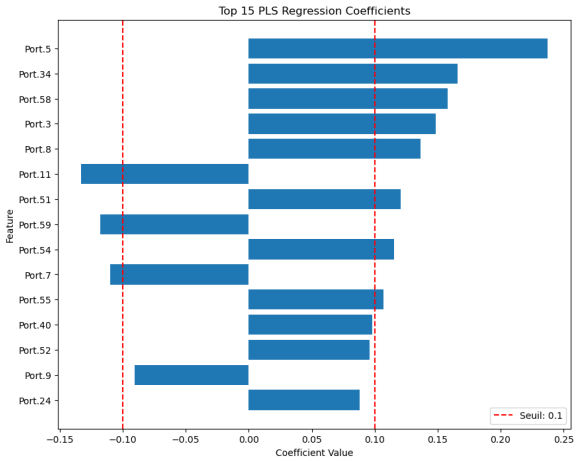

FIGURE 3.33 – Les 15 coefficients les plus importants de la régression PLS

Variable	Coefficient
Port.58	0.158038
Port.34	0.165847
Port.3	0.148326
Port.5	0.237203
Port.8	0.136357
Port.54	0.115342
Port.51	0.120846
Port.55	0.107097
Port.11	-0.132948
Port.59	-0.118077
Port.7	-0.109989

FIGURE 3.34 – Coefficients importants de la régression PLS triés par valeur absolue

donc être prises en compte de manière prioritaire dans l'analyse et l'interprétation des modèles.

La régression PLS a démontré son efficacité en identifiant les variables les plus pertinentes avec seulement deux composantes principales, contrairement à la PCR qui en nécessitait 23 pour une variance expliquée similaire. Cela indique que la PLS est plus efficace pour condenser l'information utile dans moins de composantes, offrant ainsi un modèle potentiellement plus simple et plus interprétable sans sacrifier la performance. Comparativement à la PCR, la PLS semble être un choix plus judicieux pour ce jeu de données particulier.

3.3 Modélisation du nombre d'incidents

3.3.1 Analyse de la variable Incid

Dans cette sous-section, nous allons entreprendre une analyse descriptive de la variable "Incid", qui représente le nombre d'incidents. Nous examinerons sa distribution, rechercherons des valeurs aberrantes, et évaluerons les tendances ou modèles existants. Cette étape analytique initiale est essentielle pour orienter nos méthodes de modélisation

statistique futures.

Statistique	Valeur
Nombre	37
Moyenne	34.70
Écart-type	32.41
Minimum	3
25ème percentile	13
Médiane	27
75ème percentile	42
Maximum	152

TABLE 3.2 – Statistiques descriptives de la variable 'Incid'

Incid	Fréquence
5	2
15	2
... autres lignes ...	
152	1

TABLE 3.3 – Distribution des comptages de 'Incid'

Les statistiques descriptives indiquent une variabilité significative dans les incidents avec un écart-type élevé par rapport à la moyenne. La distribution est asymétrique, comme le suggère la différence entre la médiane et la moyenne, et la présence de valeurs extrêmes, avec un maximum de 152 incidents. La distribution des fréquences montre également que certains nombres d'incidents se répètent plus fréquemment, indiquant des modèles potentiels dans les données. Ces observations peuvent influencer la sélection de modèles statistiques appropriés pour la modélisation de la variable 'Incid'.

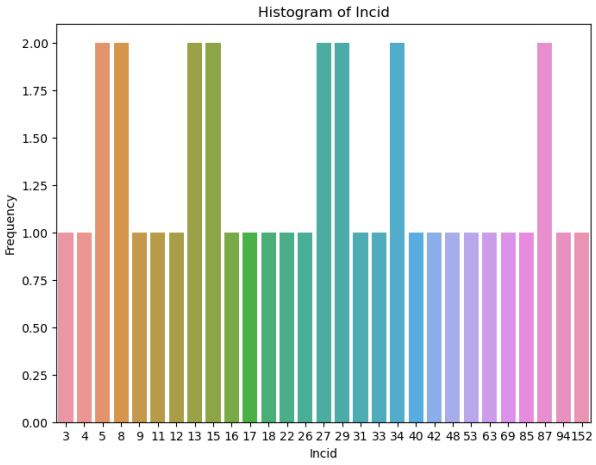

FIGURE 3.35 – Histogramme des incidents

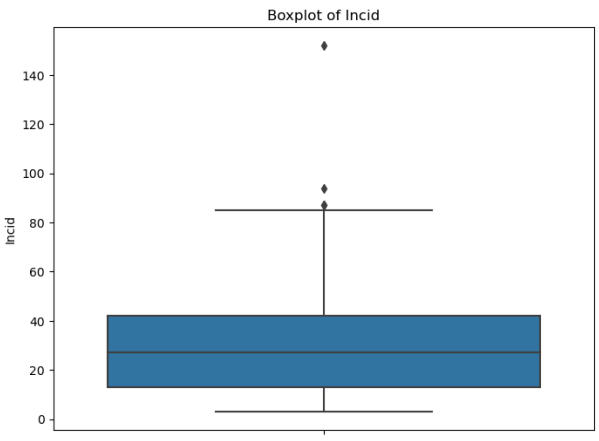

FIGURE 3.36 – Boxplot des incidents

Le boxplot et l'histogramme de la variable "Incid" offrent une vue complémentaire de

la distribution des incidents. Le boxplot révèle clairement la présence de valeurs extrêmes, suggérant quelques incidents anormalement élevés par rapport à la majorité des données. L'histogramme, quant à lui, illustre la variabilité des incidents avec certains nombres qui se répètent plus souvent. Ensemble, ces graphiques mettent en évidence la distribution asymétrique des incidents et la nécessité de considérer l'impact des valeurs extrêmes dans la modélisation.

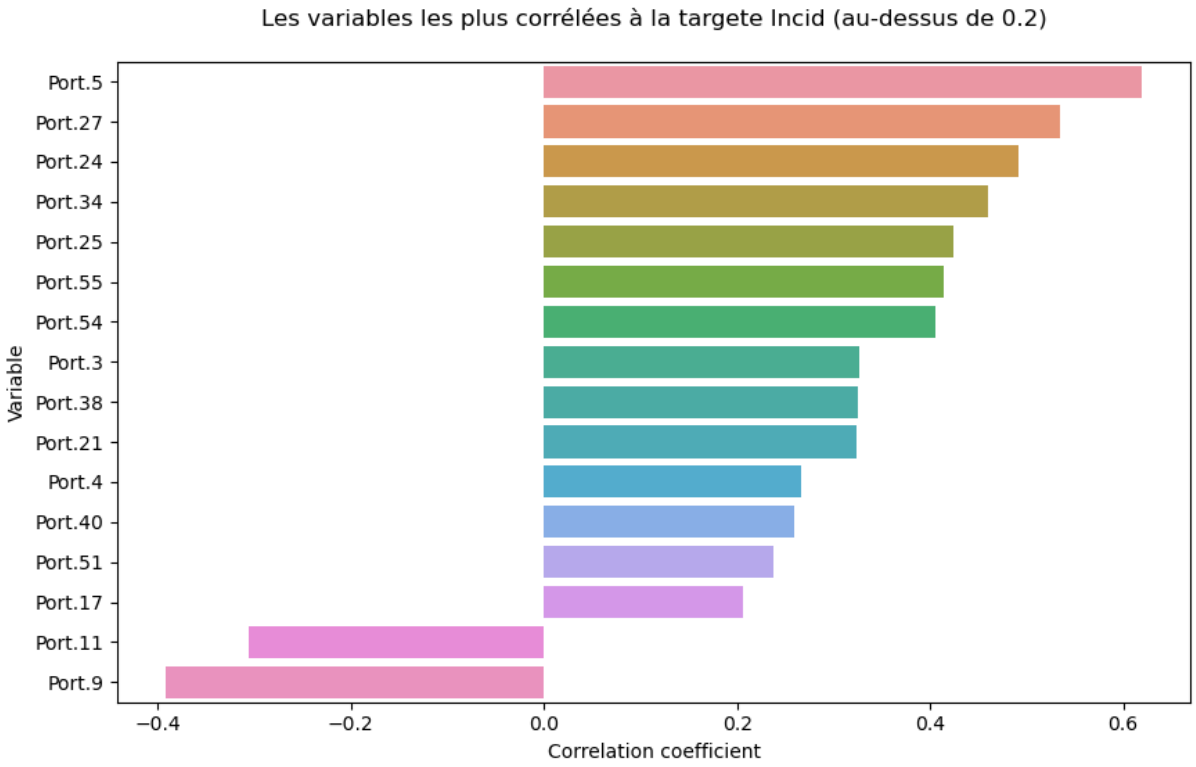

FIGURE 3.37 – Corrélations des variables avec Incid

Ce graphique illustre les variables ayant une corrélation supérieure à 0.2 avec la variable cible 'Incid'. Ces informations sont précieuses pour la sélection de caractéristiques dans la construction de modèles prédictifs, indiquant les variables potentiellement les plus informatives.

3.3.2 Modélisation du nombre d'incidents avec la régression de Poisson Ridge

Avant de procéder à la régression de Poisson Ridge, une normalisation des variables explicatives est effectuée pour une régularisation équitable via le paramètre lambda. Une

optimisation de ce dernier sera accomplie par une boucle de validation croisée. À partir de cette étape, le reste du projet est réalisé en R, ce qui peut induire des variations dans les résultats obtenus par rapport à ceux générés par Python.

Nous avons réalisé une régression de Poisson Ridge en stabilisant le choix du paramètre de régularisation λ optimal par une méthode de validation croisée. Les graphiques suivants illustrent les étapes de cette sélection. graphicx caption

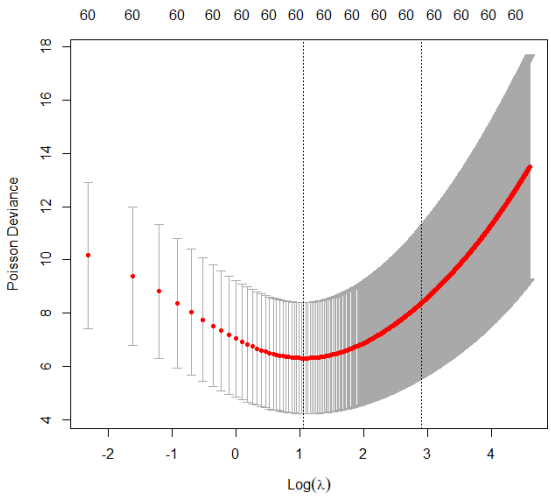

FIGURE 3.38 – Déviance de Poisson pour différentes valeurs de $\log(\lambda)$, montrant la déviance moyenne (points rouges) et l'intervalle de confiance (zone ombrée).

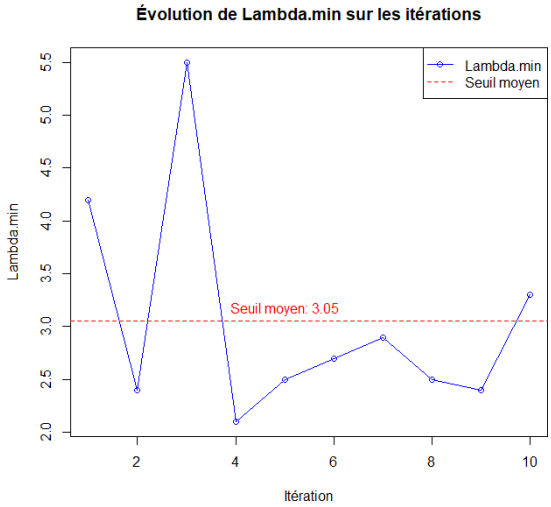

FIGURE 3.39 – Évolution de λ_{min} à travers les itérations de la validation croisée, avec le seuil moyen (ligne rouge pointillée) utilisé dans le modèle.

L'analyse des résultats indique que le choix du λ a un impact significatif sur la performance du modèle. La Figure 3.38 illustre l'effet de différents λ sur la déviance du modèle. Dans la Figure 3.39, l'évolution du λ_{min} au cours des itérations de la validation croisée est représentée, où le seuil moyen de 3.05 est établi comme le λ final pour le modèle.

Après avoir mené la régression de Poisson Ridge avec le seuil moyen pour le paramètre λ, nous avons identifié les 20 coefficients les plus significatifs. Le graphique ci-dessous illustre ces coefficients, mettant en évidence ceux qui dépassent la valeur de 0.1, comme indiqué par les lignes rouges verticales.

La Figure 3.40 montre les coefficients des variables, indiquant leur influence relative

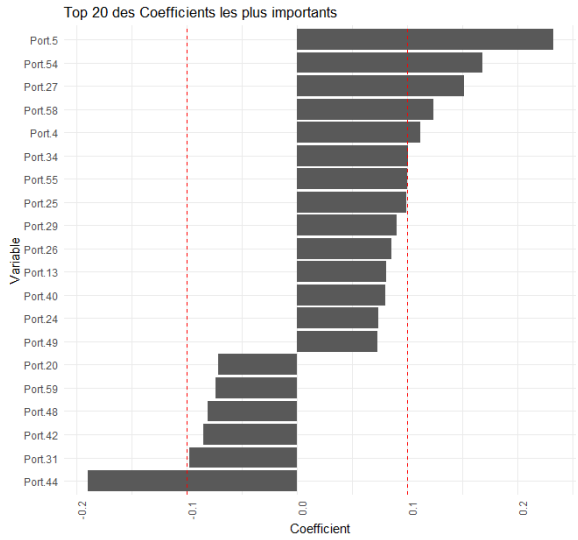

FIGURE 3.40 – Les 20 coefficients les plus importants issus de la régression de Poisson Ridge.

Variable	Coefficient
Port.5	>0.1
Port.44	>0.1
Port.54	>0.1
Port.27	>0.1
Port.58	>0.1
Port.4	>0.1
Port.34	>0.1
Port.55	>0.1

FIGURE 3.41 – Variables sélectionnées

dans le modèle de régression. Les variables listées dans le Tableau 3.41 sont celles qui ont un impact substantiel, avec des coefficients dépassant le seuil de 0.1. Cette identification aide à comprendre les variables les plus pertinentes influençant la réponse prédite par le modèle.

La matrice de corrélation présentée ci-dessous fournit un aperçu des relations linéaires entre les variables sélectionnées par la régression de Poisson Ridge. Chaque cellule indique le coefficient de corrélation entre deux variables, variant de -1 à 1. Une valeur proche de 1 suggère une forte corrélation positive, une valeur proche de -1 indique une forte corrélation négative, et une valeur autour de 0 implique l'absence de corrélation.

L'analyse de cette matrice est cruciale, car la régression de Poisson Ridge est sensible à la multicollinéarité entre les prédicteurs. En identifiant les paires de variables hautement corrélées, je peux prendre des décisions éclairées concernant l'exclusion de prédicteurs redondants ou la nécessité de les combiner. La régression Ridge est bénéfique dans ce contexte, car elle réduit la variance des estimations des coefficients en pénalisant les coefficients de corrélation élevés, ce qui est un avantage lorsqu'il y a des prédicteurs corrélés. Cependant, un inconvénient est que la méthode n'effectue pas de sélection de variables, ce

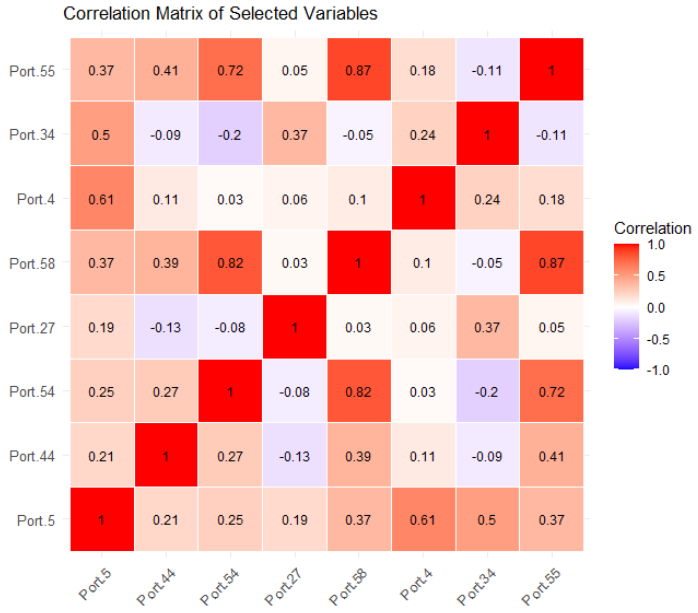

FIGURE 3.42 – Matrice de corrélation des variables sélectionnées.

qui signifie que toutes les variables sélectionnées, même celles avec de faibles coefficients, restent dans le modèle final.

Il est à noter que des variables comme ‘Port.5‘ et ‘Port.55‘ présentent des corrélations significatives avec plusieurs autres prédicteurs. Cela suggère leur centralité dans le réseau de corrélation et leur potentiel impact sur le modèle. La compréhension de ces interactions est essentielle pour interpréter correctement le modèle de régression et pour formuler des conclusions solides sur les facteurs influençant le nombre d'incidents.

3.3.3 Application de la Régression de Poisson LASSO

Dans cette sous-section, l'objectif est de mettre en œuvre une régression de Poisson LASSO pour identifier les variables les plus influentes dans la prédiction du nombre d'incidents. Le LASSO, ou Least Absolute Shrinkage and Selection Operator, est une technique de régularisation qui non seulement pénalise la complexité du modèle avec un terme de régularisation proportionnel à l'absolu des coefficients, mais permet également une sélection de variables en réduisant directement certains coefficients à zéro.

Le choix du paramètre de régularisation λ est déterminant dans la régression LASSO. Deux méthodes courantes pour sélectionner λ seront explorées : λ_{min}, qui correspond au lambda où l'erreur de prédiction est minimisée, et λ_{1se}, qui est le plus grand λ tel que l'erreur est dans une fourchette d'une erreur standard de l'erreur minimale. La décision entre ces deux seuils sera prise en considérant le compromis entre la précision de prédiction et la simplicité du modèle.

Finalement, après avoir déterminé le seuil de λ approprié, les variables retenues par le modèle LASSO seront présentées. Ces variables représentent les 'ports' les plus pertinents qui influencent le nombre d'incidents, selon les critères de sélection du LASSO. Cette approche vise à fournir un modèle à la fois parcimonieux et performant en termes de prévision.

La régression de Poisson LASSO a été appliquée pour identifier les variables les plus pertinentes dans le modèle. Les graphiques suivants présentent les résultats obtenus pour le choix du paramètre de régularisation λ.

La Figure 3.43 montre la variation de la déviance pour différentes valeurs de λ, permettant de visualiser le point où la déviance est minimisée. Ceci est essentiel pour comprendre à quel point la complexité du modèle affecte la prédiction. La Figure 3.44 trace l'évolution du λ_{1se} sur plusieurs itérations, indiquant la robustesse du seuil de sélection des variables. Le seuil moyen de λ est utilisé pour garantir que le modèle reste généralisable et non surajusté aux données d'entraînement.

La régression LASSO a permis de réduire le nombre de variables en assignant un coefficient nul à celles qui sont les moins significatives. Le graphique ci-dessous montre les 20 coefficients les plus importants qui sont différents de zéro, soulignant l'efficacité de LASSO dans la sélection des variables.

Les variables présentées dans le tableau 3.47 sont celles identifiées comme ayant un impact significatif sur le modèle. Cette sélection indique leur pertinence dans la prédiction du nombre d'incidents, faisant de la régression LASSO un outil efficace pour la simplifi-

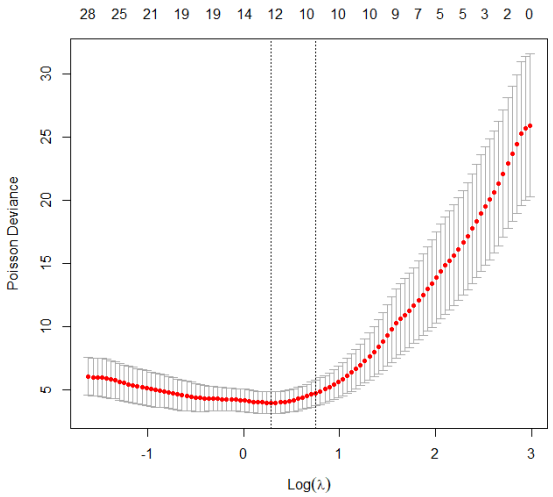

FIGURE 3.43 – Déviance de Poisson en fonction de log(λ) pour LASSO. Les points rouges marquent la déviance moyenne à chaque valeur de λ, avec l'intervalle de confiance indiqué par les barres d'erreur.

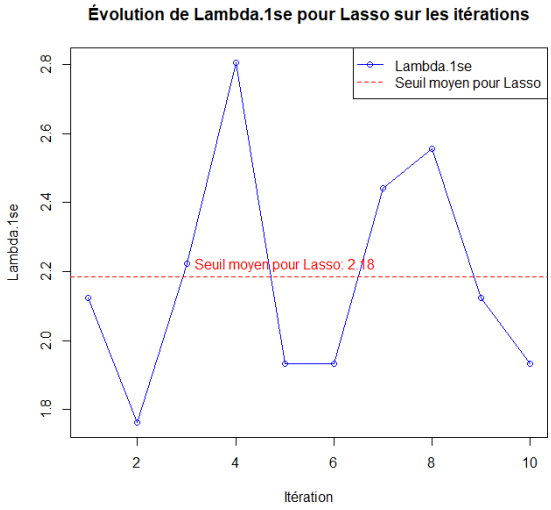

FIGURE 3.44 – Évolution de λ_{1se} à travers les itérations pour LASSO. Le seuil moyen indiqué par la ligne rouge pointillée est de 2.18.

FIGURE 3.45 – Analyse de la déviance et du choix de λ dans la régression LASSO.

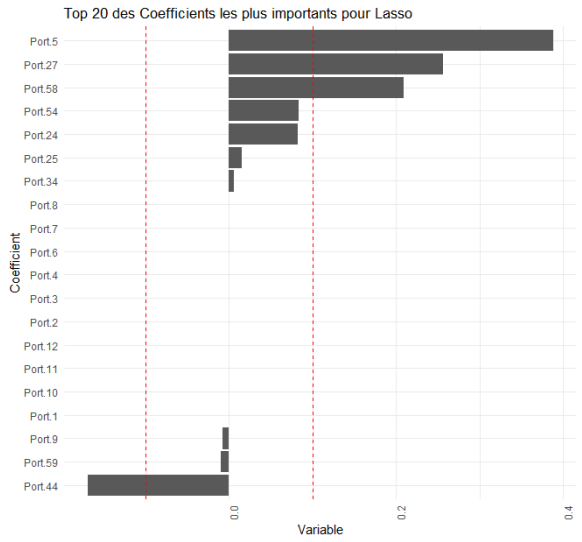

FIGURE 3.46 – Top 20 des coefficients les plus importants pour LASSO.

Variable	Coefficient
Port.5	Non nul
Port.27	Non nul
Port.58	Non nul
Port.44	Non nul
Port.54	Non nul
Port.24	Non nul
Port.25	Non nul
Port.59	Non nul
Port.9	Non nul
Port.34	Non nul

FIGURE 3.47 – Variables avec coefficients non nuls sélectionnées par LASSO.

FIGURE 3.48 – Comparaison des coefficients significatifs et des variables sélectionnées par LASSO.

cation du modèle tout en conservant les prédicteurs clés.

Après la sélection de variables par la régression LASSO, une matrice de corrélation a été construite pour évaluer les relations linéaires entre les prédicteurs retenus.

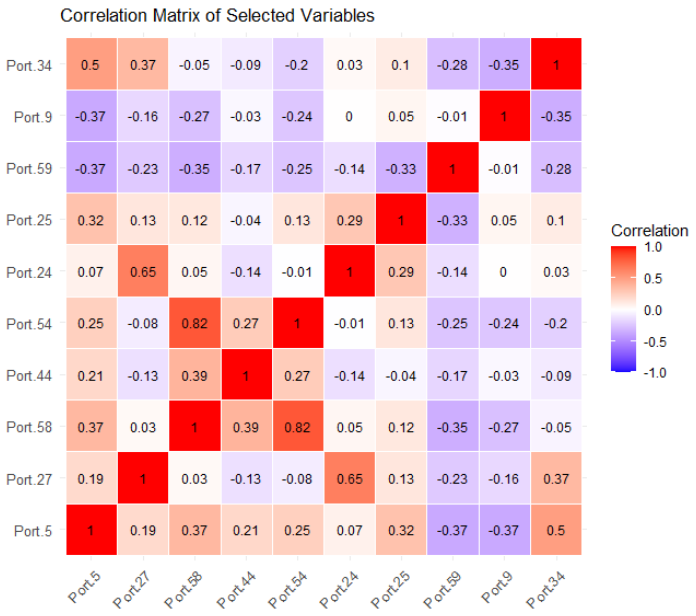

FIGURE 3.49 – Matrice de corrélation des variables sélectionnées par LASSO.

La Figure 3.49 montre la matrice de corrélation pour les variables sélectionnées par LASSO. Contrairement à la régression Ridge, où tous les prédicteurs restent dans le modèle avec des coefficients potentiellement diminués, LASSO a l'avantage de réduire explicitement certains coefficients à zéro, ce qui simplifie le modèle en excluant les variables non significatives. Cependant, cela peut aussi être considéré comme un inconvénient si la suppression de certaines variables entraîne la perte d'informations importantes. Les corrélations fortes entre certaines variables, comme 'Port.5' et 'Port.27', sont conservées, ce qui indique qu'elles partagent une quantité significative d'information. En revanche, des corrélations négatives, telles que celles entre 'Port.9' et 'Port.34', suggèrent des relations inverses.

Cette analyse des corrélations est essentielle pour comprendre les dynamiques entre les variables et pour assurer que les prédicteurs finaux fournissent une vue équilibrée du

système étudié, sans redondance ni omission critique.

3.3.4 Optimisation du Modèle par Régression ElasticNet

La régression ElasticNet combine les propriétés de régularisation du LASSO et de Ridge pour améliorer la sélection de variables et la robustesse du modèle. Dans cette section, une régression ElasticNet sera effectuée pour déterminer les variables qui ont un impact significatif sur le nombre d'incidents. Cette méthode est particulièrement utile lorsque les données présentent des corrélations entre les prédicteurs et lorsque le compromis entre la sélection de variables et la réduction de la variance est crucial.

L'objectif ici est de trouver le meilleur paramètre alpha, qui équilibre les pénalités de LASSO et Ridge dans la fonction de perte d'ElasticNet. Un alpha optimal signifie un modèle qui n'est ni trop complexe, risquant de surajuster, ni trop simplifié, ce qui pourrait omettre des informations importantes. Les démarches à suivre seront similaires à celles utilisées pour LASSO et Ridge, comprenant la normalisation des variables, la sélection de λ, et l'analyse des coefficients résultants.

En fin de compte, cette approche vise à obtenir un modèle prédictif fiable qui conserve les avantages de LASSO et Ridge tout en atténuant leurs limites respectives.

La régression ElasticNet a été appliquée et les résultats sont illustrés ci-dessous. Le graphique de gauche montre l'évolution de λ_{1se} sur plusieurs itérations, avec le seuil moyen indiqué, tandis que le graphique de droite présente les 20 coefficients les plus importants identifiés par le modèle.

En bas, un tableau résume les variables dont les coefficients sont significativement supérieurs à 0.1, indiquant une influence notable sur le modèle.

Les résultats montrent une sélection rigoureuse des variables, réduisant la complexité du modèle tout en conservant les prédicteurs pertinents. Le seuil de λ_{1se} indique le compromis entre biais et variance, assurant la généralisabilité du modèle.

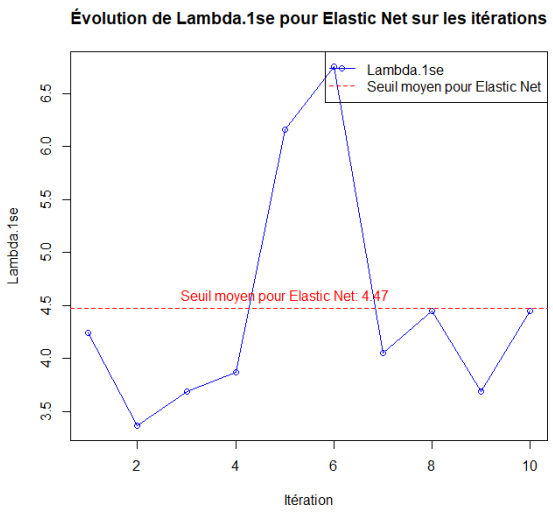

FIGURE 3.50 – Évolution de λ_{1se} pour ElasticNet.

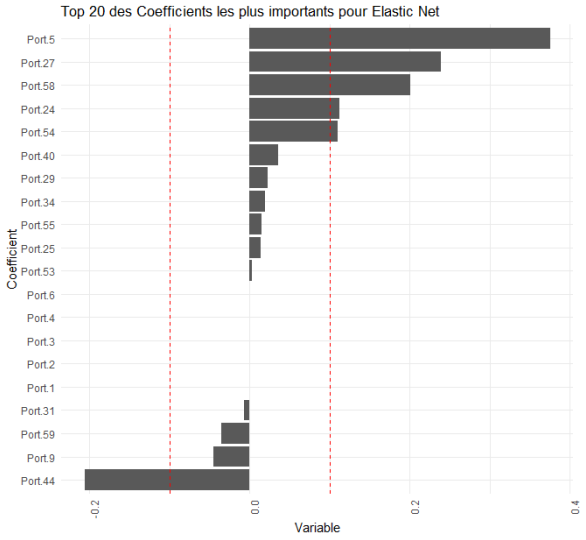

FIGURE 3.51 – Top 20 des coefficients les plus importants pour ElasticNet.

Variable	Coefficient
Port.5	>0.1
Port.27	>0.1
Port.44	>0.1
Port.58	>0.1
Port.24	>0.1
Port.54	>0.1

TABLE 3.4 – Variables avec coefficients supérieurs à 0.1 dans le modèle ElasticNet.

3.3.5 Comparaison entre Régression de Poisson et Binomiale Négative

Dans cette section, l'objectif est d'évaluer la performance de deux modèles statistiques couramment utilisés pour modéliser des données de comptage : la régression de Poisson et la régression binomiale négative. Ces modèles seront comparés pour déterminer lequel est le plus adapté aux données en présence de surdispersion.

Tout d'abord, une régression de Poisson sera réalisée en utilisant les variables identifiées comme étant les plus pertinentes dans les analyses précédentes. Le modèle de Poisson suppose que la moyenne et la variance des données sont égales, une condition qui n'est pas toujours satisfaite dans la pratique. Pour évaluer la présence de surdispersion, c'est-à-dire si la variance des données dépasse la moyenne, les résidus de Pearson seront calculés et examinés.

Si une surdispersion est détectée, ce qui est indiqué par la somme des carrés des résidus de Pearson divisée par le nombre d'observations moins le nombre de paramètres (n-p), un modèle binomial négatif sera alors proposé. Ce dernier peut accommoder la surdispersion en permettant à la variance de dépasser la moyenne, ce qui pourrait améliorer la qualité de l'ajustement du modèle aux données.

Enfin, l'effet de la variable la plus significative sera interprété pour comprendre son influence sur le nombre d'incidents. Cela aidera à déterminer non seulement la qualité de l'ajustement du modèle, mais aussi la pertinence des variables incluses pour expliquer la variation observée dans les données de comptage.

Une régression de Poisson a été réalisée en utilisant les variables sélectionnées précédemment pour déterminer leur effet sur le nombre d'incidents. L'examen des résidus de Pearson permet de vérifier l'hypothèse d'équidispersion inhérente au modèle de Poisson, où la moyenne est égale à la variance.

Un indicateur de surdispersion a été calculé comme la somme des carrés des résidus de Pearson divisée par le nombre de degrés de liberté (n-p). La valeur de l'indicateur est présentée dans le tableau ci-dessous :

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.18598    0.03772  84.470  < 2e-16 ***
Port.5       0.48773    0.03230  15.100  < 2e-16 ***
Port.27      0.31752    0.04268   7.440 1.01e-13 ***
Port.44     -0.36847    0.04179  -8.818  < 2e-16 ***
Port.58      0.41853    0.06108   6.852 7.27e-12 ***
Port.24      0.06534    0.04561   1.433    0.152
Port.54      0.03957    0.04865   0.813    0.416
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 919.026  on 36  degrees of freedom
Residual deviance:  47.321  on 30  degrees of freedom
AIC: 245.97

Number of Fisher Scoring iterations: 4
```

FIGURE 3.52 – Résultats de la régression de Poisson classique.

Mesure	**Valeur**
Indicateur de Surdispersion	1.563

TABLE 3.5 – Indicateur de surdispersion pour le modèle de Poisson.

La valeur obtenue de l'indicateur de surdispersion est de 1.563, ce qui suggère que la variance des données est plus grande que la moyenne, indiquant la présence de surdispersion dans le modèle. Cela peut conduire à des estimations de la variance des coefficients qui sont sous-évaluées, ce qui rend les tests d'hypothèses moins fiables. Pour remédier à cela, un modèle binomial négatif sera considéré, car il peut accommoder une variance excédant la moyenne, offrant ainsi une meilleure adaptation aux données présentant de la surdispersion.

Face à la surdispersion mise en évidence dans le modèle de Poisson, un modèle binomial négatif a été ajusté pour mieux s'adapter aux données. Cette approche permet de gérer la variance des données qui est supérieure à la moyenne, offrant ainsi des estimations plus fiables.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.18592    0.03775  84.399  < 2e-16 ***
Port.5       0.48780    0.03236  15.076  < 2e-16 ***
Port.27      0.31752    0.04274   7.428 1.10e-13 ***
Port.44     -0.36861    0.04184  -8.809  < 2e-16 ***
Port.58      0.41843    0.06117   6.841 7.89e-12 ***
Port.24      0.06540    0.04567   1.432    0.152
Port.54      0.03979    0.04876   0.816    0.414
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(14211.32) family taken to be 1)

    Null deviance: 916.375  on 36  degrees of freedom
Residual deviance:  47.231  on 30  degrees of freedom
AIC: 247.97

Number of Fisher Scoring iterations: 1
```

FIGURE 3.53 – Résultats de la régression binomiale négative.

Les résultats montrent que certaines variables ont des coefficients estimés avec une grande significativité, comme indiqué par les valeurs de p faibles. La variable ‘Port.5‘, par exemple, avec un coefficient estimé de 0.48780 et une valeur de p extrêmement faible, est très significative. Cela indique que, toutes choses étant égales par ailleurs, une augmentation unitaire de ‘Port.5‘ est associée à une augmentation exponentielle du nombre d’incidents.

Les autres variables significatives, telles que ‘Port.27‘, ‘Port.44‘, et ‘Port.58‘, montrent également des effets importants sur la réponse. En particulier, ‘Port.44‘ avec un coefficient de -0.36861 et une valeur de p significative, suggère qu’une augmentation de cette variable est associée à une diminution du nombre d’incidents.

Ces résultats offrent une compréhension approfondie des facteurs qui influencent le nombre d’incidents, permettant d’orienter des stratégies efficaces pour leur gestion. Le modèle binomial négatif se présente ainsi comme un outil robuste pour analyser des données de comptage avec surdispersion.

3.4 Modélisation de la Variable Chiffre d'Affaires (CA)

Dans cette partie du rapport, l'objectif est d'étudier l'impact des quantités livrées dans différents ports sur le niveau du chiffre d'affaires. Pour ce faire, différentes techniques de régression logistique seront appliquées pour modéliser et comprendre les relations entre les quantités livrées et le CA.

Une régression logistique Ridge sera mise en œuvre en premier lieu. Cette méthode permettra de gérer la multicollinéarité potentielle entre les prédicteurs et de stabiliser la sélection des variables importantes. Une attention particulière sera accordée au choix du paramètre de régularisation λ, en utilisant une méthode de validation croisée pour trouver la valeur optimale. Si le λ optimal se trouve à l'extrémité de l'intervalle, la grille des valeurs de λ sera ajustée. Une méthode de sélection de variables sera proposée et les ports les plus influents sur le CA seront identifiés.

Ensuite, une régression logistique LASSO sera réalisée. Cette approche vise à réduire le nombre de variables en poussant les coefficients de certaines d'entre elles vers zéro. Un seuil pour le choix de λ sera défini, en considérant à la fois λ_{min} et λ_{1se}, pour sélectionner un ensemble de variables optimal. Les ports retenus par cette méthode seront également présentés.

Si nécessaire, une régression ElasticNet sera effectuée pour combiner les avantages des régressions Ridge et LASSO. Cette technique sera utilisée pour affiner davantage le modèle, en recherchant le meilleur paramètre alpha qui équilibre les deux types de régularisation.

La performance du modèle final sera évaluée en calculant l'aire sous la courbe (AUC) de la courbe caractéristique de fonctionnement du récepteur (ROC). Cette métrique aidera à quantifier la capacité du modèle à distinguer entre les différents niveaux de CA. Enfin, l'interprétation se concentrera sur la variable la plus significative, explorant son effet sur le CA pour fournir des insights pertinents sur les dynamiques du marché.

Les analyses effectuées permettront de tirer des conclusions éclairées sur la relation

entre le volume de livraison dans les ports et le chiffre d'affaires, offrant ainsi une base solide pour les décisions commerciales et logistiques futures.

3.4.1 Présentation de la Variable Chiffre d'Affaires (CA)

Cette section présente une analyse initiale de la variable CA, qui est le chiffre d'affaires des ports, et explore la corrélation entre le CA et les quantités livrées par les ports.

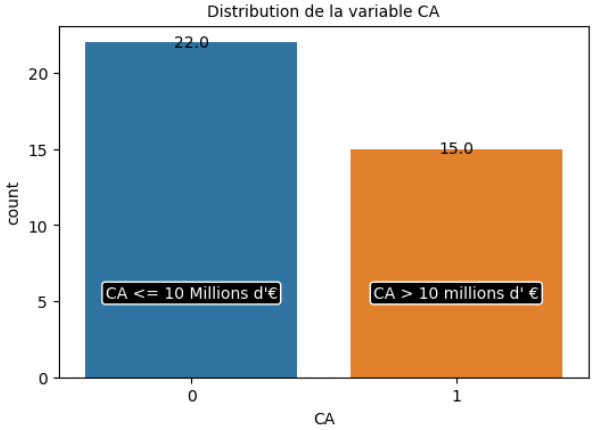

FIGURE 3.54 – Distribution de la variable CA, divisée en deux catégories basées sur un seuil de 10 millions d'euros.

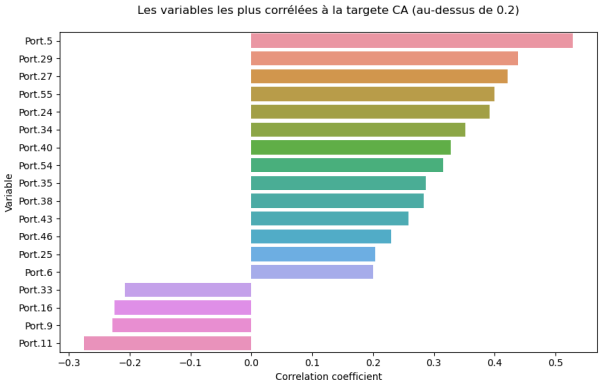

FIGURE 3.55 – Corrélation entre les quantités livrées par les ports et le CA, montrant les ports avec une corrélation supérieure à 0.2.

La Figure 3.54 montre que la majorité des ports ont un CA inférieur ou égal à 10 millions d'euros. Moins de ports ont un CA supérieur à ce seuil, indiquant une distribution inégale des niveaux de chiffre d'affaires.

Dans la Figure 3.55, les ports sont classés par l'intensité de leur corrélation avec le CA. Des valeurs de corrélation plus élevées suggèrent une relation plus forte entre les quantités livrées et le chiffre d'affaires généré. Par exemple, 'Port.5' montre la corrélation la plus forte, indiquant un lien significatif entre ses livraisons et le CA généré.

Ces analyses fournissent une compréhension de base pour la modélisation future du CA, en identifiant quels ports ont le plus grand impact et méritent donc une attention particulière dans les modèles prédictifs.

3.4.2 La Régression Logistique Ridge

Dans le cadre de la régression logistique Ridge, le choix du paramètre de régularisation λ est crucial. Le graphique suivant montre l'évolution de la déviance binomiale en fonction de différents niveaux de $\log(\lambda)$.

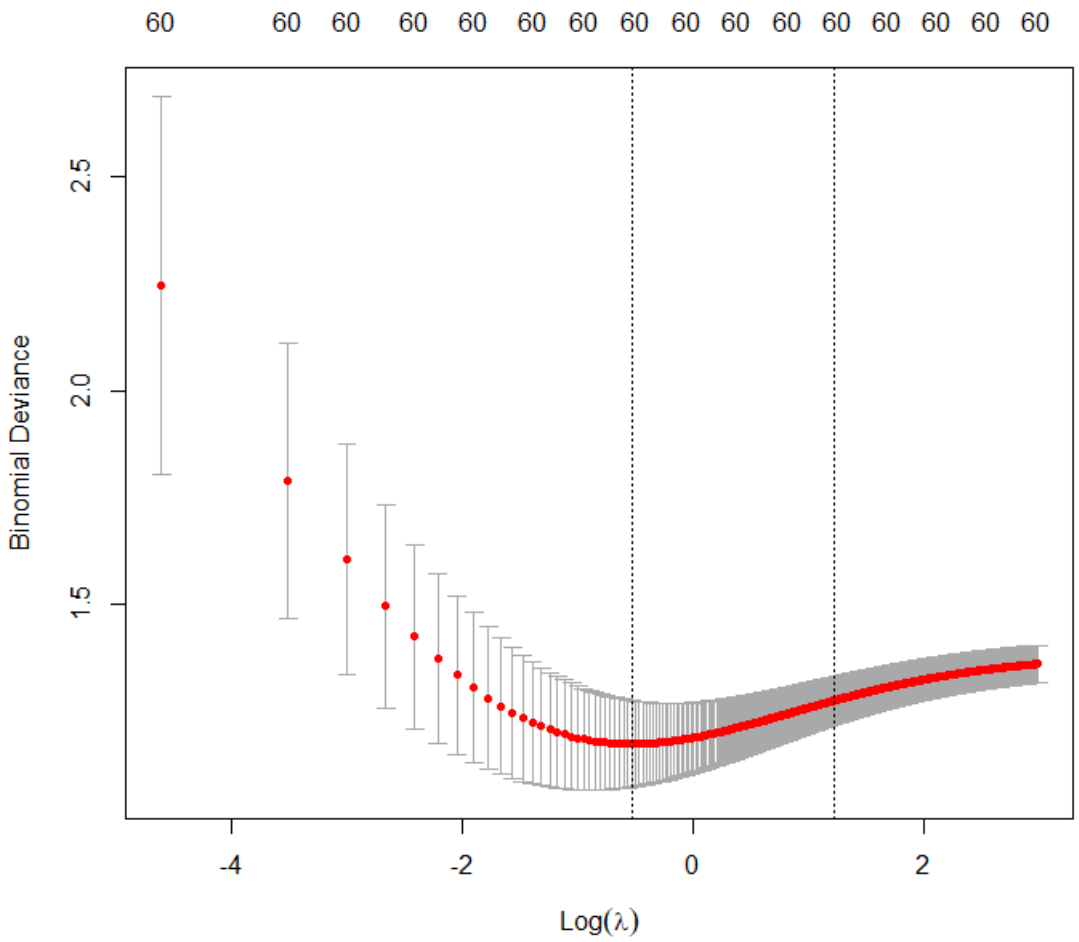

FIGURE 3.56 – Optimisation de λ pour la régression logistique Ridge. Les points rouges représentent la déviance binomiale moyenne pour chaque valeur de λ, avec l'intervalle de confiance représenté par les barres verticales.

La déviance binomiale mesure l'ajustement du modèle ; des valeurs plus basses indiquent un meilleur ajustement. Les points rouges illustrent la déviance moyenne obtenue par validation croisée pour chaque valeur $\log(\lambda)$, et les barres indiquent l'incertitude associée à cette estimation. La déviance diminue et atteint un plateau, ce qui suggère qu'au-delà d'un certain point, augmenter la pénalité de régularisation n'améliore pas significativement l'ajustement du modèle. La sélection de λ se fait en cherchant le point le

plus à gauche où la déviance est minimale avant de remonter, indiquant ainsi le compromis optimal entre biais et variance.

Pour garantir la stabilité de la sélection de λ, une boucle a été utilisée. Si le λ sélectionné est à l'extrémité de l'intervalle des valeurs testées, la grille de λ est ajustée pour explorer de nouvelles valeurs, assurant ainsi que le λ optimal n'est pas en bordure de l'intervalle exploré.

L'analyse des coefficients obtenus par la régression logistique Ridge est illustrée dans le graphique ci-dessous et permet d'identifier les variables les plus influentes.

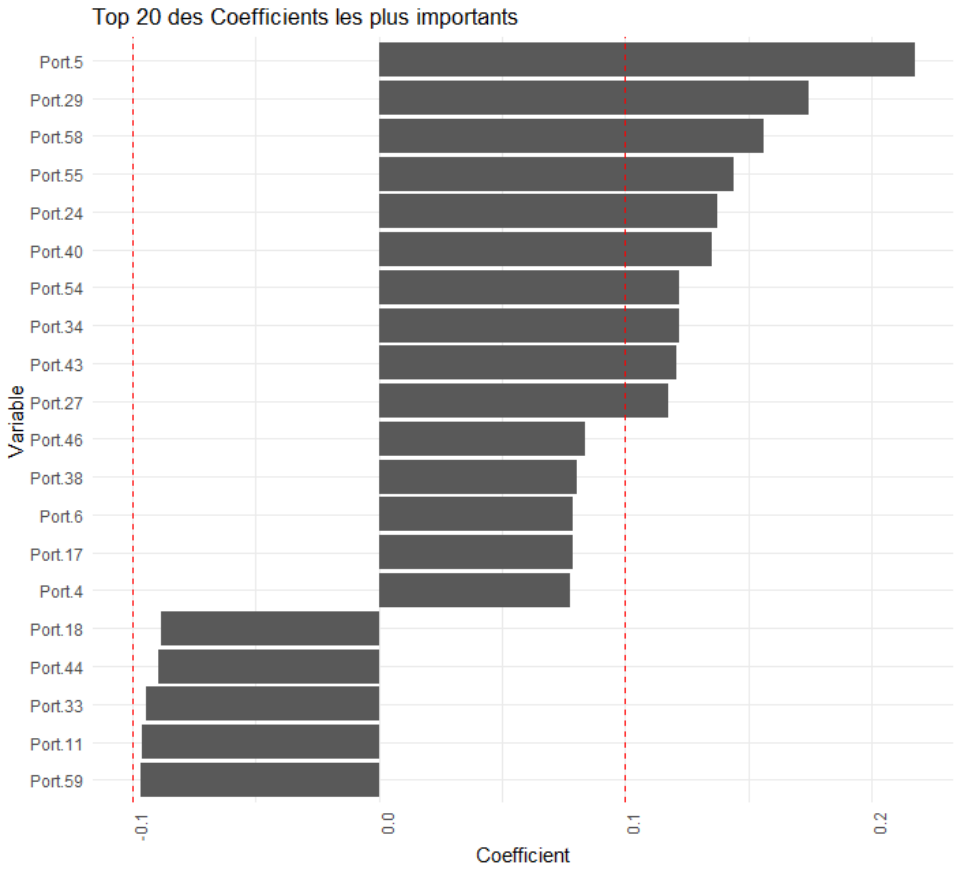

FIGURE 3.57 – Les coefficients les plus importants issus de la régression logistique Ridge.

Le graphique montre les 20 variables avec les plus grands coefficients en valeur absolue, indiquant leur importance relative dans le modèle. Une méthode de sélection visuelle est utilisée ici pour retenir les variables les plus significatives, avec un seuil prédéfini basé sur la magnitude des coefficients.

Les variables suivantes ont été retenues pour leur contribution significative, comme l'indiquent leurs coefficients respectifs :

Variable	**Coefficient**
Port.5	0.2173
Port.29	0.1745
Port.58	0.1560
Port.55	0.1442
Port.24	0.1373
Port.40	0.1352
Port.54	0.1222
Port.34	0.1219
Port.43	0.1207
Port.27	0.1177
Port.59	-0.0965
Port.11	-0.0957

TABLE 3.6 – Les 12 variables les plus importantes basées sur les coefficients de la régression Ridge.

Ces variables sélectionnées représentent les ports qui ont un impact majeur sur le modèle de prédiction du chiffre d'affaires, avec Port.5 montrant le plus grand effet positif et Port.59 et Port.11 ayant les impacts négatifs les plus notables.

Après la sélection des variables par la régression Ridge, il est important de comprendre les relations entre elles. La matrice de corrélation suivante permet d'évaluer ces interactions.

Dans la Figure 3.58, chaque cellule représente le coefficient de corrélation entre deux variables, allant de -1 à 1. Une valeur proche de 1 indique une forte corrélation positive, tandis qu'une valeur proche de -1 indique une forte corrélation négative. Des valeurs proches de zéro suggèrent qu'il n'y a pas de corrélation linéaire.

Les corrélations fortes et positives entre certains ports, comme entre 'Port.58' et 'Port.55', peuvent indiquer des comportements similaires ou des effets synergiques en

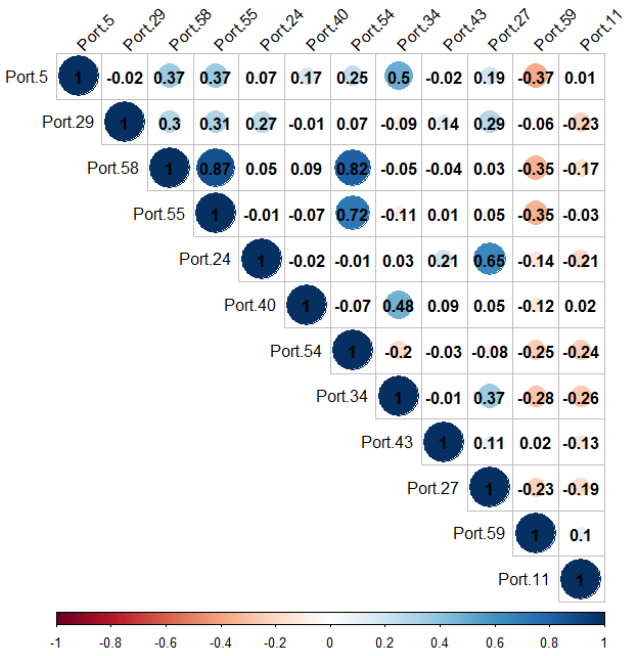

FIGURE 3.58 – Matrice de corrélation des variables sélectionnées par la régression Ridge.

termes de contribution au chiffre d'affaires. À l'inverse, une corrélation négative, comme entre 'Port.5' et 'Port.27', pourrait refléter des différences dans leurs influences sur le CA. Cette analyse est essentielle pour s'assurer que le modèle final n'inclut pas de variables redondantes et pour comprendre la dynamique entre les variables retenues.

3.4.3 La Régression Logistique Lasso

La régression logistique LASSO est une méthode de régularisation et de sélection de variables qui permet de construire un modèle prédictif en conservant uniquement les variables les plus pertinentes. Cette sous-section détaillera l'application de la régression LASSO pour identifier les variables ayant le plus grand impact sur le chiffre d'affaires.

— Une régression logistique LASSO sera mise en œuvre pour affiner davantage la sélection des variables effectuée précédemment par la régression Ridge.

— Le paramètre de régularisation λ sera choisi en utilisant la méthode de validation croisée. Les options λ_{min} et λ_{1se} seront évaluées pour déterminer un seuil approprié pour la sélection des variables.

— Après avoir déterminé le seuil optimal, les ports qui contribuent significativement au modèle seront identifiés. Ces ports seront les variables retenues pour le modèle

final.

L'accent sera mis sur l'élaboration d'un modèle à la fois précis et simple, éliminant les variables superflues tout en conservant celles qui sont essentielles pour prédire le chiffre d'affaires. L'analyse vise à fournir une compréhension claire de la relation entre les quantités livrées par les ports et le chiffre d'affaires généré.

Dans le processus de régression LASSO, le choix du paramètre de régularisation lambda est déterminant pour la qualité du modèle. Le graphique suivant montre l'évolution cumulée des résidus au fur et à mesure des itérations de la validation croisée et la valeur finale de lambda choisie, λ_{1se}.

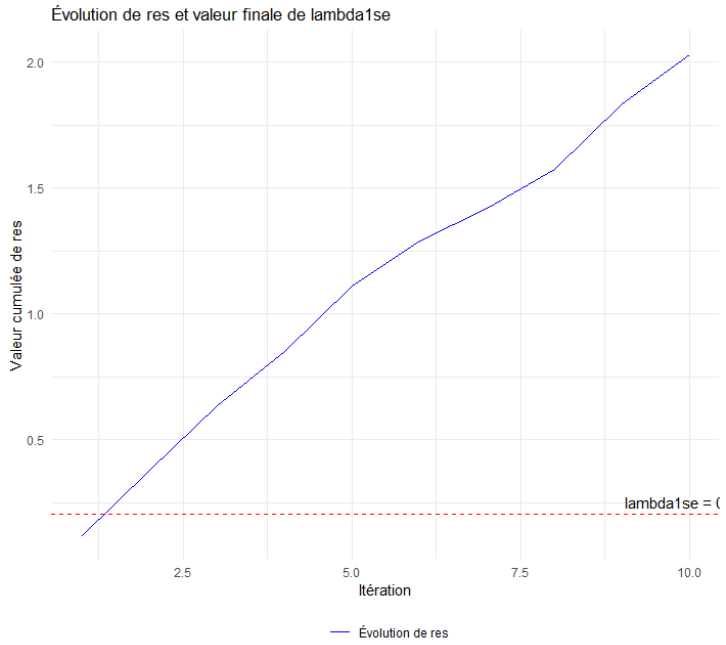

FIGURE 3.59 – Évolution cumulative des résidus et valeur finale de λ_{1se} pour la régression LASSO.

La Figure 3.59 montre que la valeur de λ_{1se} est choisie en fonction de la stabilité des résidus à travers les itérations. Cette valeur représente un compromis entre la réduction de la variance et la conservation d'un modèle suffisamment riche en informations. En optant pour λ_{1se}, le modèle tend vers plus de généralité, préférant ainsi éviter le surajustement potentiel qui pourrait survenir avec λ_{min}.

L'utilisation de λ_{1se} pour la sélection des variables garantit que seules les variables

ayant une influence significative sur le modèle soient retenues, conduisant à un modèle plus robuste et performant pour la prédiction du chiffre d'affaires.

Afin d'obtenir un modèle efficace, une régression logistique LASSO a été réalisée avec un paramètre de régularisation lambda minutieusement sélectionné. Le lambda choisi est de 0.07, préféré à λ_{1se} car il retient un nombre plus élevé de variables pertinentes.

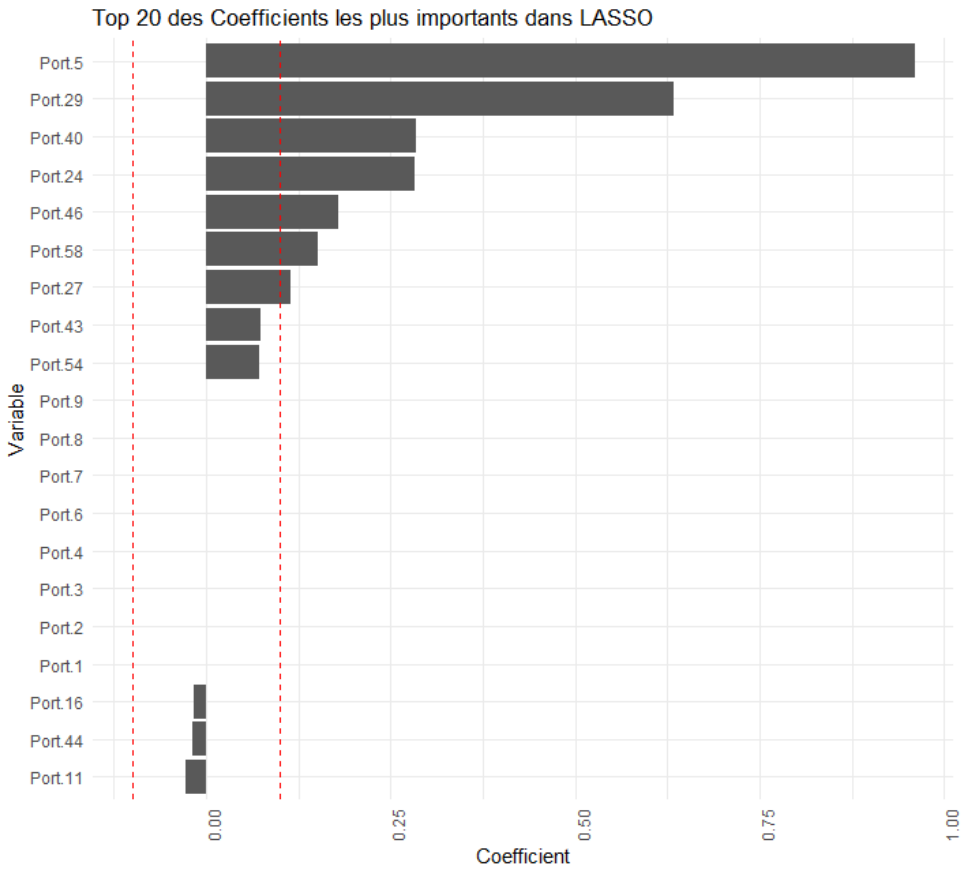

FIGURE 3.60 – Coefficients des variables dans la régression LASSO avec λ_{min}.

La Figure 3.60 montre les 20 variables les plus importantes selon le modèle LASSO, avec des coefficients non nuls. Les barres représentent l'ampleur de l'effet de chaque port sur la variable cible, avec la ligne rouge verticale indiquant la valeur de λ_{min} utilisée pour la sélection.

Les ports suivants sont retenus dans le modèle final, triés par l'importance de leur coefficient en valeur absolue :

La sélection de ces variables indique leur importance relative dans le modèle prédictif du chiffre d'affaires, avec 'Port.5' montrant le plus grand effet positif sur la variable cible.

Variable	Coefficient
Port.29	0.6336
Port.5	0.9599
Port.40	0.2854
Port.24	0.2828
Port.46	0.1801
Port.58	0.1522
Port.27	0.1153
Port.43	0.0744
Port.54	0.0730
Port.44	-0.0187
Port.16	-0.0168
Port.11	-0.0282

TABLE 3.7 – Variables retenues par la régression LASSO avec leurs coefficients respectifs.

3.4.4 Régression Logistique ElasticNet

Dans cette section, je vais appliquer la régression ElasticNet pour affiner le modèle prédictif du chiffre d'affaires. ElasticNet, en combinant les forces de la régression Ridge et LASSO, permet de sélectionner des variables de manière plus équilibrée. Je chercherai le meilleur paramètre alpha qui contrôle le compromis entre les deux techniques de régularisation. L'objectif est de sélectionner les variables les plus prédictives tout en maintenant la généralité du modèle pour éviter le surajustement.

Le graphique illustre la progression de la valeur cumulée des résidus au fil des itérations pendant le processus de validation croisée pour ElasticNet. La ligne rouge pointillée marque la valeur finale retenue pour λ_{1se}, qui est le compromis entre complexité et performance du modèle.

Cette figure montre que la sélection de λ_{1se} est basée sur la stabilisation des résidus, ce qui est une indication que le modèle ne gagnerait pas en performance en augmentant la complexité au-delà de ce point. Le choix de ce paramètre vise à obtenir un modèle qui généralise bien sur des données non vues, en évitant de surajuster les données d'appren-

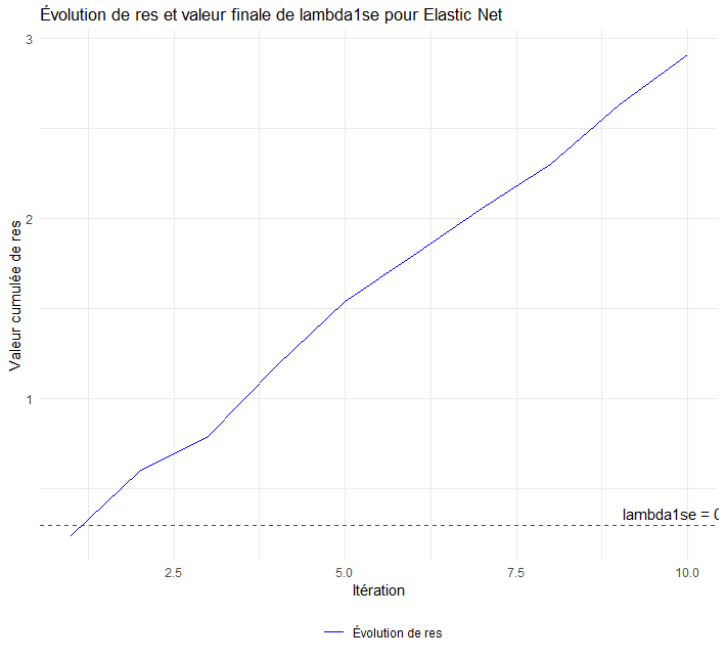

FIGURE 3.61 – Courbe de validation pour le choix de λ_{1se} dans ElasticNet.

tissage.

Le modèle ElasticNet a été appliqué pour sélectionner les variables ayant un impact significatif sur le chiffre d'affaires. Le graphique suivant illustre les coefficients des variables retenues par le modèle.

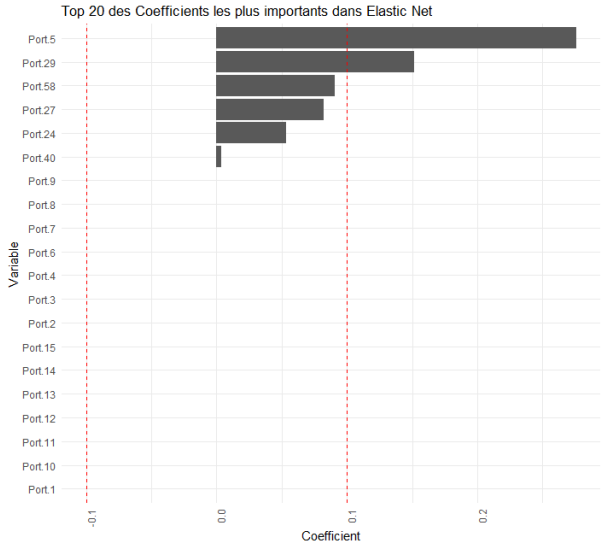

FIGURE 3.62 – Coefficients des variables sélectionnées par la régression ElasticNet.

Les barres représentent la magnitude de l'influence de chaque variable, avec les valeurs positives indiquant une augmentation et les valeurs négatives une diminution du chiffre d'affaires. Les variables sélectionnées sont celles avec les coefficients les plus éloignés de zéro.

À côté du graphique, le tableau suivant récapitule les variables et leurs coefficients correspondants :

Variable	**Coefficient**
Port.5	0.2757
Port.29	0.1517
Port.58	0.0906
Port.27	0.0825
Port.24	0.0537
Port.40	0.0034

TABLE 3.8 – Variables et coefficients sélectionnés par la régression ElasticNet.

Ces coefficients, obtenus après un processus de sélection rigoureux, mettent en lumière les ports les plus importants pour la prédiction du chiffre d'affaires dans le modèle ElasticNet. 'Port.5' se distingue comme la variable ayant le plus grand impact positif.

Suite à la sélection des variables par la régression ElasticNet, il est crucial d'examiner les corrélations entre elles pour s'assurer de l'indépendance des prédicteurs.

La Figure 3.63 présente la matrice de corrélation, avec des nuances de rouge représentant différentes intensités de corrélation positive et le bleu indiquant des corrélations négatives. Une valeur absolue de 1 signifie une corrélation parfaite, soit positive soit négative, tandis qu'une valeur proche de 0 indique peu ou pas de corrélation directe.

Une analyse attentive révèle que certaines variables, comme 'Port.58' et 'Port.29', montrent une corrélation forte, suggérant une relation linéaire entre elles. Ces informations sont essentielles pour comprendre les dynamiques complexes au sein du modèle et

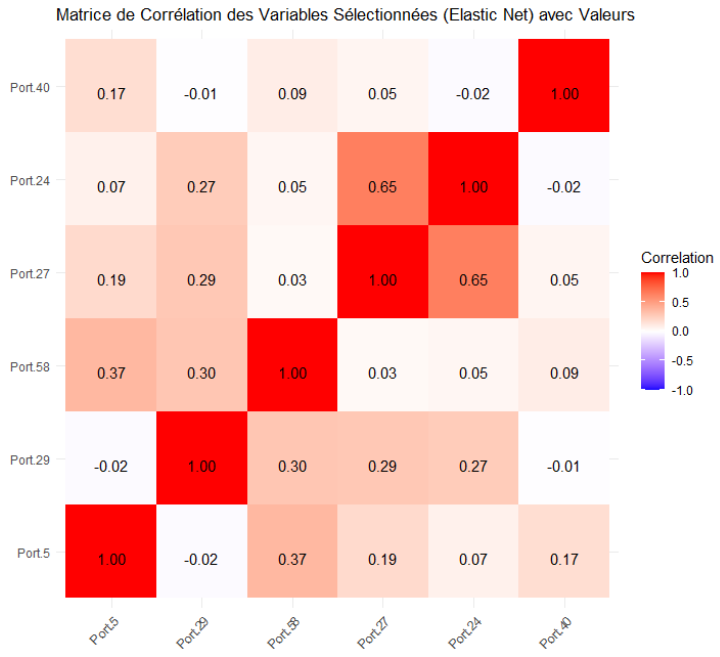

FIGURE 3.63 – Corrélations entre les variables choisies par ElasticNet.

pour envisager des ajustements si nécessaire pour minimiser la redondance et maximiser l'indépendance des variables.

Le fait de repérer ces corrélations permet d'anticiper l'impact des interactions entre variables sur la prédiction et la stabilité du modèle final. Cela garantit que les variables retenues contribuent de manière unique à la prévision du chiffre d'affaires, ce qui est un aspect fondamental de la modélisation prédictive.

3.4.5 AUC Interprétation

Le modèle final, construit à partir des variables sélectionnées par la régression ElasticNet, a été évalué en calculant l'aire sous la courbe ROC (AUC). La courbe ROC et l'AUC sont des indicateurs clés de la performance du modèle.

Comme le montre la Figure 3.64, l'AUC de 0.97 indique une excellente capacité du modèle à distinguer entre les différentes classes de chiffre d'affaires. Cela suggère que le modèle est très performant et fiable.

En examinant les coefficients, 'Port.5' apparaît comme la variable la plus significative avec le plus grand coefficient positif, indiquant une forte influence positive sur la

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4580     1.0116  -1.441   0.1495
Port.5        4.0086     1.7959   2.232   0.0256 *
Port.29       2.4264     1.1069   2.192   0.0284 *
Port.24       2.0235     1.0586   1.911   0.0560 .
Port.40       1.1695     0.7681   1.523   0.1279
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49.961  on 36  degrees of freedom
Residual deviance: 15.440  on 32  degrees of freedom
AIC: 25.44

Number of Fisher Scoring iterations: 8
```

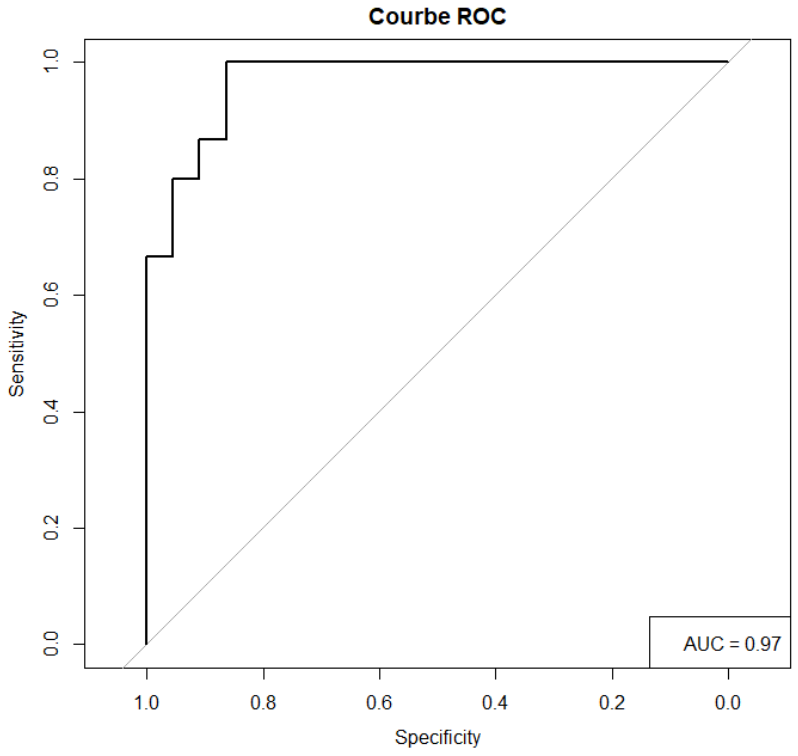

FIGURE 3.64 – À gauche : Coefficients du modèle ElasticNet. À droite : Courbe ROC avec AUC pour le modèle final.

probabilité d'augmenter le chiffre d'affaires. Cette variable est donc considérée comme la plus prédictive et pourrait être un point d'intérêt pour des stratégies d'optimisation commerciale.

La performance élevée du modèle et l'influence notable de 'Port.5' soulignent l'efficacité de la sélection de variables via ElasticNet et la puissance prédictive du modèle développé.

3.5 Modélisation Polytomique de la Variable CA3

Dans cette sous-section, je vais entreprendre la modélisation de la variable CA3, qui représente le chiffre d'affaires réparti en catégories plus détaillées. L'approche choisie est une régression multinomiale pénalisée, qui ne prend pas en compte l'ordre potentiel entre les catégories. Cette méthode permettra d'identifier les variables influençant les différentes classes de CA3 sans présupposer une relation ordonnée entre elles.

Après avoir sélectionné les variables pertinentes à l'aide de la régression multinomiale pénalisée, j'explorerai également une régression polytomique ordonnée. Cela permettra de comparer si la prise en compte de l'ordre naturel des catégories de chiffre d'affaires peut améliorer la performance du modèle.

Ces analyses fourniront des insights sur la structure du chiffre d'affaires et sur la pertinence de traiter CA3 comme une variable ordonnée ou non ordonnée dans le contexte de la modélisation prédictive.

3.5.1 Présentation de la Variable CA3

La variable CA3 catégorise le chiffre d'affaires en trois classes distinctes, permettant une analyse plus détaillée de sa distribution. Le premier graphique montre la répartition des données de chiffre d'affaires dans ces trois catégories, tandis que le second graphique met en évidence les variables les plus corrélées à CA3.

Dans la Figure 3.65 à gauche, nous observons que la distribution des classes de chiffre d'affaires est relativement équilibrée, ce qui offre une bonne base pour la modélisation polytomique. Le deuxième graphique, à droite, illustre la force de la corrélation entre chaque

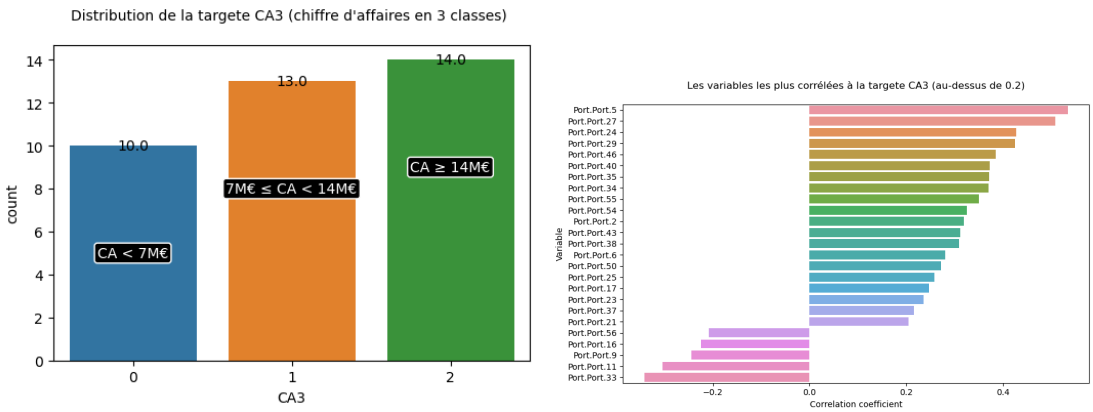

FIGURE 3.65 – À gauche : Distribution de CA3. À droite : Variables corrélées avec CA3.

variable et CA3, avec des coefficients de corrélation supérieurs à 0.2 considérés comme significatifs. Ces variables pourront être des candidats clés pour la régression multinomiale pénalisée et pourront influencer de manière importante le modèle de prévision de CA3.

L'analyse de ces graphiques fournit une compréhension initiale de la structure des données et guide la sélection des variables pour la modélisation subséquente.

3.5.2 Régression Multinomiale avec Elastic Net

L'application d'une régression multinomiale pénalisée Elastic Net sur la variable CA3 a permis de sélectionner les variables influençant les différentes catégories de chiffre d'affaires sans considérer l'aspect ordonné. Le graphique suivant illustre les coefficients pour les classes résultantes de la modélisation.

La Figure 3.66 présente les coefficients des variables pour deux classes distinctes. Il est notable que la classe 2 n'est pas représentée, ce qui peut résulter d'un manque de différenciation significative de cette classe par rapport aux variables sélectionnées, ou que les variables choisies ne distinguent pas efficacement cette classe dans l'espace des caractéristiques actuel.

Les coefficients indiquent l'influence de chaque variable sur la probabilité d'appartenance à chaque classe, avec 'Port.Port.5' ayant le coefficient le plus élevé pour la classe 1, suggérant une forte influence positive sur la probabilité d'appartenir à cette catégorie de CA3.

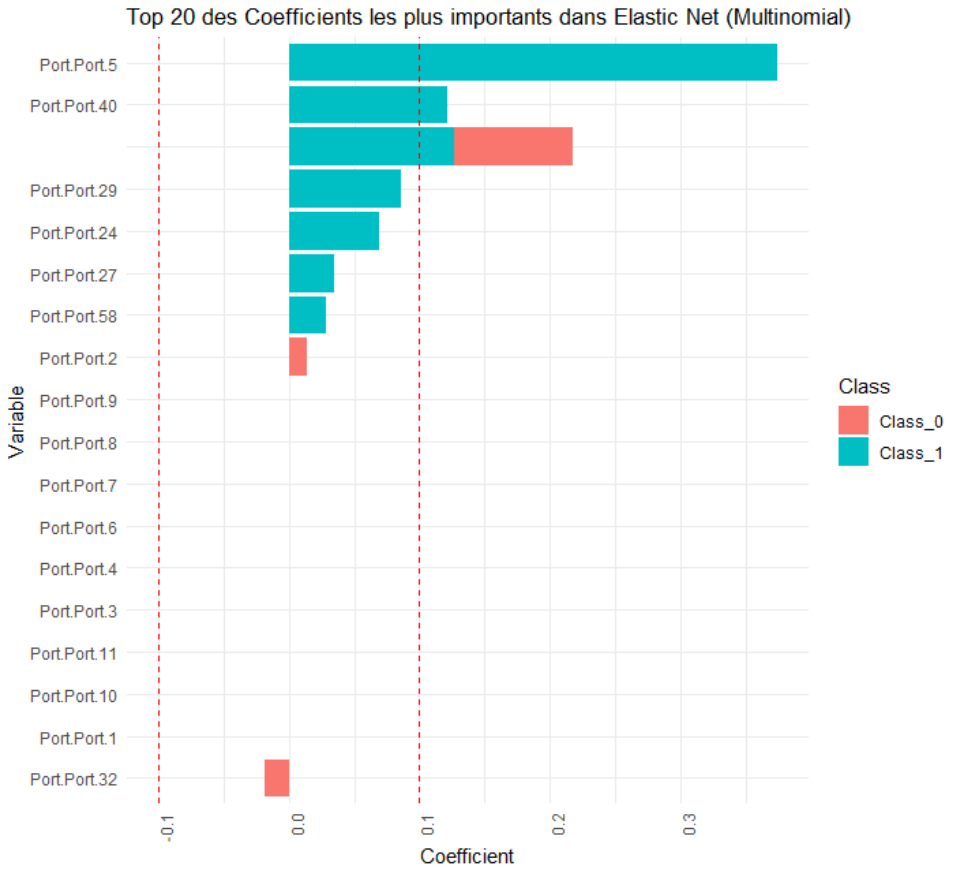

FIGURE 3.66 – Coefficients les plus importants dans la régression multinomiale Elastic Net.

Classe	**Variable**	**Coefficient**
Class_1	Port.Port.5	0.3738
Class_1	Port.Port.40	0.1210
Class_1	Port.Port.29	0.0855
Class_1	Port.Port.24	0.0692
Class_1	Port.Port.27	0.0348
Class_1	Port.Port.58	0.0282
Class_0	Port.Port.32	-0.0180
Class_0	Port.Port.2	0.0144

TABLE 3.9 – Variables et coefficients sélectionnés par la régression multinomiale Elastic Net.

Cette analyse des coefficients et la performance du modèle fournissent des informations essentielles pour la compréhension des facteurs influençant les différentes gammes de chiffre d'affaires et pourront guider des décisions stratégiques commerciales.

Interprétation de la Régression Polytomique Ordonnée

Après avoir sélectionné les variables pertinentes par une régression multinomiale pénalisée, j'ai procédé à une régression polytomique ordonnée. Cette méthode est appropriée lorsque les catégories de la variable dépendante ont un ordre naturel, mais qu'on ne suppose pas cet ordre dans le modèle de base.

Résultats du modèle Les coefficients obtenus du modèle sont présentés ci-dessous, montrant l'impact estimé de chaque variable sur les probabilités de chaque catégorie de CA3.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
Port.Port.5    2.2279     1.0421   2.138  0.03253 *
Port.Port.40   1.7015     0.6107   2.786  0.00533 **
Port.Port.29   1.2878     0.6977   1.846  0.06492 .
Port.Port.24   1.0974     0.6903   1.590  0.11192
Port.Port.27   1.1144     0.8530   1.306  0.19140
Port.Port.58   0.8691     0.6915   1.257  0.20881
Port.Port.32  -1.2091     0.7523  -1.607  0.10802
Port.Port.2    1.3488     0.7160   1.884  0.05961 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
0|1  -2.9272     0.9337  -3.135
1|2   1.7159     0.7625   2.250
```

FIGURE 3.67 – Coefficients estimés par la régression polytomique ordonnée.

Le modèle a identifié 'Port.Port.5' et 'Port.Port.40' comme ayant les plus grands coefficients positifs, ce qui suggère qu'ils sont des prédicteurs significatifs de la probabilité d'appartenir à une catégorie supérieure de chiffre d'affaires. Cependant, l'absence de la classe 2 parmi les coefficients peut indiquer que les variables sélectionnées ne sont pas aussi discriminantes pour cette classe ou que la classe 2 est peut-être sous-représentée dans l'échantillon.

Variable	Estimate	Std. Error	z value	Pr(>|z|)
Port.Port.5	2.2279	1.0421	2.138	0.0325*
Port.Port.2	1.3488	0.7160	1.884	0.0596

TABLE 3.10 – Coefficients estimés de la régression polytomique ordonnée.

Les seuils de classe indiquent les points de transition entre les catégories. Avec un premier seuil négatif et un deuxième seuil positif, cela suggère que les probabilités basculent de la classe 0 vers la classe 1, puis vers la classe 2, à mesure que les valeurs des prédicteurs augmentent.

Cette interprétation des coefficients et des seuils fournit des insights sur la manière dont les différentes catégories de chiffre d'affaires sont influencées par les variables prédictives, offrant ainsi une base pour des décisions stratégiques informées.

Chapitre 4

Discussion

Cette étude a employé des méthodes statistiques avancées pour analyser les données issues de 37 entreprises de transport maritime, en se concentrant sur les liens entre les émissions de CO2, les chiffres d'affaires, les retards dus à des incidents, et les quantités transportées dans les ports. L'utilisation des régressions pénalisées Ridge, LASSO, et ElasticNet a révélé des insights précieux, notamment en identifiant les ports et autres facteurs influençant significativement ces variables.

La régression sur composantes principales et la régression PLS ont également joué un rôle clé, en fournissant une perspective alternative et en réduisant la complexité des données. La comparaison de ces méthodes a permis d'évaluer leurs avantages et limites respectifs dans le contexte des données spécifiques de cette étude.

L'analyse du nombre d'incidents a mis en évidence l'utilité des régressions de Poisson Ridge et LASSO pour identifier les ports contribuant le plus à ces incidents. La découverte de la surdispersion dans ces données a conduit à l'adoption d'une approche binomiale négative, plus adaptée à cette situation.

L'étude du chiffre d'affaires a bénéficié d'une approche similaire, avec l'ajout de modèles polytomiques ordonnés pour une compréhension plus nuancée des catégories de revenus. La régression multinomiale a fourni des informations supplémentaires, soulignant l'importance de considérer les spécificités de chaque classe de chiffre d'affaires.

La diversité des méthodes statistiques utilisées dans cette étude démontre la richesse des approches disponibles pour l'analyse de données complexes. Elle souligne également la nécessité d'une sélection soignée des techniques, en fonction des particularités des données et des objectifs de recherche.

Chapitre 5

Conclusion

Cette étude a fourni une analyse approfondie des données de 37 entreprises de transport maritime, en se focalisant sur les relations entre les émissions de CO2, les incidents, et les chiffres d'affaires. L'utilisation de méthodes de régression pénalisée a permis de distinguer les variables les plus significatives, contribuant à une meilleure compréhension des facteurs impactant ces variables clés.

Les techniques de régression sur composantes principales et PLS ont offert des perspectives enrichissantes, permettant de traiter efficacement la multicollinéarité et la complexité des données. La régression de Poisson, adaptée aux données de comptage avec surdispersion, a révélé des facteurs influençant les incidents, tandis que la modélisation polytomique a permis d'examiner le chiffre d'affaires sous différents angles.

Ces analyses ont révélé l'importance cruciale de certains ports dans l'émission de CO2, la génération de revenus, et la survenue d'incidents. Ces informations sont précieuses pour les entreprises de transport maritime, offrant des pistes pour optimiser les opérations, réduire les émissions, et maximiser les profits.

Cette étude illustre l'importance d'approches analytiques diversifiées et rigoureuses dans l'examen des données complexes. Les résultats obtenus peuvent servir de base pour des décisions stratégiques éclairées dans le secteur du transport maritime, contribuant ainsi à des opérations plus efficaces et durables.

Chapitre 6

Références

1. Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer Series in Statistics.
2. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning : with Applications in R*. Springer Texts in Statistics.
3. Kuhn, M., Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
4. Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer.

Remerciements

Je tiens à exprimer ma profonde gratitude à mon professeur, Denys POMMERET. Votre expertise et votre passion pour l'enseignement ont été une source d'inspiration constante.

Gaoussou Diakite

15 décembre 2023