Mastère Spécialisé Data Science pour la connaissance client

REGRESSIONS PENALISEES

Auteurs:
ADUAYOM MESSAN Messan Daniel
SADIO Ndeye Salimata

Professeur:
Mr. Denys POMMERET

Contents

1 Introduction

Le transport maritime, dominant les échanges mondiaux à hauteur d'environ 80% confère aux ports une position prépondérante en tant que carrefours du commerce, ancrés au cœur de la dynamique économique. Toutefois, les répercussions environnementales, les performances économiques et logistiques revêtent une importance cruciale, nécessitant une attention soutenue.

Au sommet des préoccupations mondiales, la problématique urgente du changement climatique et des émissions de gaz à effet de serre, imputables aux activités humaines, accentue la responsabilité du secteur des transports. Ce dernier, du fait de sa dépendance à la technologie du moteur à combustion interne, émerge comme l'un des principaux contributeurs aux émissions de gaz à effet de serre, notamment le CO2 provenant de la combustion des carburants fossiles. À l'échelle mondiale, les transports représentaient 25% des émissions totales de CO2 en 2003, avec une augmentation significative de 31% entre 1990 et 2003 (OECD, 2007). Cette considération environnementale majeure positionne les activités portuaires au centre des préoccupations liées à la durabilité, incitant à des initiatives visant à atténuer l'impact écologique du transport maritime.

L'interconnexion croissante entre la conteneurisation, moteur du commerce international, et l'économie mondiale dans le contexte de la globalisation souligne l'importance cruciale des ports maritimes dans la logistique mondiale. L'étude systématique de l'offre de transport des plus grands armements conteneurisés mondiaux offre une vision approfondie de la circulation maritime du trafic conteneurisé, dévoilant les caractéristiques spatiales essentielles de cette activité (Frémont et Soppé). Ainsi, l'efficacité économique des ports est intrinsèquement liée à leur rôle dans le commerce international et à la manière dont ils s'intègrent dans les flux mondiaux.

Par ailleurs, la performance logistique et le chiffre d'affaires des entreprises sont étroitement liés à la gestion des ports, qui jouent un rôle essentiel en tant que points de passage clés pour les marchandises. Dans le cadre de cette étude approfondie sur les performances environnementales, économiques et logistiques des ports maritimes, nous explorerons ces trois dimensions clés : les émissions de CO2, le chiffre d'affaires généré, et les retards dans le transport maritime. Notre approche analytique sera centrée sur l'utilisation de techniques de régression, en particulier des régressions pénalisées, pour modéliser ces relations complexes.

En particulier, notre investigation explorera l'existence de relations entre :

- Les émissions de CO2 et les quantités de marchandises transportées, dans le but d'apprécier l'impact écologique.
- Le chiffre d'affaires des entreprises et les quantités transportées, dans le but de saisir la manière dont les résultats financiers sont liés à l'efficacité opérationnelle.
- Le nombre de retards imputables aux incidents portuaires et les volumes transportés, dans le but de distinguer l'efficacité et la fiabilité des opérations de transport maritime.

L'objectif du rapport est donc, au moyen d'une analyse méthodique, de présenter des éclairages pertinents susceptibles de guider les stratégies en matière de durabilité, d'optimiser l'efficacité opérationnelle et de contribuer à l'élaboration de politiques de gestion pour les entreprises de transport maritime et les responsables de ports. Les informations que nous fournirons seront potentiellement utiles aux parties prenantes du secteur maritime, aux décideurs politiques, ainsi qu'à la communauté académique intéressée par ce sujet et son impact sur le développement durable.

2 Méthodologie

2.1 Présentation de la donnée

Dans le cadre de notre étude, nous disposons d'une base de données complète provenant de mesures effectuées auprès de 37 entreprises majeures du secteur du transport maritime. Cette base de données englobe des paramètres cruciaux reflétant divers aspects environnementaux, économiques et logistiques.

Les principales variables sont les suivantes :

- **CO2 :** Émissions de CO2 en milliers de tonnes par entreprise sur une année.
- **CA (Chiffre d'Affaires) :** Classifié 1 si le chiffre d'affaires est supérieur à 10 millions d'euros, 0 sinon.
- **Incid (Incidents) :** Nombre de retards imputés au port.

En parallèle, nous disposons de données sur les quantités de marchandises transportées (en millions de tonnes) dans 60 principaux ports. Cette base de données exhaustive constitue la pierre angulaire de notre analyse, permettant une exploration approfondie des relations entre les émissions de CO2, le chiffre d'affaires, les retards et les quantités de marchandises transportées.

2.2 Méthodes d'analyse: Cadre non standard

Avant d'entamer notre étude, il est impératif de déterminer le cadre dans lequel nous évoluons. Deux possibilités se présentent à nous:

- Un cadre standard: caractérisé par une matrice $X'X$ inversible, permet généralement d'assurer la qualité et la stabilité des estimations dans le contexte de l'analyse statistique.
- Un cadre non standard : cas où il est souvent nécessaire d'appliquer des techniques adaptées aux particularités du problème étudié, telles que des méthodes robustes, des modèles non linéaires, ou des méthodes de bootstrap, entre autres. Ainsi, une première inspection de notre jeu de données révèle que le nombre de variables ($p = 60$) excède le nombre d'individus ($n = 37$), soit $p > n$. Ainsi, la matrice $X'X$, conçue comme une matrice carrée $p \times p$, présente un rang inférieur à $\max(p, n)$, ce qui induit potentiellement une instabilité ou une mauvaise condition de la matrice.

Pour approfondir notre compréhension de la stabilité de la matrice, nous avons choisi d'explorer les valeurs propres de $X'X$. L'analyse de ces valeurs révèle une disparité significative entre les plus grandes et les plus petites. Plus précisément, les premières valeurs propres sont nettement supérieures aux dernières, ces dernières tendant vers zéro. Cette disparité montre la présence potentielle de colinéarité ou de multicollinéarité parmi les variables explicatives de la matrice X.

Une observation intrigante est la proximité de certaines valeurs propres à zéro, indiquant une matrice proche de la singularité. En pratique, la représentation exacte de zéro étant souvent impossible pour la plupart des logiciels numériques, ces valeurs très petites (de l'ordre de 10^{-12} ou moins) peuvent être interprétées comme des indicateurs de quasi-singularité.

En règle générale, la proximité de zéro pour une valeur propre suggère une matrice mal conditionnée. Dans notre cas, la dernière valeur propre est très proche de zéro (1.456791×10^{-11}), laissant entrevoir la possibilité que la matrice $X'X$ soit singulière ou presque singulière.

En conclusion, la matrice $X'X$ est potentiellement mal conditionnée ou singulière en raison des valeurs propres proches de zéro. Dans le contexte de la modélisation statistique, cela pourrait engendrer des problèmes lors de l'inversion de la matrice, conduisant à des estimations peu fiables ou instables. Il serait donc prudent d'examiner de plus près la nature de la colinéarité dans nos données et d'ajuster notre approche statistique en conséquence, envisageant éventuellement l'utilisation de méthodes de régularisation ou l'élimination de variables fortement corrélées.

Pour orienter notre analyse de manière judicieuse, nous optons donc pour l'emploi de méthodes de régressions pénalisées. Ces approches se révèlent particulièrement adaptées pour mieux appréhender les relations entre les variables dans notre jeu de données. En effet, elles nous offriront la possibilité de :

- Sélectionner de manière efficiente les variables significatives,
- Gérer efficacement la multicolinéarité présente dans notre ensemble de variables,
- Prévenir tout phénomène de surajustement, améliorant ainsi la généralisation des modèles.

Nous avons fait le choix d'explorer trois méthodes spécifiques, à savoir LASSO, RIDGE, et Elastic Net, afin de tirer parti de leurs avantages respectifs dans la modélisation de notre problème. Ces méthodes, en introduisant des termes de pénalité, permettront ainsi d'affiner nos modèles tout en répondant aux défis inhérents à la dimensionnalité de nos données et aux caractéristiques observées dans la matrice $X'X$.

3 Analyse Descriptive

3.1 Tableau récapitulatif sur CO2, CA et Incid

Table 1: Tableau des statistiques

CO2	Incid	CA	CA3
Min. : 4.970	Min. : 3.0	Min. :0.0000	Min. :0.000
1st Qu.: 6.710	1st Qu.: 13.0	1st Qu.:0.0000	1st Qu.:0.000
Median : 7.410	Median : 27.0	Median :0.0000	Median :1.000
Mean : 7.465	Mean : 34.7	Mean :0.4054	Mean :1.108
3rd Qu.: 8.290	3rd Qu.: 42.0	3rd Qu.:1.0000	3rd Qu.:2.000
Max. :10.550	Max. :152.0	Max. :1.0000	Max. :2.000

L'analyse des statistiques descriptives du tableau 1 offre des insights significatifs sur les caractéristiques de notre ensemble de données. En ce qui concerne les émissions de CO2, la moyenne d'environ 7.465 milliers de tonnes sert de mesure centrale, avec une médiane de 7.410 milliers de tonnes. L'étendue des émissions varie de 4.970 à 10.550 milliers de tonnes, suggérant une certaine variabilité.

Pour le nombre d'incidents imputés au port (Incid), la moyenne de 34.7 incidents indique une fréquence moyenne relativement élevée, avec une médiane de 27 incidents. La distribution semble étirée vers des valeurs plus élevées, comme le suggère l'écart entre la moyenne et la médiane. Le nombre d'incidents varie de 3 à 152, soulignant la diversité des situations.

En ce qui concerne le chiffre d'affaires (CA), environ 40.54% des années ont un chiffre d'affaires supérieur à 10 millions d'euros, selon la moyenne. La médiane de 0 indique que la moitié des années ont un chiffre d'affaires inférieur à ce seuil. La distribution semble bimodale, concentrée à la fois sur 0 et 1, ce qui peut refléter une dichotomie dans les chiffres d'affaires, avec certaines années en dessous et d'autres au-dessus de 10 millions d'euros.

En résumé, ces statistiques descriptives fournissent un aperçu des tendances centrales et de la dispersion des émissions de CO2, du nombre d'incidents, et du chiffre d'affaires. Ces informations sont cruciales pour comprendre la dynamique de notre ensemble de données et orienter des analyses plus approfondies.

3.2 Matrice de corrélation

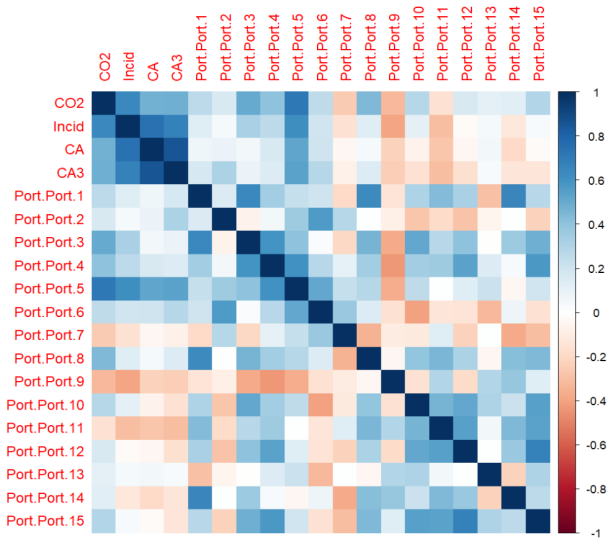
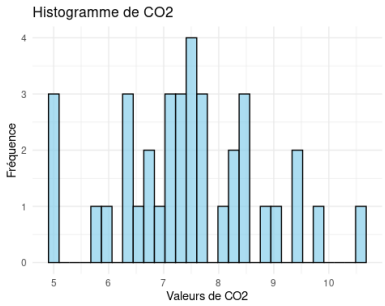


Figure 1: Matrice de corrélation

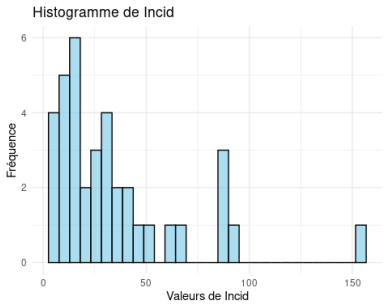
La matrice ci-dessous nous permet de voir que le nombre de retards et le chiffre d'affaires sont très corréllés. On note également que certains ports émettent plus de CO2 que d'autres, accusent un nombre de retards plus élevés que dans d'autres également.

3.3 Distributions

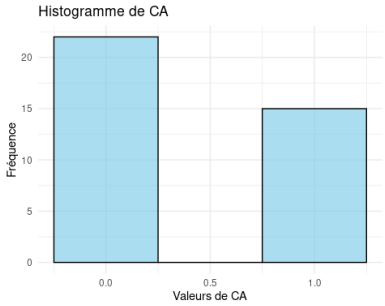
Avant d'entamer notre analyse approfondie, il est judicieux d'effectuer des analyses plus avancées sur nos variables, en particulier en ce qui concerne leur distribution. Cela nous permettra d'obtenir des insights approfondis sur le comportement des entreprises incluses dans notre étude. Nous pourrons éventuellement déterminer si des différences significatives existent entre les entreprises ou si une hétérogénéité notable se dégage de notre ensemble de données.



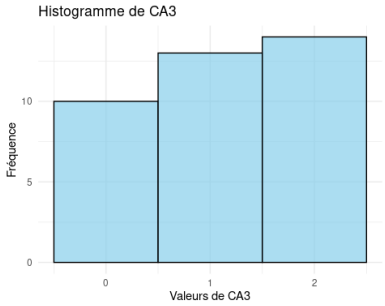
(a) Emission du CO2



(b) Nombres d'incidents



(c) Chiffres d'affaires



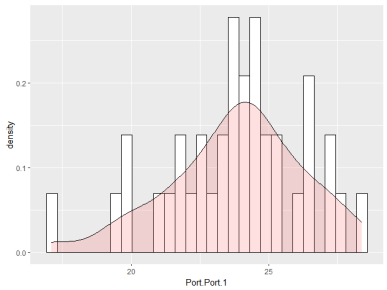
(d) Chiffres d'affaires3

Figure 2: Distribution et histogramme

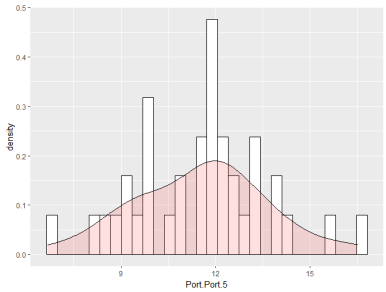
Nous observons une concentration significative des émissions de CO2 dans l'intervalle de 6 à 8 milliers de tonnes par an, indiquant une tendance commune parmi les entreprises. Cependant, quelques-unes parviennent à dépasser cet intervalle, enregistrant des émissions allant jusqu'à 11 milliers de tonnes. D'un autre côté, il est notable que peu d'entreprises enregistrent des émissions relativement faibles.

En ce qui concerne les incidents, leur fréquence est généralement basse, avec la majorité des observations se situant du côté gauche de l'histogramme, indiquant que la plupart des entreprises ont un faible nombre d'incidents, tournant autour de 0 à 25 fois. Bien que cela reste significatif, quelques entreprises enregistrent parfois jusqu'à 100 accidents, soulignant des cas exceptionnels.

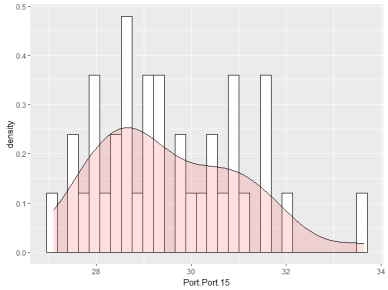
Pour les variables qualitatives, nous observons une certaine homogénéité des catégories. Pour la variable CA, la modalité 0 se démarque, indiquant qu'il y a davantage d'entreprises avec un chiffre d'affaires inférieur à 7 millions. En ce qui concerne la variable CA3, les groupes enregistrent à peu près le même nombre d'entreprises, mais on observe que celles avec un chiffre d'affaires dépassant 14 millions sont plus nombreuses. Une analyse plus approfondie à un niveau de granularité plus fin pourrait fournir des insights différents sur la situation financière des entreprises.



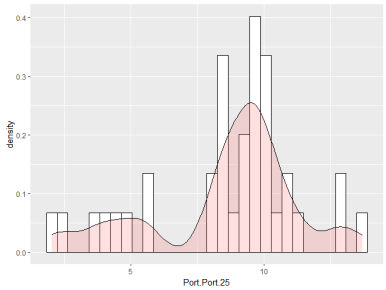
(a) Port 1



(b) Port 5



(c) Port 15



(d) Port 25

Figure 3: Distribution et histogramme des ports

Dans le cadre de notre étude, notre attention s'est portée sur les variables explicatives, et face à la complexité d'un ensemble volumineux de variables, nous avons décidé de restreindre notre analyse à quelques ports spécifiques. Il est important de noter que la quantité transportée peut varier considérablement d'un port à l'autre.

L'observation des distributions de quantités transportées pour différents ports révèle des comportements distincts. En particulier, les ports 1 et 5 affichent des distributions qui présentent des similitudes avec une distribution de loi normale, suggérant une certaine régularité dans la répartition des quantités transportées. En revanche, la distribution du port 15 semble s'écarter davantage d'une distribution normale, tout comme celle des deux autres ports.

Cette constatation met en lumière la diversité des distributions de quantités transportées entre les ports étudiés. Certains ports semblent suivre une distribution plus symétrique et centrée, tandis que d'autres présentent des variations plus significatives. Cette disparité peut avoir des implications importantes dans notre compréhension des dynamiques de transport dans ces ports spécifiques.

En résumé, notre analyse souligne que les ports peuvent exhiber des distributions de quantités transportées assez différentes les unes des autres. Cette diversité peut être influencée par divers facteurs tels que la taille du port, la nature des marchandises transportées, ou d'autres variables pertinentes. Il serait intéressant d'approfondir cette observation pour mieux appréhender les spécificités de chaque port et les motifs sous-jacents à ces variations.

4 Analyse du lien entre les variables

Tout d'abord, nous allons centrer et réduire toutes les variables explicatives. Ainsi, nous pourons rendre les résultats interprétables, assurer une pénalisation équitable, améliorer la convergence des algorithmes d'optimisation et faciliter la comparaison entre les coefficients de régression Ridge, Lasso ou ElasticNet.

4.1 Modélisation du CO2

Dans cette section de notre projet, nous visons à modéliser l'émission de dioxyde de carbone (CO_2) en fonction des quantités transportées dans divers ports. L'objectif est d'explorer plusieurs approches de modélisation statistique pour mieux comprendre les relations entre les émissions de CO_2 et les activités portuaires. Nous utiliserons différentes techniques de régression, notamment la régression linéaire Ridge, LASSO, ElasticNet, un modèle linéaire classique, la régression sur composantes principales (PCA), et la régression PLS (Partial Least Squares).

4.1.1 Régression Linéaire Ridge:

Nous avons amorcé notre analyse par le modèle de régression linéaire Ridge.

De prime abord, nous allons faire une validation croisée afin de déterminer le paramètre de régularisation optimal (λ). Cette approche revêt une importance capitale dans le processus de modélisation, notamment lorsqu'il s'agit de prévenir le surajustement et d'assurer la robustesse du modèle face à des données variées.

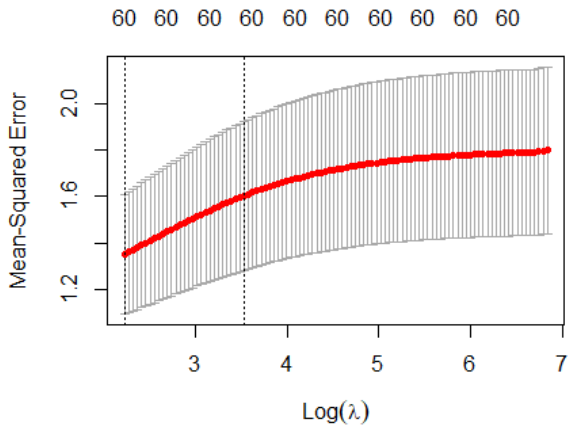


Figure 4: Légende de l'image 1

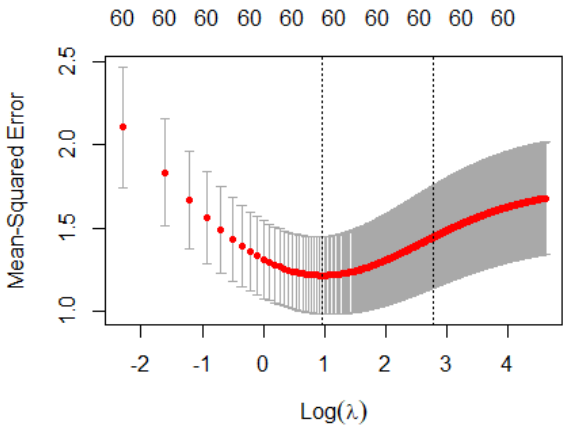


Figure 5: Légende de l'image 2

Figure 6: Légende de l'image 1

En utilisant un processus itératif et en le combinant avec la validation croisée, nous avons déterminé un seuil moyen en calculant la moyenne des valeurs de seuil de chaque validation croisée, retenu sur la base de lambda.1se. Cela se traduit par la formule seuil = $\frac{\sum \min(\lambda)}{10}$. Nous avons ainsi établi un seuil de 2.4.

Df	%Dev	Lambda
60	73.37	2.4

Table 2: Résumé du modèle

En résumé, notre modèle a été ajusté avec 60 coefficients non nuls, expliquant environ 73,37% de la déviance, avec un paramètre de régularisation lambda égal à 2.4. Donc, le modèle a été régularisé pour éviter un surajustement, et le choix du paramètre lambda peut être basé sur une validation croisée ou un autre critère.

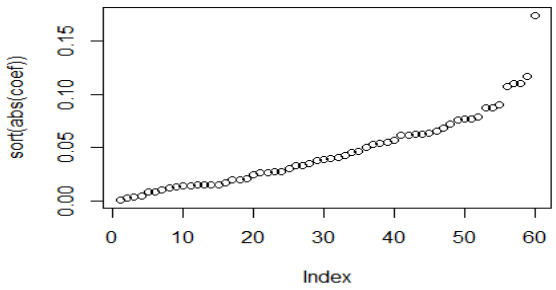


Figure 7: Coefficient du modèle de ridge pour les différentes ports

La régularisation Ridge vise à conserver tous les prédicteurs dans le modèle tout en réduisant leurs coefficients afin d'éviter le surajustement. Bien que les coefficients ne soient pas nuls, leur magnitude relative sert à évaluer l'importance relative des ports dans la prédiction des émissions de CO2. En ce sens, en se basant sur la Figure 5, nous avons choisi de retenir les 8 premières variables qui sont les coefficients associés aux ports 3, 5, 8, 34 et 58 avec une valeur absolue supérieure à 0.6. Cette décision découle de l'observation que ces ports exercent une influence significative sur les émissions de CO2 des entreprises de transport maritime. Ainsi, selon

ce modèle, ces ports émergent comme les principaux contributeurs aux émissions de CO2 dans le cadre de cette étude.

4.1.2 Régression Linéaire LASSO

Nous avons continué notre exploration avec le modèle de régression Lasso.

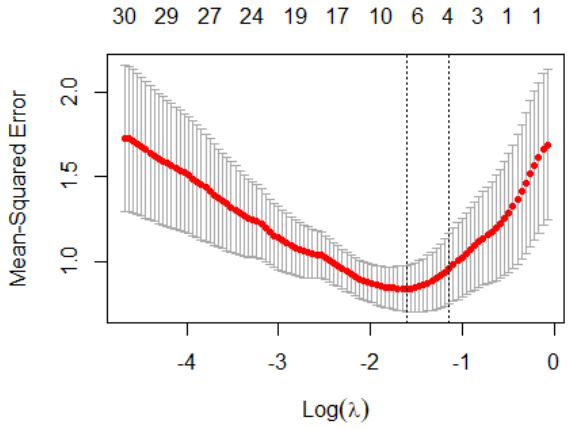


Figure 8: Légende de l'image 1

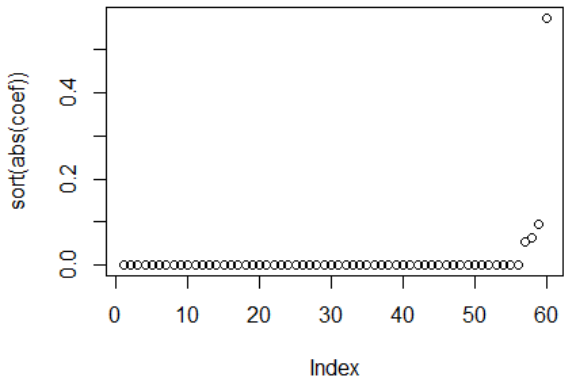


Figure 9: Légende de l'image 2

Comparativement à notre modèle Ridge, le Lasso se distingue par sa capacité à forcer certains coefficients à zéro, réalisant ainsi une sélection automatique des variables. Dans le contexte de ce modèle Lasso, les coefficients associés aux ports 3, 5, 7 et 58.

4.1.3 Régression ElasticNet

Dans la suite de notre analyse, afin de mieux comprendre comment différents prédicteurs influent sur notre réponse, nous envisageons d'établir un modèle Elastic Net. Cette méthode combine les pénalités L1 du Lasso et L2 du Ridge, offrant ainsi une flexibilité accrue dans la sélection des variables. L'utilisation de l'Elastic Net est pertinente dans le contexte où il peut contribuer à atténuer les limitations spécifiques du Lasso tout en conservant ses avantages en termes de sélection automatique de variables. Cette approche hybride permettra de mieux appréhender la complexité des relations sous-jacentes entre les quantités transportées dans les ports et les émissions de CO2 des entreprises de transport maritime. Naturellement, dans ce cas ci-dessus, la régression Elasticnet retient l'union des ports retenus entre Ridge et Lasso à savoir: 3, 5, 7, 8, 34 et 58.

4.1.4 Modèle Linéaire Classique

Nous allons réaliser le modèle linéaire classique à partir des variables retenues. Nous prenons celles issues de la régression LASSO car elles sont souvent moins corrélées et le modèle sera ainsi plus stable.

Table 3: Résumé du modèle de regression liénaire

	Dependent variable:
	Y
Port.Port.3	0.094
	(0.078)
Port.Port.5	0.430***
	(0.074)
Port.Port.7	-0.313***
	(0.093)
Port.Port.58	0.060
	(0.040)
Constant	5.065***
	(1.792)
Observations	37
R^2	0.729
Adjusted R^2	0.695
Residual Std. Error	0.723 (df = 32)
F Statistic	21.555*** (df = 4; 32)
Note:	*p<0.1; **p<0.05; ***p<0.01

Le modèle est statistiquement significatif avec une p-value¡0.05 et il semble également bien ajusté avec un R^2 de 0.69, ce qui suggère qu'il explique bien une partie de la variabilité des données. La variable la plus significative est le : Port5 estimé à une valeur de 0.42989. Cette dernière indique que, toutes choses étant égales par ailleurs, une unité d'augmentation de marchandise transportée en millions de tonne de dans le Port.5 est associée à une augmentation d'environ 0.42989 unité du CO2.

4.1.5 Régression sur Composantes Principales (PCA)

La régression sur composantes principales (PCA) émerge comme une approche intéressante dans les modélisations où la matrice des variables explicatives est volumineuse ou instable. En régressant la variable CO2 sur les composantes principales, nous réduisons efficacement la dimension de l'espace des caractéristiques tout en conservant l'information cruciale. Cette méthode devient particulièrement avantageuse lorsque le nombre initial de variables explicatives est élevé, permettant ainsi une modélisation concise et interprétable, tout en minimisant les risques de surajustement. La PCA offre ainsi une solution élégante pour traiter des ensembles de données complexes tout en préservant l'essentiel de l'information liée à la variable CO2.

Table 4: Résumé du modèle

	TRAINING: % variance explained																
Components	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
X	11.91	21.74	30.67	38.54	45.42	51.98	57.54	62.55	67.17	71.21	75.09	78.12	80.87	83.17	85.12	87.02	88.73
Y	10.07	15.90	24.69	40.01	50.56	53.59	54.61	63.65	64.96	66.15	67.54	67.90	69.36	75.77	76.02	78.59	78.80
Components	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	
X	90.30	91.63	92.82	93.88	94.88	95.77	96.45	97.10	97.64	98.16	98.55	98.89	99.20	99.43	99.58	99.71	
Y	78.87	78.99	80.46	80.49	80.50	80.51	82.61	82.77	82.82	89.36	89.38	89.65	91.38	92.27	96.67	97.66	
Components	34	35	36														
X	99.83	99.92	100														
Y	98.64	98.72	100														

La première composante principale pour X explique 11.91 % de la variance totale, et pour Y, elle explique 10.07 %Ces résultats suggèrent que la première composante capture une proportion significative de la variabilité dans les données. En ajoutant des composantes successives, on observe une augmentation cumulative de la variance expliquée. Par exemple, les 10 premières composantes expliquent environ 75.09 % de la variance pour X et 66.15 % pour Y. Nous pouvons noter que, même avec un nombre restreint de composantes, une proportion substantielle de la variance totale peut être expliquée.

Au-delà de l'implémentation de la régression, le choix du nombre optimal de composantes revêt une importance cruciale. Bien que l'inclusion de toutes les 36 composantes puisse garantir la préservation totale de l'information, notre analyse révèle que bien avant d'atteindre ce nombre maximal, nous avons déjà capturé une quantité significative d'informations. Pour étayer notre sélection de composantes, nous nous appuyons sur des diagnostics, en particulier le RMSEP (Racine Carrée de l'Erreur Quadratique Moyenne de Prédiction). Le RMSEP mesure la précision des prédictions d'un modèle sur un ensemble de données de test. En ajustant le modèle pour divers nombres de composantes, notre objectif est de choisir le nombre optimal qui minimise le RMSEP, indiquant ainsi un équilibre optimal entre la complexité du modèle et sa capacité à généraliser sur de nouvelles observations.

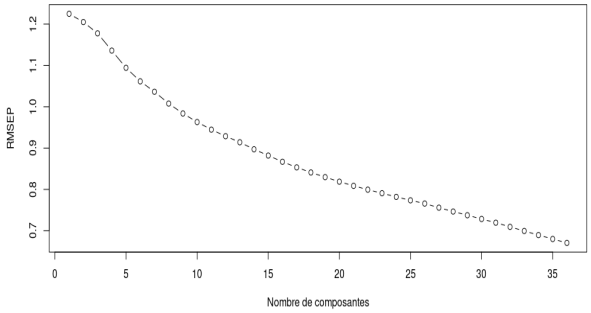


Figure 10: RMSE pour le choix du nombre de composantes

À la lumière des résultats, la rétention de 16 composantes semble être une option judicieuse. En comparaison avec d'autres techniques de régression pénalisées, nous conservons davantage de variables. Bien que nous ayons envisagé d'explorer plus avant en considérant jusqu'à 20 composantes, opter pour 16 demeure une choix très viable.

Pour la sélection des coefficients, nous avons choisi de fixer un seuil à 0,3 et de retenir les coefficients dont la valeur absolue dépasse ce seuil. Cette approche nous permet de conserver tant les coefficients positifs que négatifs qui surpassent ce seuil. En conséquence, les coefficients des ports les plus importants sont ceux correspondant aux positions 3, 11, 5, 26, 37, et 50.

4.1.6 Régression PLS (Partial Least Squares)

Nous allons maintenant explorer une autre approche dans le cadre standard, à savoir la méthode des moindres carrés partiels (PLS, pour Partial Least Squares). Cette méthode constitue une alternative que nous allons examiner plus en détail dans la suite de cette section.

Table 5: Résumé de la regression PLS

Composantes	1	2	3	4	5	6	7	8	9	10
X	9.075	17.28	23.17	28.79	34.25	38.51	44.43	49.21	54.01	57.50
Y	69.775	80.21	86.23	89.98	92.86	95.56	96.56	97.35	98.11	98.85
Composantes	11	12	13	14	15	16				
X	60.37	64.54	67.84	71.85	74.51	77.17				
Y	99.39	99.62	99.76	99.82	99.89	99.93				

Les valeurs présentées dans le tableau représentent les contributions cumulatives des composantes pour les variables X et Y. Ces contributions cumulatives indiquent la proportion de variation expliquée dans les variables à mesure que l'on considère un nombre croissant de composantes. Par exemple, la première composante explique 9.08% de la variation dans les quantités transportées et 69.78% de la variation dans les émissions de CO2.

La lecture des contributions cumulatives permet de sélectionner le nombre optimal de composantes pour la modélisation. Dans notre exemple, on observe que les dix premières composantes expliquent respectivement 57.50% des quantités transportées et 98.85% des émissions de CO2. Cette information peut orienter la décision quant au nombre de composantes à retenir pour une modélisation efficace.

4.2 Modélisation du nombre d'incidents

Dans la continuité de notre analyse, nous explorons la relation entre la performance logistique et la quantité transportée. Notre objectif dans cette section est d'identifier les ports ayant une influence significative sur le nombre d'incidents logistiques. Pour assurer une comparaison cohérente, nous appliquerons les modèles Lasso et Ridge, utilisés précédemment, tout en adaptant notre approche à la nature de notre variable de réponse, une variable de comptage. Ainsi, nous opterons pour la loi de Poisson pour modéliser ce phénomène.

Le choix de la loi de Poisson se justifie par sa pertinence dans le contexte des variables de comptage, où l'on mesure le nombre d'occurrences d'un événement dans un intervalle fixe. Dans notre cas, cela correspond au décompte du nombre d'incidents dans différents ports. Cette approche simplifiée permettra d'apporter des éclairages supplémentaires sur les ports ayant un impact notable sur le volume de marchandises. Ces résultats, intégrés dans notre démarche sur Overleaf, contribueront à une analyse concise et pertinente.

4.2.1 Régression Linéaire Ridge

Df	%Dev	Lambda
60	90.5	2.09

Table 6: Résumé du modèle Poisson

Le modèle de régression de Poisson a été ajusté avec 60 coefficients non nuls, expliquant environ 90.5% de la déviance, avec un paramètre de régularisation lambda égal à 2.09. Dans le contexte de notre étude, cela suggère que les quantités transportées dans ces ports ont une influence significative sur le nombre d'incidents, et le modèle a été régularisé pour éviter un surajustement. Par ailleurs, rappelons que le choix du paramètre lambda a été basé sur une validation croisée.

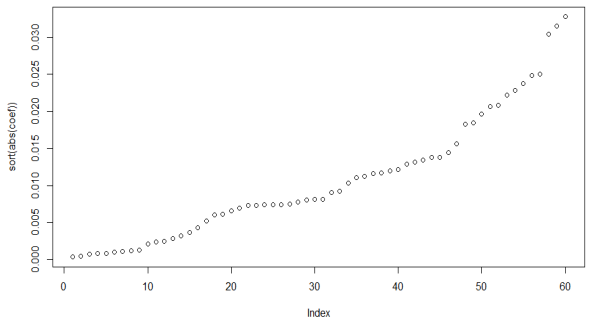


Figure 11: Coefficient du modèle de ridge-poisson pour les différentes ports

Dans cette étape, l'application de la technique du seuil prédéfini s'avère délicate, contrairement à notre modèle de Ridge précédent où la détermination du seuil adéquat était relativement précise. En revanche, dans le contexte de la modélisation des nombres d'incidents, le choix du seuil devient complexe. Malgré cette complexité, nous avons fixé notre seuil à 0.02, ce qui nous a permis de retenir les ports 3, 5, 8, 11, 15, 16, 34, 37, 50, et 58 en tant que ceux ayant un impact significatif sur le nombre d'incidents.

Il est important de souligner que notre choix de seuil repose sur notre expertise subjective. Idéalement, l'utilisation d'un modèle Lasso offrirait une approche automatique du choix du seuil, une piste que nous explorerons. Cette démarche nous permettra de confronter une fois de plus les résultats obtenus par les deux modèles, enrichissant ainsi notre analyse comparative.

4.2.2 Régression Linéaire Lasso

Nous mettons en place notre modèle lasso comme précédement cette fois si la loi est une loi de poisson.

Df	%Dev	Lambda
4	55.78	0.32

Table 7: Résumé du modèle Poisson

En ce qui concerne l'interprétation, le modèle de régression de Poisson a été ajusté avec 4 coefficients non nuls, expliquant environ 55.78% de la déviance, avec un paramètre de régularisation lambda égal à 0.32.

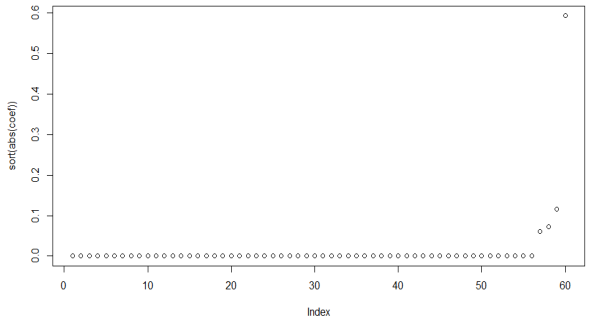


Figure 12: Coefficient du modèle de lasso-poisson pour les différentes ports

Cette fois-ci, la sélection des ports s'est opérée de manière plus sélective, se limitant aux ports 3, 5, 7, et 58. Ce choix restreint a été guidé par l'application d'un modèle de Lasso, une approche qui privilégie une sélection automatique de variables en favorisant la parcimonie.

4.2.3 Régression de Poisson « classique »

Dans le cadre de notre analyse, nous avons entrepris la réalisation d'une régression de Poisson « classique » en utilisant les meilleures variables que nous avions précédemment sélectionnées. Cette démarche s'inscrit dans notre volonté de modéliser de manière robuste et précise le nombre d'incidents dans les ports maritimes, en mettant en lumière l'influence respective de ces variables sur la réponse.

Table 8: Régression de poisson classique

	Dependent variable:
	Y
ports_selectionneesPort.Port.3	0.015
	(0.040)
ports_selectionneesPort.Port.5	0.056
	(0.037)
ports_selectionneesPort.Port.7	−0.040
	(0.047)
ports_selectionneesPort.Port.58	0.008
	(0.021)
Constant	1.639*
	(0.922)
Observations	37
Log Likelihood	−Inf.000
Akaike Inf. Crit.	Inf.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Chaque coefficient indique la variation attendue dans du nombre d'incidents par unité d'augmentation de la variable respective. Par exemple, le coefficient associé à "Port_5" suggère qu'une augmentation d'une unité dans cette variable est associée à une augmentation d'environ 5.7% du nombre d'incidents. Ces résultats fournissent des insights précieux sur l'impact relatif de chaque port sur la fréquence des incidents.

4.3 Modélisation de la variable CA

Dans cette section, notre objectif est d'explorer la relation entre le chiffre d'affaires des entreprises de transport maritime et les quantités de marchandises livrées dans les principaux ports. La variable clé ici est le chiffre d'affaires, que nous avons catégorisé en deux modalités : 1 pour les entreprises dont le chiffre d'affaires dépasse 10 millions d'euros et 0 sinon. Cette distinction nous permet de modéliser la relation entre le chiffre d'affaires et les quantités livrées dans les ports à l'aide de méthodes de régression logistique pénalisée, notamment la régression logistique Ridge, la régression logistique LASSO, et Elastic Net.

L'approche reste assez similaire à celle implémentée dans les sections précédentes, tout en ajustant notre modèle pour tenir compte de la nature binaire de la variable réponse, le chiffre d'affaires. Cette variable suit une distribution binomiale, où chaque observation est catégorisée comme ayant un chiffre d'affaires supérieur à 10 millions d'euros (1) ou non (0). La loi binomiale est particulièrement adaptée pour modéliser ces situations où chaque observation peut appartenir à l'une de deux catégories exclusives.

À travers cette approche analytique, notre objectif est d'apporter des insights significatifs dans le cadre de l'analyse du transport maritime. Nous mettrons en lumière les facteurs les plus déterminants, notamment les ports dans le succès financier de ces acteurs majeurs du secteur, en se concentrant spécifiquement sur les résultats obtenus à partir des modèles de régression pénalisée.

4.3.1 Régression logistique Ridge

Dans le cadre de notre modélisation de rdige, nous nous sommes tout d'abord assuré de définir une modalité de référence et comme celle de référence, nous avons pris la modalité 1. En soit, dans le contexte d'une régression logistique, le choix de la catégorie de référence est arbitraire et n'affecte pas la validité des résultats ou l'interprétation des coefficients. La régression logistique modélise les log-odds (rapports des chances) d'appartenir à une catégorie par rapport à une catégorie de référence. La catégorie de référence est simplement utilisée comme point de comparaison pour les autres catégories.

Df	%Dev	Lambda
60	44.85	0.7

Table 9: Régression Logistique Ridge

Le modèle de régression logistique Ridge a été ajusté avec succès, montrant que toutes les 60 variables explicatives sont actives dans le modèle, indiquant une complexité relative. Environ 44.85% de la déviance est expliquée, ce qui dénote une capacité significative à rendre compte de la variabilité du chiffre d'affaires. La valeur optimale du paramètre de pénalité (lambda) est de 0.7, indiquant une régularisation modérée. Ces résultats soulignent l'équilibre atteint par le modèle entre complexité et pouvoir explicatif dans la modélisation de la relation entre les quantités livrées dans les ports et le chiffre d'affaires des entreprises de transport maritime.

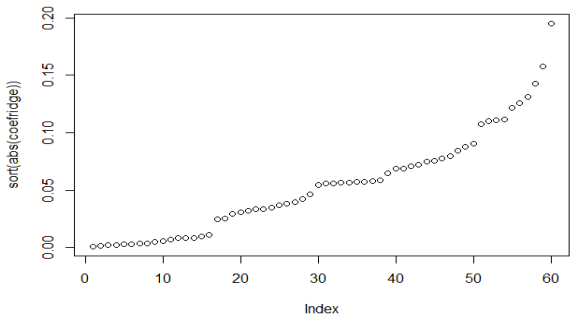


Figure 13: Coefficient du modèle de logistique ridge pour les différentes ports

L'étape de sélection des variables a été réalisée en ordonnant les coefficients absolus du modèle de Ridge de manière décroissante. Pour une sélection restreinte de trois ports significatifs, nous avons identifié les ports 5, 29 et 58 comme les plus importants. Pour une sélection élargie de six ports, les ports 5, 29, 58, 55, 24 et 40 ont été retenus. Cette approche offre une manière rigoureuse de cibler les ports les plus significatifs, fournissant ainsi des informations précieuses pour orienter les décisions stratégiques des entreprises de transport maritime.

4.3.2 Régression logistique LASSO

Une nouvelle fois, nous avons appliqué la technique de LASSO dans le contexte de la modélisation logistique, en comparant la méthode automatique de sélection de coefficients à notre approche. Le modèle a été ajusté en utilisant une stabilisation basée sur la cross-validation, et la sélection du lambda optimal a été réalisée à partir de lambda.1se.

Df	%Dev	Lambda
8	41.3	0.1

Table 10: Résultats de la Régression Logistique LASSO

L'interprétation des résultats indique que le modèle final sélectionne 8 variables explicatives, expliquant environ 41.3% de la déviance du modèle. Le lambda optimal, choisi à partir de lambda.1se lors de la cross-validation, est de 0.1. Ce choix de lambda permet d'obtenir un équilibre approprié entre la complexité du modèle et son pouvoir explicatif, fournissant ainsi une sélection de variables robuste pour la relation étudiée.

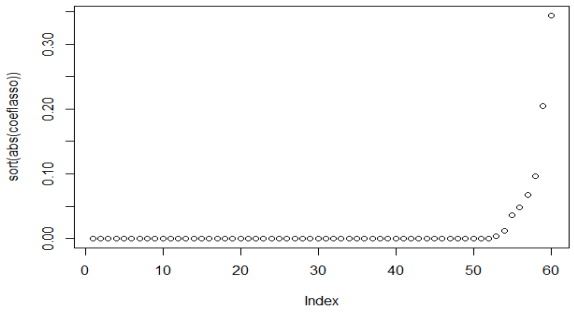


Figure 14: Coefficient du modèle de logistique LASSO pour les différentes ports

L'application de la technique de LASSO dans la modélisation logistique a conduit à la sélection automatique des ports les plus pertinents dans la relation entre les quantités livrées dans les ports et le chiffre d'affaires des entreprises de transport maritime. Les ports retenus par le modèle LASSO sont "5", "29", "24", "27", "40", et "58". En comparaison avec la technique de Ridge utilisée précédemment, qui a identifié les ports "5", "29", et "58" comme les plus importants, le modèle LASSO a inclus des ports supplémentaires dans sa sélection. Cette différence peut s'expliquer par la nature de la pénalité L1 dans LASSO, qui tend à favoriser la sélection de variables individuelles, conduisant potentiellement à une sélection plus parcimonieuse et diversifiée des ports. Ainsi, la comparaison des résultats entre ces deux techniques offre des perspectives intéressantes sur les ports considérés comme les plus déterminants dans la relation étudiée.

4.3.3 AUC et Interprétation.

Nous avons opté pour l'utilisation du modèle de régression logistique LASSO afin d'évaluer la probabilité que le chiffre d'affaires des entreprises de transport maritime dépasse les 10 millions d'euros. Ce choix s'avère particulièrement pertinent en raison de la capacité automatique du modèle à sélectionner les coefficients significatifs. À partir de ce modèle, nous procéderons à l'interprétation de l'Aire Sous la Courbe (AUC) ainsi que des coefficients significatifs qui contribuent à la compréhension de la relation entre les ports et la probabilité de dépassement du seuil de chiffre d'affaires.

L'AUC mesure la capacité d'un modèle à discriminer entre les deux classes, en l'occurrence, les entreprises dont le chiffre d'affaires est supérieur à 10 millions d'euros et celles dont le chiffre d'affaires est inférieur ou égal à cette valeur seuil. Une AUC proche de 1 indique une excellente capacité discriminante. Nous avons évalué notre modèle LASSO en calculant l'AUC, et le résultat obtenu est de 0.9667. Cette valeur élevée suggère que le modèle est robuste dans sa capacité à distinguer entre les deux catégories de chiffre d'affaires.

Dans le cadre du modèle logistique LASSO, les coefficients associés à chaque port représentent la contribution à la log-odds de la probabilité que le chiffre d'affaires soit supérieur à 10 millions d'euros.

- **Port 5 :** Un coefficient de 0.3445 indique que, toutes choses égales par ailleurs, la log-odds de la probabilité que le chiffre d'affaires dépasse 10 millions d'euros est augmentée de 0.3445 unité pour une entreprise opérant dans le port 5 par rapport à la modalité de référence.

- **Port 24 :** Un coefficient de 0.0966 indique une augmentation de 0.0966 unité dans la log-odds pour une entreprise opérant dans le port 24 par rapport à la modalité de référence.
- **Port 27 :** Avec un coefficient de 0.0678, le port 27 contribue positivement à la log-odds de la probabilité de chiffre d'affaires supérieur à 10 millions d'euros.
- **Port 29 :** Un coefficient de 0.2045 pour le port 29 indique une contribution positive à la log-odds.
- **Port 40 :** Le port 40 contribue positivement avec un coefficient de 0.0478.
- **Port 43 :** Bien que le coefficient soit faible (0.0035), le port 43 contribue positivement.
- **Port 46 :** Avec un coefficient de 0.0125, le port 46 a une contribution positive.
- **Port 58 :** Un coefficient de 0.0365 indique une contribution positive du port 58.

La modalité de référence est définie pour une entreprise dont le chiffre d'affaires est inférieur ou égal à 10 millions d'euros. Ces coefficients permettent de comprendre l'impact de chaque port sur la probabilité du dépassement de ce seuil, contribuant ainsi à une meilleure compréhension des facteurs influençant le chiffre d'affaires dans le secteur du transport maritime.

4.4 Modélisation de la variable CA3 polytomique

4.4.1 Régression multinomiale pénalisée

Dans le contexte de notre étude, nous avons entrepris une régression multinomiale pénalisée sur la variable d'intérêt CA3, initialement sans prendre en compte l'aspect ordonné (sans contrainte de glmnet). Pour ce faire, nous avons adopté une approche reposant sur le modèle de régression LASSO, faisant usage de la fonction `glmnet` de R.

Notre démarche s'est articulée autour de la mise en place d'une modélisation séquentielle visant à déterminer le seuil optimal pour la construction du modèle. Cette approche itérative nous a permis d'affiner le choix du lambda, un paramètre de régularisation, afin d'obtenir un modèle robuste tout en évitant la surajustement.Le modèle final, ajusté avec la fonction `glmnet`, nous fournit les résultats suivants :

Df	%Dev	Lambda
13	37.72	0.1

Table 11: Résultats de la Régression Multinomiale Pénalisée

En interprétant ces résultats, nous pouvons constater que le modèle final, avec un degré de liberté de 13, explique 37.72% de la déviance dans les données. Le paramètre de régularisation lambda optimal est fixé à 0.1, démontrant ainsi la capacité du modèle à généraliser tout en conservant une complexité suffisante.

Dans le processus de sélection des variables, nous avons identifié les ports suivants comme étant les plus significatifs dans la prédiction de la variable CA3 : Port 5, Port 24, Port 29, Port 40, Port 43, et Port 54. Ces ports ont démontré une contribution significative à la compréhension du phénomène étudié, soulignant ainsi leur importance dans l'explication des variations observées dans les niveaux de chiffre d'affaires selon les catégories définies.

4.4.2 Régression polytomique ordonnée

Dans cette phase de l'analyse, une fois que nous avons sélectionné les variables les plus pertinentes, nous avons procédé à la réalisation d'une régression polytomique ordonnée pour modéliser la relation entre les quantités de marchandises transportées dans différents ports et la variable d'intérêt CA3. Nous avons utilisé la fonction `clm` de R pour ajuster un modèle de régression polytomique ordonnée. Les résultats obtenus sont les suivants :

Table 12: Régression polytomique ordonnée

	Dependent variable: CA3
Port.Port.5	1.258^{***}
	(0.453)
Port.Port.24	0.639^{*}
	(0.328)
Port.Port.29	0.659^{**}
	(0.259)
Port.Port.40	0.685^{***}
	(0.250)
Port.Port.43	0.393^{*}
	(0.207)
Port.Port.54	0.987^{**}
	(0.483)
Observations	37
Log Likelihood	−16.258
Note:	*p<0.1; **p<0.05; ***p<0.01

Les coefficients pour chaque port indiquent la contribution de chaque port à la probabilité d'appartenir à une catégorie spécifique de CA3. Par exemple, un coefficient positif pour "Port.Port.5" suggère une augmentation de la probabilité de passer de CA3=0 à CA3=1, tandis qu'un coefficient positif pour "Port.Port.29" suggère une augmentation de la probabilité de passer de CA3=1 à CA3=2.

Les seuils ("Threshold coefficients") déterminent les points de basculement entre les catégories de CA3. Par exemple, le seuil entre CA3=0 et CA3=1 est estimé à 39.76, et entre CA3=1 et CA3=2 est estimé à 44.70.

Ces résultats nous permettent de comprendre comment les quantités de marchandises transportées dans différents ports influent sur les niveaux de chiffre d'affaires des entreprises de transport maritime, en distinguant les transitions entre les différentes catégories de CA3. Les ports significatifs identifiés contribuent de manière notable à ces transitions, offrant ainsi des informations cruciales pour les décisions stratégiques dans le secteur maritime.

5 Conclusion

Cette étude propose une vision globale des performances des entreprises de transport maritime en examinant les liens entre des facteurs clés tels que les émissions de CO2, le chiffre d'affaires, le nombre de retards, et les volumes de marchandises transportées. Les résultats de ces analyses peuvent fournir des insights essentiels pour orienter les décisions stratégiques visant à optimiser l'efficacité opérationnelle et à encourager des pratiques plus durables au sein du secteur du transport maritime.

References

[1] Marcel Masson. 14. la société du canal de provence et le transport d'eau par voie maritime. *Journées de l'hydraulique*, 22(4):1–10, 1992.

[2] Charles Raux. Réduire les émissions de co2 dans le transport: un marché de permis pour les automobilistes et le frêt. *Transports*, (445):pp–285, 2007.

[1] [2]